GUIDELINE DEVELOPMENT ON HEALTHCARE RELATED TESTING

Mariska Tuut

Guideline development on healthcare related testing

Mariska Tuut

The research presented in this thesis was conducted at CAPHRI Care and Public Health Research Institute, department of Family Practice at Maastricht University. CAPHRI participates in the Netherlands School of Public Health and Care Research (CaRe).

Colophon
Lay-out: Mariska Tuut
Cover design: Ridderprint | www.ridderprint.nl
Printing: Ridderprint | www.ridderprint.nl

ISBN: 978-94-6506-258-7

Guideline development on healthcare related testing

Proefschrift

voor het behalen van de graad van doctor
aan de Universiteit Maastricht
onder gezag van Rector Magnificus, Prof. dr. Pamela Habibović,
overeenkomstig het besluit van het College van Decanen,
te verdedigen in het openbaar
op dinsdag 8 oktober 2024 om 13.00 uur

door

Margaretha Klaasje Tuut
geboren op 26 januari 1975
in Groningen

Promotores
Prof. dr. T. van der Weijden
Prof. dr. J.S. Burgers

Co-promotor
Dr. M.W. Langendam (University of Amsterdam)

Beoordelingscommissie
Prof. dr. M.H.J.M. Majoie (voorzitter)
Prof. dr. P.J. van der Wees (Radboud Universiteit Nijmegen)
Prof. dr. S.M.A.A. Evers
Dr. M.M. Tabbers (University of Amsterdam)

Table of contents

Preface

'Good guidelines can only make you better' [1].

'The challenge of scientific research is not to find answers, but to formulate the question.' [2].

'Guideline development reveals the dilemmas and uncertainties associated with the application of medical knowledge. The guideline should not cover this up, but make it transparent, and link patient decision aids to preference sensitive recommendations.' [3].

The above three propositions, cited from my supervisors have been published decades ago, but still underpin the urgency of this thesis. These statements not only show confidence in the profession of guideline development, but also enlighten ongoing challenges in guideline development methods. But foremost, these propositions inspire. They align with my experience as a guideline methodologist, in which I had, and have the opportunity to work with so many dedicated healthcare professionals and patient(representative)s, methodologists/process leads and guideline panel chairs, in whom I saw enthusiasm and expertise, but in whom I also saw their struggles in using the right ingredients in the right way to 'cook the right guidelines'.

It is my personal ambition to improve and facilitate guideline development methods - especially in the area of recommendations about healthcare related testing – and thereby to be able to contribute to the improvement of healthcare quality. Therefore, this thesis focuses on knowledge and tools that can help guideline developers (in the broadest sense) in appropriately developing recommendations about healthcare related testing.

References

1. Burgers JS. Quality of clinical practice guidelines. Nijmegen: Catholic University Nijmegen; 2002.
2. Langendam MW. The impact of harm reduction-based methadone treatment on HIV infection and mortality. Amsterdam: University of Amsterdam; 2000.
3. Van der Weijden T. Richtlijnen in de spreekkamer, van dogma naar dans. Maastricht: Maastricht University; 2010.

Chapter 1.

General introduction

General introduction

This introduction chapter guides through the various pillars that are essential for addressing challenges in guideline development and healthcare related testing in practice. It sets the rationale for this thesis, outlining and bringing together the worlds of guideline development, testing in practice, and test evaluation in research to finally arrive at the aim of this thesis and the research questions.

Guidelines

Guidelines, including clinical practice guidelines and public health guidelines, are documents providing recommendations intended to optimize patient care. They are developed using a systematic review of the available evidence and an analysis of benefits and harms of alternative care options. To be regarded as trustworthy according to the Institute of Medicine, guidelines should:

- be based on a systematic review of the existing evidence;
- be developed by a knowledgeable, multidisciplinary panel of experts and representatives from key affected groups;
- consider important patient subgroups and patient preferences, as appropriate;
- be based on an explicit and transparent process that minimizes distortions, biases, and conflicts of interest;
- provide a clear explanation of the logical relationships between alternative care options and health outcomes;
- provide ratings of both the quality of evidence and the strength of recommendations; and
- be reconsidered and revised as appropriate when important new evidence warrants modifications of recommendations [1].

Guideline development follows a clear process, which is crucial for acceptance and implementation. The first step concerns an analysis of problems to be addressed, the identification of the specific topic, target group(s) and target population of the guideline. Next, a guideline panel (also known as a guideline development group/committee) is established, consisting of representatives from all relevant professional groups, patient/consumer/people representatives and methodologists. Following that, the scope of the guideline is defined including the formulation of key questions that need to be addressed. After that, a draft guideline is developed. This process includes a series of steps, in which available guidelines are reviewed, scientific evidence is identified and critically assessed, and relevant expertise and experience is considered, after which draft recommendations are formulated. Next, the draft guideline is disseminated to all relevant stakeholders and target groups for

comments and feedback. This step may include pilot testing of the draft guideline to identify barriers for implementation. Then, the final version of the guideline can be submitted for endorsement or authorisation. Finally, the guideline, and any related materials, such as summaries, patient versions and decision support tools are published. The guideline outlines specific criteria for reviewing and updating the guideline [2]. Note that endorsement and authorisation is not universal in guideline development worldwide. In the Netherlands, authorised guidelines become part of the professional standard for healthcare providers. This guarantees legal embedding of guidelines in the healthcare process and fosters their implementation.

Several manuals and guides are available for the development of guidelines [3-5]. The Grades of Recommendation, Assessment, Development and Evaluation (GRADE) Working Group was established in 2000 to provide assistance for the process of guideline development. The GRADE approach highlights the importance of evaluating the certainty of evidence in the development of recommendations, for example by assessing risk of bias and indirectness [6]. Another crucial aspect of this methodology is its emphasis on clinical relevant differences in outcomes that are regarded as important by patients and consumers, so-called people-important outcomes [7]. The GRADE evidence-to-decision framework systematically considers relevant issues such as balance of benefits and harms, values, resources, and acceptability [8, 9]. The GRADE Working Group has produced and continues to produce comprehensive guides for guideline development [7, 10-25]. The GRADE approach has been adopted by many organisations worldwide, including the Netherlands [26].

In the GRADE approach, special attention is given to the development of guideline recommendations on testing, as the link between testing and the impact on people-important outcomes is indirect and requires a specific approach [27-30]. This includes consideration of the consequences of false positive, false negative, and inconclusive test results, specific risk of bias assessment, moving from test results to people-important outcomes (so-called linked evidence), and the need for formal or informal modelling.

Competencies needed for guideline development

While the essential steps for guideline development have been outlined [31-33], there is limited understanding of the competencies required for the appropriate development of guidelines, particularly those that feature recommendations about testing.

Some research has been conducted in this area: Sultan et al. provided a theoretical framework for competencies and educational milestones that should be acquired by

guideline developers for example through training. The authors identified three core competencies:

1. Facilitate the development of guideline structure and setup
2. Make judgments about the quality or certainty of the evidence
3. Transform evidence to a recommendation

These core competencies are divided into subcompetencies and milestones. Additionally, the authors acknowledge that a guideline panel includes various roles, i.e. chair, methodologist, and panel members, with different competencies [34]. The specific knowledge and competencies needed for creating guidelines on testing are not explicitly incorporated in this framework.

Testing and people-important outcomes

A test refers to any procedure performed on a person to detect, diagnose or monitor a condition. This includes testing of a person's fluids, cells, tissue, functioning and subjective experience. The final objective of testing is to improve people-important outcomes (and/or to prevent deterioration of people-important outcomes). Additional objectives may include offering other benefits (such as simplifying healthcare organisation or reducing expenses) without worsening people-important outcomes.

People-important outcomes, also known as people relevant outcomes, patient important outcomes, patient relevant outcomes or patient-centered outcomes, are components of people's (health) status following an intervention. These outcomes serve to evaluate the effectiveness of the intervention [35]. People-important outcomes may differ depending on the condition and the individual. Common examples include mortality, morbidity, quality of life, and quality of life subscales such as functioning capacity and societal participation.

When assessing the effectiveness of a specific treatment, the link between treatment and change in people-important outcomes is usually clear. For example, antibiotic treatment is related to curing bacterial pneumonia (and reducing mortality), radiotherapy is linked to reducing pain in patients with bone metastases, and hip replacement surgery to improved walking function (although side effects and complications should be considered in all cases).

Unlike treatment, testing itself typically has no immediate effect on people-important outcomes, although reassurance when a serious illness is ruled out, and the occurrence of serious burden (such as serious adverse events) due to testing are common exceptions to this statement. In general, to progress from testing to people-

important outcomes, a series of essentials steps – such as treatment of a certain condition – should be taken.

Testing in clinical practice

Clinical decision-making with the use of a test or testing strategy is daily practice. Healthcare professionals may consider the use of tests after history taking and physical examination. Patients may also demand for tests for various reasons, such as family history of disease, concern about physical conditions, or the need for regular testing. Most patients have high expectations regarding the value of tests: they do not expect false positive or false negative test results and do believe that test results are reliable. In other words, test results would give them certainty about their health status and reassure them in case of test results in the normal range [36].

Testing is frequently used for diagnostic purposes. In clinical practice, the diagnostic process is an empirical iterative process [37]. It has inductive and deductive elements, based on Bayes' theorem [38]. Bayes' theorem, also known as Bayes' rule, states that the a posteriori probability of an event (such as a disease or condition) is conditional and depends on the a priori probability of that event and test results. Taking medical history (anamnesis), conducting physical examination and routine medical testing (such as routine laboratory tests) are generally inductive processes for making a general diagnosis ('rough selection'). Clinicians use signs and symptoms and combine them inductively to move in a diagnostic direction. This can be seen as hypothesis generation. In addition, specific tests (such as spirometry or a dementia test) can be conducted as part of deductive processes. These are targeted tests, intended to confirm or rule out a specific diagnosis. These can be seen as hypothesis testing. The entire diagnostic process in the clinical practice is called the hypothetico-deductive method [39-41]. The diagnostic process includes both sense (including clinical reasoning, understanding, experience and common sense) and science (including evidence, theory and testing) [42]. Clinical experience, which includes gut feelings ('pluis/niet pluis'), is a crucial element of patient care during consultations [43]. Accordingly, tests serve as complementary tools in clinical practice.

Testing in healthcare

In this thesis, a test or testing refers to all healthcare related tests and testing strategies that are used for different purposes and roles [44]. Thus, this thesis extends beyond the use of tests for diagnostic purposes by healthcare providers in the consultation room to encompass the entire healthcare, including public health.

Healthcare related tests can be used for several purposes: screening, surveillance, risk classification, diagnosis, staging, treatment triage, determination of prognosis and monitoring/follow-up [44, 45]. Examples of these purposes are shown in *table 1*. A single test can serve multiple purposes, such as an MRI for women with increased risk or suspected of, or diagnosed breast cancer. It can be used for screening, risk classification, diagnosis, staging, and monitoring/follow-up.

Table 1. Testing purpose and examples

Testing purposes	Examples
Screening	▪ Faecal occult blood testing in people aged 55-75 years to screen for colorectal cancer ▪ Anoscopy in people with HIV to screen for anal intraepithelial neoplasia to reduce the risk of anal cancer-related mortality ▪ Hip examination in youth care to select infants at high risk of having hip dysplasia
Surveillance	▪ Influenza surveillance to gain insight in the spread and typology of influenza viruses, and their impact ▪ Antimicrobial surveillance to understand antibiotic resistance patterns
Risk classification	▪ Measurement of blood cholesterol levels and blood pressure in primary care patients to stratify the risk of a cardiovascular event ▪ Bone mineral density measurement using DEXA scanning to determine the risk of an osteoporotic fracture
Diagnosis	▪ Urine dipstick to diagnose urinary tract infection in primary care ▪ Amniocentesis including chromosomal testing to rule out trisomy 21 (Down's syndrome) ▪ X-ray to diagnose bone fracture ▪ Vision test to detect visual impairment
Staging	▪ Histology to stage cancer disease ▪ CT scanning in patients with breast cancer to detect metastases ▪ Beck Depression Inventory to assess level of depression
Treatment triage	▪ Allergen testing in patients with asthma to guide asthma management ▪ Bacteriological test to guide antibiotic treatment
Prognosis	▪ 6-minute walk distance test (6MWD) to estimate risk of death in patients with heart failure ▪ Advanced Dementia Prognostic Tool (ADEPT) to estimate survival in people with dementia
Monitoring/ follow-up	▪ Blood glucose monitoring to monitor diabetes mellitus ▪ Weight measurement to monitor weight loss therapy ▪ Spirometry to monitor COPD ▪ Cardiac ultrasound to follow-up patients with heart failure

As illustrated in *table 1*, there is a variety of tests, including self-tests, laboratory tests, imaging, functional tests, and questionnaires, as well as a variety of settings in which testing can be performed, such as public health, primary care, secondary care and long-term care.

History taking and physical examination can also be considered as tests but are outside the focus of this thesis due to their general nature and routine application. Additionally, tests unrelated to healthcare are also outside the scope of this thesis. Such tests include e.g. weight and muscle measurements in gyms, or genealogy tests to trace one's ancestors.

Scientific evaluation of a test

To assess the value of a healthcare related test, different aspects should be taken into account [46]:

- Analytical performance
- Clinical performance
- Clinical effectiveness
- Cost-effectiveness
- Broader impact

These concepts are elaborated on in *box 1*.

Box 1. Components of test evaluation

Analytical performance: this refers to the ability of the test to accurately detect or measure a particular
- measurand. Parameters of analytical performance include:
- trueness: the determination whether the test measures the variable of interest
- precision: the assessment of the reproducibility of the test.
- detection limits: a test might not detect a measurand below or above a certain level or might not be specific enough.
- cross-reactivity: the influence of factors on the test result beyond the measurand of interest.

Clinical performance: this refers to the ability of a test in correctly classifying individuals with and without the target condition (such as a disease). This is also called the diagnostic accuracy of a test. Parameters of diagnostic accuracy can be established by comparing the index and reference tests. The index test is the test of interest, while the reference test (also known as the reference standard) is the test to which the index test is compared. The reference test can be the gold standard, but also other options (such as the test in usual care/practice) are used.
Clinical performance measures can be obtained by categorizing people with and without the target condition according to their test results in a 2x2 table (*table 2*):

Table 2. Clinical performance of a test in a 2x2 table

	People with the target condition	People without the target condition	Total
Positive test result	TP	FP	TP+FP
Negative test result	FN	TN	FN+TN
Total	TP+FN	FP+TN	Total

TP: true positives, FP: false positives, FN: false negatives, TN: true negatives

Such a table provides insight into the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) test results. A test can have an inconclusive result as well. Other frequently used parameters of the clinical performance of a test include:
- sensitivity: the probability of getting a positive test result in people with the target condition (TP/(TP+FN))

- specificity: the probability of getting a negative test result in people without the target condition (TN/(TN+FP))
- positive predictive value (PPV): the probability of having the target condition in people with a positive test result (TP/(TP+FP))
- negative predictive value (NPV): the probability of not having the target condition in people with a negative test result (TN/(TN+FN))

Clinical effectiveness (also known as clinical utility): this refers to the ability of a test to improve people-important outcomes.
Cost-effectiveness: this refers to the assessment of changes in costs and people relevant outcomes resulting from the introduction of a test. There are several perspectives from which costs can be determined, such as the individual patient perspective (e.g. costs that patients have to pay to undergo a test), the healthcare perspective (e.g. costs because of time invested by healthcare professionals and other resources needed for performing a test) and societal perspective (e.g. costs of testing covered by health insurance).
In all perspectives, direct costs (such as costs of the tests), and indirect costs can be taken into account. Indirect costs could include travel expenses and costs for childcare for the patient while travelling to the hospital, loss of income, and social security expenses due to absence at work.
Broader impact: this refers to consequences of the test beyond clinical effectiveness and cost-effectiveness, such as acceptability, implementability, and consequences on legal, ethical, and organisational issues.

Besides the above-mentioned evaluation scenarios, it is essential to define the role of a new test in comparison to the existing test, as this influences the interpretation of the new test's value. Various roles are acknowledged [47, 48]:

- Triage
- Replacement of a reference standard or an existing test
- Add-on
- Parallel/combined

These roles are described in detail in *table 2*.

All of the above factors can be relevant when considering the benefits and harms of testing in specific circumstances and for specific populations.

Impact of inappropriate testing

There is considerable practice variation in test usage in practice, with both underuse and overuse of tests being common [49, 50]. Sullivan et al. conducted a systematic review on over- and undertesting in primary care, in which they explored the frequency of inappropriate ordering of 103 diagnostic tests in relation to their respective guidelines. The results showed a wide range of non-compliance to the testing recommendations in guidelines (median: 40.0%; range: 0.2-100%). Examples of underuse (inappropriately not performed tests) include echocardiography for heart failure (89% underuse) or atrial fibrillation (56% underuse), and pulmonary function tests for COPD (73% underuse). Examples of overuse (inaccurately performed tests) include echocardiography in people with no symptoms or signs of cardiovascular disease (77-92% overuse), urine cultures (77% overuse), upper gastrointestinal

endoscopy (37-54% overuse) and colonoscopy (52% overuse). Besides, an increase in overuse of CT and MRI scans for headaches was seen in the United States [51].

Table 2. Roles of a new test compared to an existing test

Role	Explanation	Examples
Triage	The new test is intended to be used before the existing test, and the existing test is then solely offered to patients with a specific result on the new test. The new test may have reduced accuracy compared to the existing test, but it can offer other advantages such as less burden or costs.	Screening all persons aged 55-75 years for faecal occult blood. Only those who have a positive test will receive colonoscopy
Replacement	The new test is intended to replace the existing test when it is more accurate or offers other advantages (such as reduced burden or costs) compared to the existing test.	▪ Magnetic resonance imaging (MRI) instead of mammography in women suspected of having familial breast cancer ▪ Polymerase chain reaction testing to detect herpes simplex virus instead of viral culture
Add-on	The new test is intended to be performed after the existing test, which restricts the test's application to a subset of people, for instance those who evaluate positive on the existing test. Implementing the new test may increase the accuracy of the testing pathway, but it could also have drawbacks such as increased burden and costs.	Positron emission tomography (PET) in patients with cancer after having a negative computed tomography (CT) scan for metastases
Parallel/ combined	The new test is intended to be used together with an existing test.	Determination of eGFR and albumin creatinine ratio to diagnose chronic kidney disease

Healthcare spending on laboratory diagnostics among both American and German oncologists and cardiologists was investigated by Rohr et al. They found that laboratory diagnostics accounted for 2.3% and 1.4% of healthcare spending in the United States and Germany respectively, influencing 64% and 67% of clinical decisions [52]. Incorrect testing can result in high healthcare costs, and in unnecessary test burden and anxiety [53].

Physicians acknowledge that unnecessary testing is a significant problem. Reasons for unnecessary test ordering include concerns of liability, providing reassurance, patient demands, keeping patients satisfied, and insufficient time to consult with patients. Most physicians have a sense of responsibility to prevent unnecessary testing. A majority of physicians also state that providing evidence-based recommendations in a format intended for patient communication (e.g. with icon arrays or graphs), would be effective in reducing unnecessary testing [54].

Challenges in guideline development about testing

Guideline panel members face challenges when interpreting test accuracy measures, such as sensitivity and specificity. Recalculating these measures to determine the number of true positives, true negatives, false positives and false negatives per 1000 people tested provides greater clarity, which is easier to understand [55]. Formulating key questions about testing that include people-important outcomes can be challenging as well. Moreover, there are barriers in searching and synthesizing the evidence, such as a lack of valid search filters, complex meta-analysis methods and the inclusion of outcomes beyond diagnostic accuracy. Interpreting and applying GRADE criteria for the evaluation of the clinical performance of a test can be difficult because the assessment of inconsistency and imprecision differs from the evaluation of intervention studies on clinical performance of a treatment [56]. Formulating recommendations about testing is challenging due to a lack of evidence, conflicting expert opinions, and insufficient knowledge and competencies [57]. Given the numerous challenges, it is suspected that consequences of testing on people-important outcomes are hardly considered when developing recommendations on healthcare related testing.

Aim and research questions

Developing guidelines comes with various issues, particularly when focusing on developing recommendations about testing, as described in the previous sections. There are indications from evidence and experience from guideline methodologists that the process of guideline development related to testing is suboptimal, which may lead to inaccurate consideration of the benefits and harms of testing. It is not yet known which knowledge or tools are necessary and/or helpful in appropriately developing guideline recommendations about testing.

Therefore, this thesis focuses on barriers and solutions in the development of guideline recommendations about testing, with specific attention to the required expertise for developing these recommendations and tools to facilitate this process. The aim of this thesis is to facilitate and improve guideline development concerning healthcare related testing. The first objective is to identify problems by exploring current practice and challenges in developing guidelines for healthcare related testing. The second objective is to improve this process by identifying the knowledge needed to develop testing recommendations in guidelines. The third objective is to facilitate the guideline development process by developing and testing a tool to support the formulation of appropriate guideline questions on healthcare related testing.

This has led to the following research questions:

1. What are challenges and possible solutions when assessing the certainty of evidence of a test-management pathway?
2. Which types of evidence (diagnostic accuracy, burden of the test, natural course, treatment effectiveness, link between test result and administration of treatment) are used to support guideline recommendations about testing?
3. What is the minimum knowledge required for guideline panel members involved in developing recommendations about testing?
4. Can a step-by-step guide aid guideline developers in formulating key questions about testing?

Outline of the thesis

After this introduction chapter, chapter 2 presents findings from a case study on the application of GRADE for tests and test strategies, including the identification of methodological challenges, and suggestions for solutions to these challenges (research question 1). This study evaluated the full test-management pathway for the net benefit of IgE (immunoglobulin E) in the diagnosis of allergic rhinitis. Chapter 3 presents a systematic document analysis including quality assessment of publicly available guidelines on three diagnostic tests: C-reactive protein, colonoscopy, and fractional exhaled nitric oxide. This study evaluated the incorporation of the various components of the test-management pathway in the evidence base for the guideline recommendation, including factors contributing to the comprehensiveness of the evidence as well as explanations for eventual differences between the guidelines (research question 2). Chapter 4 presents the results of a developmental study with the aim of defining the minimum knowledge required by guideline panel members who are involved in developing recommendations about testing. This study used a literature review and expert interviews to formulate a list of required knowledge components (research question 3). During the development and presentation of the required knowledge components, it became clear that practical examples of test-management pathways were needed. Chapter 5 provides detailed examples that can aid in the understanding and implementation of the test-management pathway concept. Chapter 6 presents the outcomes of developing and testing a step-by-step guide for guideline developers. The guide's objective was to assist guideline panel members in formulating key questions regarding testing (research question 4). Finally, chapter 7 offers a general discussion summarising the results of the studies, reflecting on these results, and outlining implications for practice. Additionally, it provides suggestions for further research.

References

1. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Clinical Practice Guidelines We Can Trust. In: Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US). Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011 isbn: doi:10.17226/13058.
2. Burgers JS, Van der Weijden T, Grol R. Richtlijnen als hulpmiddel bij de verbetering van de zorg. In: Wensing M, Grol R, editors. Implementatie. 7 ed. Houten: Bohn Stafleu van Loghum; 2017. p. 99-124. isbn: 978-90-368-1731-8. doi:10.1007/978-90-368-1732-5.
3. Adviesgroep Kwaliteitsstandaarden Zorginstituut Nederland. AQUA-Leidraad. Zorginstituut Nederland; 2021. Available from: https://www.zorginzicht.nl/binaries/content/assets/zorginzicht/ontwikkeltools-ontwikkelen/aqua-leidraad.pdf.
4. National Institute for Health and Care Excellence. How we develop NICE guidelines. National Institute for Health and Care Excellence; 2021. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-guidelines/how-we-develop-nice-guidelines.
5. Scottish Intercollegiate Guidelines Network. SIGN 50. A guideline developer's handbook. Scottish Intercollegiate Guidelines Network; 2019. Available from: https://www.sign.ac.uk/media/2038/sign50_2019.pdf.
6. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6. doi:10.1136/bmj.39489.470347.AD.
7. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011;64(4):395-400. doi:10.1016/j.jclinepi.2010.09.012.
8. Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. BMJ. 2016;353:i2089. doi:10.1136/bmj.i2089.
9. Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016. doi:10.1136/bmj.i2016.
10. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011;64(12):1283-93. doi:10.1016/j.jclinepi.2011.01.012.
11. Guyatt GH, Ebrahim S, Alonso-Coello P, Johnston BC, Mathioudakis AG, Briel M, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. J Clin Epidemiol. 2017;87:14-22. doi:10.1016/j.jclinepi.2017.05.005.
12. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94. doi:10.1016/j.jclinepi.2010.04.026.
13. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6. doi:10.1016/j.jclinepi.2010.07.015.
14. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407-15. doi:10.1016/j.jclinepi.2010.07.017.
15. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64(12):1277-82. doi:10.1016/j.jclinepi.2011.01.011.
16. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011;64(12):1294-302. doi:10.1016/j.jclinepi.2011.03.017.
17. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011;64(12):1303-10. doi:10.1016/j.jclinepi.2011.04.014.
18. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6. doi:10.1016/j.jclinepi.2011.06.004.

19. Brunetti M, Shemilt I, Pregno S, Vale L, Oxman AD, Lord J, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. J Clin Epidemiol. 2013;66(2):140-50. doi:10.1016/j.jclinepi.2012.04.012.
20. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2013;66(2):151-7. doi:10.1016/j.jclinepi.2012.01.006.
21. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013;66(2):158-72. doi:10.1016/j.jclinepi.2012.01.012.
22. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol. 2013;66(2):173-83. doi:10.1016/j.jclinepi.2012.08.001.
23. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-25. doi:10.1016/j.jclinepi.2012.03.013.
24. Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66(7):726-35. doi:10.1016/j.jclinepi.2013.02.003.
25. Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
26. GRADE Working Group. GRADE Working Group. Available from: https://www.gradeworkinggroup.org/#.
27. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.
28. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. Evidence-based medicine. 2008;13(6):162-3. doi:10.1136/ebm.13.6.162-a.
29. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009;64(8):1109-16. doi:10.1111/j.1398-9995.2009.02083.x.
30. Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
31. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2010;182(18):E839-42. doi:10.1503/cmaj.090449.
32. GIN-McMaster. GIN-McMaster Guideline Development Checklist (GDC). 2014. Available from: https://cebgrade.mcmaster.ca/guidelinechecklistprintable.pdf.
33. Schunemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2014;186(3):E123-42. doi:10.1503/cmaj.131237.
34. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schünemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
35. GRADE Working Group. GRADE Handbook. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013. Available from: https://gdt.gradepro.org/app/handbook/handbook.html.
36. Van Bokhoven MA, Pleunis-Van Empel MC, Koch H, Grol RP, Dinant GJ, Van der Weijden T. Why do patients want to have their blood tested? A qualitative study of patient expectations in general practice. BMC Fam Pract. 2006;7:75. doi:10.1186/1471-2296-7-75.
37. Norman G, Barraclough K, Dolovich L, Price D. Iterative diagnosis. BMJ. 2009;339:b3490. doi:10.1136/bmj.b3490.

38. Wulff HR. (eds.). Principes van klinisch denken en handelen; Nederlandse bewerking. Utrecht: Bohn, Scheltema & Holkema; 1980. isbn:90 313 0399 2.
39. Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. BMJ. 2002;324(7339):729-32. doi:10.1136/bmj.324.7339.729.
40. Hopayian K. Why medicine still needs a scientific foundation: restating the hypotheticodeductive model - part two. The British journal of general practice : the journal of the Royal College of General Practitioners. 2004;54(502):402-3; discussion 4-5.
41. Hopayian K. Why medicine still needs a scientific foundation: restating the hypotheticodeductive model - part one. The British journal of general practice : the journal of the Royal College of General Practitioners. 2004;54(502):400-1; discussion 4-5.
42. Van Leeuwen YD, Baggen JL. De medische beslissing: juist én zinnig? Huisarts en Wetenschap. 2002;45(2):66-9.
43. Stolper E. Gut feelings in general practice. Maastricht: Maastricht University; 2010.
44. Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, et al. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PLoS One. 2015;10(10):e0134553. doi:10.1371/journal.pone.0134553.
45. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001;323(7305):157-62. doi:10.1136/bmj.323.7305.157.
46. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta. 2014;427:49-57. doi:10.1016/j.cca.2013.09.018.
47. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92. doi:10.1136/bmj.332.7549.1089.
48. Mustafa RA, Wiercioch W, Cheung A, Prediger B, Brozek J, Bossuyt P, et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 2: a review of methodological and practical challenges. J Clin Epidemiol. 2017;92:18-28. doi:10.1016/j.jclinepi.2017.09.003.
49. Welch HG, Black WC. Overdiagnosis in cancer. J Natl Cancer Inst. 2010;102(9):605-13. doi:10.1093/jnci/djq099.
50. Jacobs TS, Forno E, Brehm JM, Acosta-Perez E, Han YY, Blatter J, et al. Underdiagnosis of allergic rhinitis in underserved children. Journal of Allergy & Clinical Immunology. 2014;134(3):737-9.e6. doi:10.1016/j.jaci.2014.03.028.
51. O'Sullivan JW, Albasri A, Nicholson BD, Perera R, Aronson JK, Roberts N, et al. Overtesting and undertesting in primary care: a systematic review and meta-analysis. BMJ Open. 2018;8(2):e018557. doi:10.1136/bmjopen-2017-018557.
52. Rohr UP, Binder C, Dieterle T, Giusti F, Messina CG, Toerien E, et al. The Value of In Vitro Diagnostic Testing in Medical Practice: A Status Report. PLoS One. 2016;11(3):e0149856. doi:10.1371/journal.pone.0149856.
53. Choosing wisely. Do you really need that medical test or treatment? The answer may be no. Choosing wisely; 2017. Available from: https://www.choosingwisely.org/files/Do-You-Need-That-Test_4x9-Eng.pdf.
54. PerryUndem Research/Communication. Unnecessary Tests and Procedures in the Health Care System. 2014. Available from: https://www.choosingwisely.org/files/Final-Choosing-Wisely-Survey-Report.pdf.
55. Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implement Sci. 2011;6:62. doi:10.1186/1748-5908-6-62.
56. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014;67(7):760-8. doi:10.1016/j.jclinepi.2014.01.006.
57. Gopalakrishna G, Leeflang MM, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6(9):e010549. doi:10.1136/bmjopen-2015-010549.

Chapter 2.

Applying GRADE for diagnosis revealed methodological challenges: an illustrative example for guideline developers

Mariska Tuut
Hans de Beer
Jako Burgers
Erik-Jonas van de Griendt
Trudy van der Weijden
Miranda Langendam

Journal of Clinical Epidemiology 2021; https://doi.org/10.1016/j.jclinepi.2020.11.021

Abstract

Objective: To identify challenges in the application of GRADE for diagnosis when assessing the certainty of evidence in the test-treatment strategy (diagnostic accuracy, test burden, management effectiveness, natural course, linked evidence) in an illustrative example and to propose solutions to these challenges.

Study design: A case study in applying GRADE for diagnosis that looked at the added value of IgE for diagnosing allergic rhinitis.

Results: Evaluation of the full test-treatment strategy showed a lack of (high-quality) evidence for all elements. In our example, we found a lack of evidence for test burden, natural course and link between test result and clinical management. Overall, systematically reviewing the evidence for all elements of a test-treatment strategy is more time-consuming than only considering test accuracy results and management effectiveness. To increase efficiency, the guideline panel could determine critical elements of the test-treatment strategy that need a systematic review of the evidence. For less critical elements, a guideline panel can rely on grey literature and professional expertise.

Conclusion: A lack of high-quality evidence and time investment if the full test-treatment strategy is assessed create challenges in applying GRADE for diagnosis. Discussion within guideline panels about critical elements that need to be reviewed might help.

Keywords: GRADE, diagnosis, guidelines, evidence, medical tests, systematic review

Introduction

Clinicians use tests to ascertain or reject a clinical diagnosis [1]. The clinical value of a test depends on various elements: the patient population characteristics (e.g. prevalence of the disease), test characteristics (e.g. sensitivity and specificity) and its downstream consequences on patient-relevant outcomes (e.g. test burden, natural course of the disease and management following the test results) [2]. Since direct evidence evaluating the impact of tests on patient important outcomes (diagnostic randomised trial) is scarce, different types of evidence (e.g. for diagnostic accuracy and management effectiveness) need to be assessed and linked.

Clinicians often have a limited ability to assess the value of a test in clinical practice [3, 4]. Therefore, clinical practice guidelines (CPG) have been developed to provide decision support to clinicians and patients[5]. The GRADE approach for diagnostic tests and test strategies facilitates this process by linking the elements of a test-treatment strategy and assessment of the certainty of the evidence for each element [6-8].

It is challenging to appropriately evaluate diagnostic tests (e.g. assessing the certainty of the evidence, including patient-important outcomes in evaluating test accuracy) [9, 10]. In this study, we aimed to identify the challenges of applying GRADE for diagnosis for all elements of the test-treatment strategy. We assessed the certainty in the evidence in an illustrative example and proposed solutions to overcome the barriers. This study may serve as an example for systematic reviewers and guideline developers.

Methods

Clinical question

The illustrative example is the clinical question: what is the value of specific immunoglobulin E (sIgE) blood testing as an add-on test to history taking (I) compared to history taking alone (C) in patients suspected of having allergic rhinitis (AR) in primary care (P), with relief of nasal or ocular symptoms as critical outcomes (O) [8, 11]? Concentration, sleep problems, work/school absence and quality of life (QoL) were considered important outcomes [12]. Consequences of true positive, true negative, false positive, false negative, and failed test results were discussed. We formulated PICOs for each element of the test-treatment strategy (see *table 1*).

Search strategy

Detailed methods for searching and assessing evidence for each evidence element are presented in *table 2*. We searched Medline and Embase databases to retrieve relevant evidence (*Appendix 1*). We searched for publications from 1998 to 11 January 2019

(because of sIgE-testing and non-sedating antihistamines were used since then) [12]. We used combinations of MeSH (medical subject headings) and key words and searched unrestricted to setting but limited the search to English, German or Dutch language publications.

Table 1. PICOs per sub question

Element	Patient (P)	Intervention (I)	Control (C)	Outcome (O)
Diagnostic accuracy	Patients suspected of having allergic rhinitis in primary care	sIgE-test for at least one of the allergens: ▪ Grass pollen ▪ Birch/tree pollen ▪ Herb pollen (any) ▪ House dust mite (any *Dermatophagoides*) ▪ Mould ▪ Cat epithelium ▪ Dog epithelium	Nasal provocation of allergens	▪ Accuracy measures (sensitivity, specificity); ▪ The target condition is allergic rhinitis, measured with nasal provocation (nasal challenge)
Test burden	Adults/children in general	Any venipuncture for diagnostic or screening purposes	-	Complications of testing (vasovagal reactions, pain, nerve injuries, haematoma)
Management	▪ Patients with confirmed allergic rhinitis (doctor diagnosed/ sIgE-testing/ provocation) ▪ Exclusion: self-diagnosed allergic rhinitis	▪ Allergen avoidance measures ▪ Antihistamines ▪ Nasal corticosteroids	▪ Other treatment ▪ No treatment ▪ Placebo	▪ Relief of nasal symptoms ▪ Relief of ocular symptoms ▪ Concentration ▪ Sleep problems ▪ Work/school absence ▪ Quality of life (QoL)
Natural course	▪ Patients with confirmed allergic rhinitis (doctor diagnosed/ sIgE-testing/ provocation) ▪ Exclusion: self-diagnosed allergic rhinitis	-	-	▪ Relief of nasal symptoms ▪ Relief of ocular symptoms ▪ Concentration ▪ Sleep problems ▪ Work/school absence ▪ Quality of life (QoL)
Link between test and management	Patients with a positive sIgE-test result	-	-	▪ Allergen avoidance ▪ Use of corticosteroids ▪ Use of antihistamines ▪ Compliance ▪ Treatment difficulties

Study selection, data collection and risk of bias assessment

Two authors (MT, HdB) screened abstracts and full-text articles for inclusion. Both read the full text of included studies. One reviewer (MT) completed predefined data extraction tables (*Appendix 2*) by extracting detailed information about study type, patient characteristics, methods, outcomes and risk of bias. The second reviewer (HdB) checked this process. Discrepancies were resolved by discussion.

We used study-design appropriate checklists for risk of bias assessment (*table 2*).

Table 2. Detailed methods per part of the test-treatment strategy

	Literature search (*see Appendices*) and selection eligibility criteria	Method of risk of bias/quality assessment
Diagnostic accuracy [6]	▪ Cross-sectional studies (or systematic reviews) that compare sIgE-test with nasal provocation ▪ Exclusion: case-control studies ▪ McMaster search filters for best balance of sensitivity and specificity in diagnosis [13]	QUADAS-2 [14]
Test burden	▪ Systematic reviews of at least moderate quality [15], reporting on adverse effects of venipunctures ▪ Search: MeSH with adverse events as free-floating subheading	AMSTAR-2 [15] with appraisal of risk of bias of RCTs and non-RCTs
Management	▪ Systematic reviews of at least moderate quality, consisting of RCTs (positive score on AMSTAR-2 items 4, 9, 11 and 15) ▪ McMaster search filters for best balance of sensitivity and specificity in reviews	AMSTAR-2 with appraisal of risk of bias of RCTs and non-RCTs
Natural course	Prospective cohort studies FPIN prognosis search filter [16]	Adapted QUIPS [17]
Link between test and management	Follow-up studies (reviews, scoping)	JBI Critical Appraisal Checklist for Studies Reporting Prevalence data [18]

RCT: randomised controlled trial

Data Analysis

We planned to pool results about diagnostic accuracy with RevMan 5.3. In case of substantial heterogeneity, we planned to present ranges of sensitivity and specificity. For the evidence elements 'test burden' and 'management' (avoidance measures, antihistamines, corticosteroids) we planned to calculate pooled (standardised) mean differences (MD or SMD) (in continuous outcomes) and risk ratios (RR) (in dichotomous outcomes). For the evidence elements 'natural course' and 'the link between test results and management', we used descriptive statistics (mean, ranges).

We planned subgroup analyses for age (children/adults), and type of allergen, since we expected that test characteristics and treatment effectiveness would differ in these groups. Sensitivity analysis was planned by excluding studies with a serious risk of bias.

We planned to model the different elements to patient-important outcomes if the certainty of evidence in each element was at least moderate. If the evidence was less certain, we assumed modelling was not applicable.

Certainty of evidence

For each evidence element, we prepared GRADE evidence profiles (*Appendix 3)* using GRADEpro [19]. An overall rating of confidence in estimates of effect is relevant in CPG development. It is based on the critical outcome providing the lowest confidence [20]. The overall certainty of the evidence of the test-treatment strategy was defined as the weakest link in the chain of evidence [7].

Identification of challenges and proposal of solutions

We made field notes for each methodological step lacking direct guidance on how to continue. All reported challenges were discussed between the authors, leading them to propose solutions that included a rationale for each practical/methodological choice.

Results

Consequences of test results

The consequences of test results are presented in *table 3*.

Literature search, selection and data synthesis

The yield of the literature search and selection is presented per element of the test-treatment strategy in *figure 1 (a, b)*. For the sub-question about diagnostic accuracy, three studies were included [21-23]. The search for test burden yielded no results. We included one systematic review about avoidance measures [24], two about antihistamines [25, 26], and three about corticosteroids [27-29]. We included seven cohort studies about natural course [30-36], and seven about the link between test result and management [37-43].

Study characteristics, including critical appraisal, are summarised in *Appendix 2*. GRADE evidence profiles with detailed judgements about the certainty of the evidence in the different comparisons are listed in *Appendix 3*.

Table 3. Consequences of the test results

Test result	Effects	Patient-important outcomes
True positive	True AR diagnosis, leading to targeted treatment, eventually with side effects	▪ Effective treatment reduces symptoms ▪ Level of avoidance measures ▪ Possible side effects of drug treatment ▪ Treatment costs for the patient
True negative	▪ True AR exclusion ▪ Possible non-AR diagnosis, or performance of additional diagnostic tests	▪ Persisting symptoms (unless non-AR is treated effectively) ▪ No allergen avoidance and side effects ▪ Additional testing and treatment risks
False positive	False AR diagnosis, leading to ineffective targeted treatment, eventually with side effects	▪ Persisting symptoms ▪ Potential avoidance measures ▪ Potential side effects ▪ Costs ▪ Additional testing risks
False negative	▪ False AR exclusion, no targeted treatment ▪ Follow-up tests	▪ Persisting symptoms ▪ Follow-up test risks

AR: allergic rhinitis

Clinical results and certainty of the evidence

Diagnostic test accuracy

Three studies estimated the accuracy of sIgE-testing on the house dust mite *D.pteronyssinus*, one of which also studied *D.farinae* [21-23]. Sensitivity varied from 84% to 100% and specificity from 52% to 100%. Because of heterogeneity, we did not pool results.

The certainty of evidence was very low due to serious risk of bias, indirectness and imprecision.

Test burden

We did not find evidence that fulfilled our eligibility criteria of assessing the burden of venipuncture.

Clinical management

Allergen avoidance measures

One systematic review comprising nine RCTs was included [24]. The effect estimates of the RCTs were not pooled because of clinical heterogeneity in interventions. HEPA (high-efficiency particulate air) filters were associated with a lower symptom score than placebo but were not statistically tested. Intensive bedroom cleaning combined with acaricides (a pesticide) might be related to lower symptom scores, but no absolute results were reported. The review reported inconsistent evidence about the effectiveness of allergy control bedding.

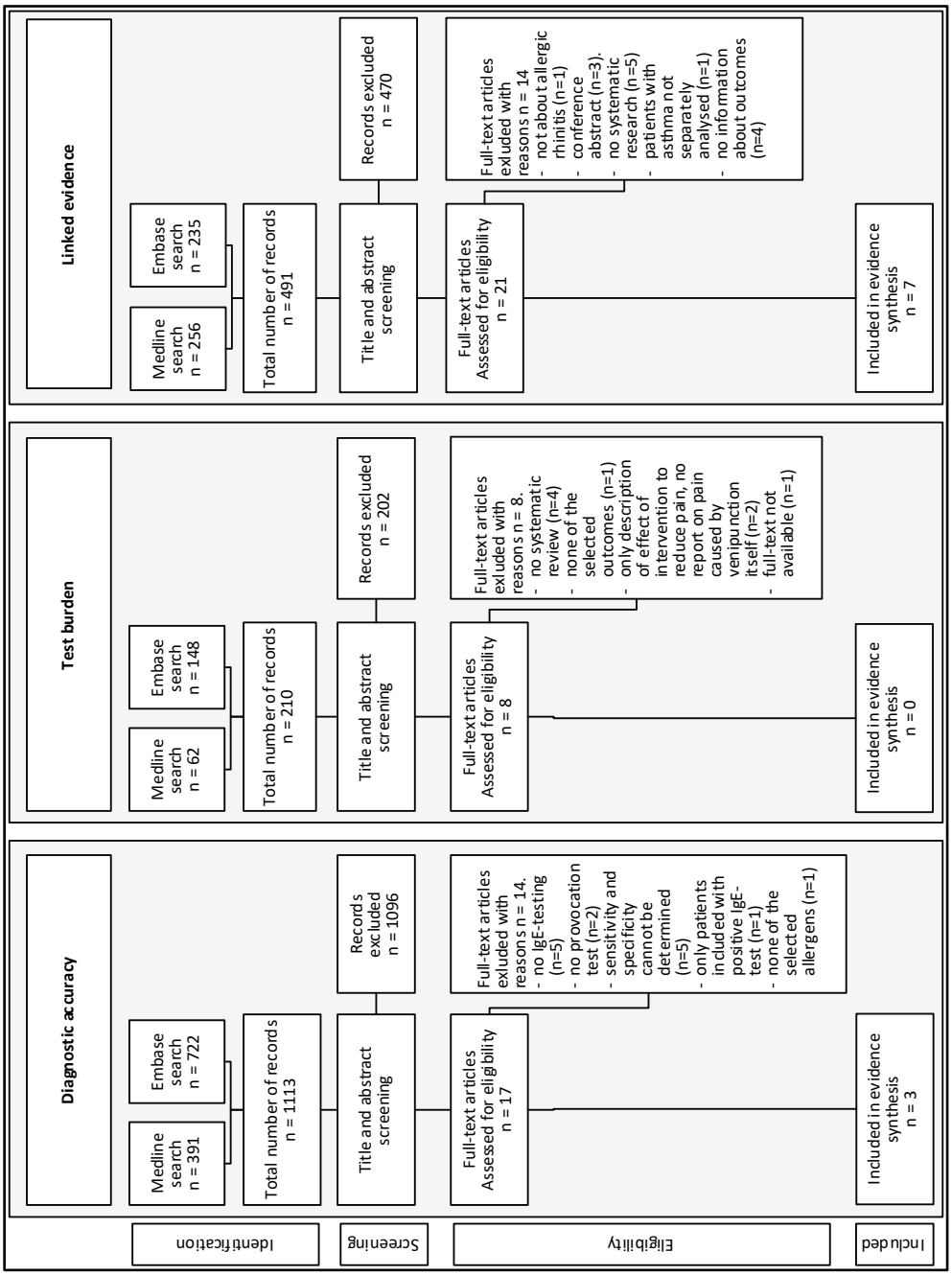


Figure 1a. Literature search and selection of the sub-questions (diagnostic accuracy, test burden, linked evidence)

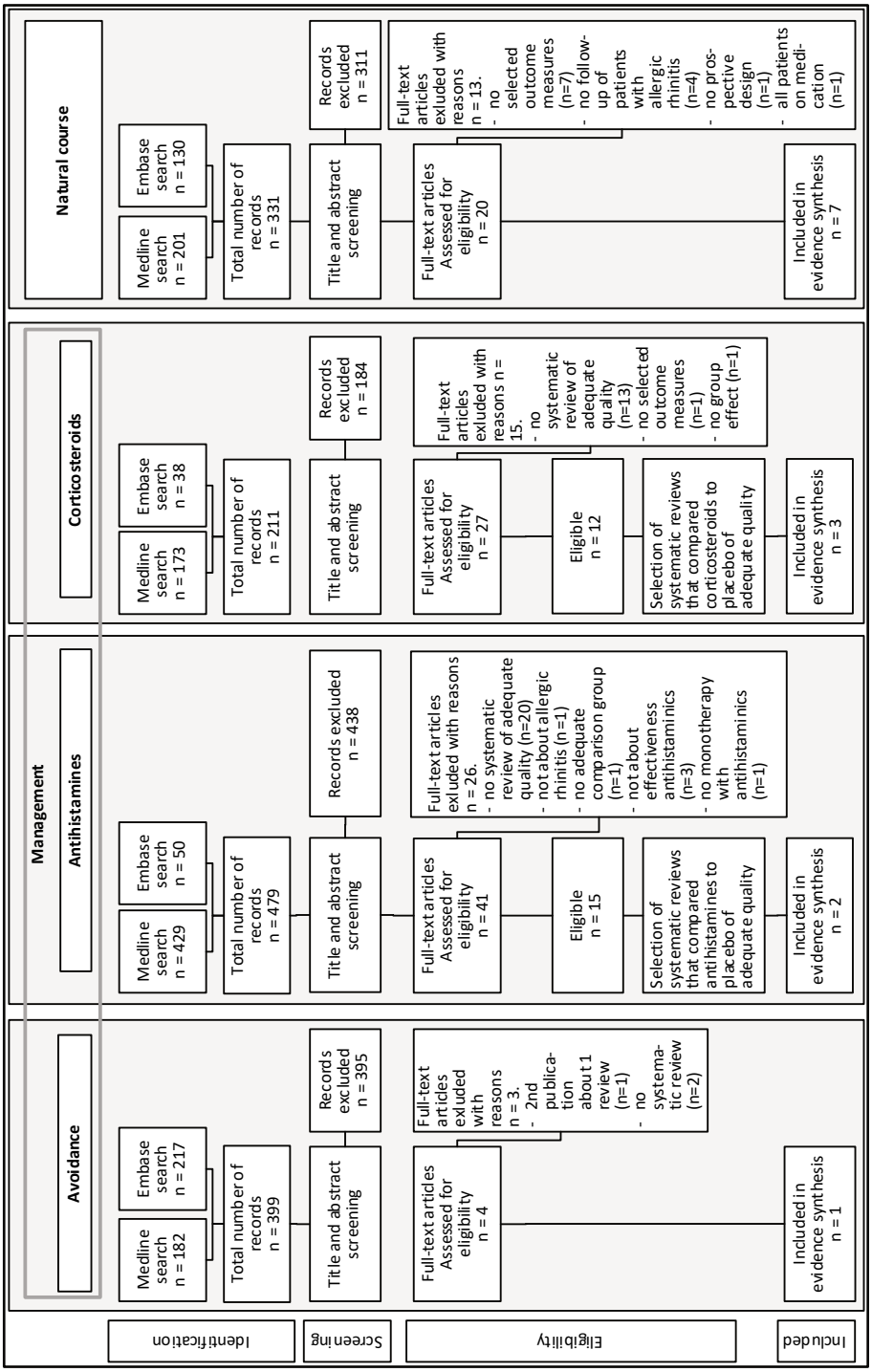


Figure 1b. Literature search and selection of the sub-questions (management, natural course)

The certainty of the evidence was very low due to very serious risk of bias and imprecision.

Antihistamines

Two systematic reviews were included [25, 26]. The first review included eight double-blind placebo-controlled RCTs in 3,532 children and adolescents with seasonal AR and compared fexofenadine (different dosages) to placebo [25]. The second review included ten double-blind placebo-controlled RCTs in 2,418 children and adults with AR (seasonal and perennial) and compared rupatadine to placebo [26]. There was moderate-quality evidence that fexofenadine had a moderate effect on total symptoms (SMD -0.42 (95%CI: -0.49 to -0.35)) and that rupatadine had a small effect on total symptoms (SMD -0.36 (95%CI: -0.48 to -0.25)) and ocular symptoms (SMD -0.29 (95%CI: -0.45 to -0.14)). No evidence was found for the other outcome measures.

The certainty of the evidence was moderate because of serious risk of bias.

Corticosteroids

Three systematic reviews about corticosteroids were included [27-29]. The first was a Cochrane review involving three placebo-controlled RCTs in 79 children with perennial AR. It compared beclomethasone dipropionate or flunisolide to placebo, without meta-analysis [27]. The second review included 16 double-blind placebo-controlled RCTs in 2,998 children and adults with seasonal or persistent AR. It compared mometasone furoate to placebo and reported symptom scores [28]. The third review included 16 double-blind placebo-controlled RCTs in 5,348 children and adults with seasonal or perennial AR [29]. It compared fluticasone furoate to placebo and ocular and reported nasal symptoms. There was moderate certainty of evidence that mometasone had a moderate effect on nasal symptoms (SMD -0.56 (95%CI: -0.71 to -0.41)) and a small effect on non-nasal symptoms (SMD -0.30 (95%CI: -0.43 to -0.18)). There was low-quality evidence that fluticasone furoate had a moderate effect on ocular symptoms in patients with seasonal AR (SMD -0.54 (95%CI: -0.70 to -0.37)) and a small effect in patients with perennial AR (SMD: -0.33 (95%CI: -0.61 to -0.05)). No evidence was found for the other pre-defined outcome measures.

Overall, the certainty of the evidence varied from low to moderate due to risk of bias and indirectness. Clinical management may result in a small to moderate reduction of symptoms of AR.

Natural course

We included seven studies reporting the natural course of AR [30-36]. All studies reported on remission, while two also described the combined outcome measure

'fewer symptoms or remission'. Follow-up varied from 2 to 23 years. Complete remission varied from 12% to 72%, while 'fewer symptoms or remission' varied from 47% to 55%.

The certainty of the evidence was very low because of (very) serious risk of bias, inconsistency (in the studies about remission), indirect evidence (since we assumed patients in the included studies might have used medication) and imprecise results.

Link between test results and management

We included seven studies reporting the link between test results and clinical management [37-43]. All studies reported about medication compliance, which we used as indirect measure for the link between a positive test result and effective management. Three studies included children [40, 41, 43], whilst the other studied adults with AR [37-39, 42].

In children, medication adherence varied from 12.5% to 70% [40, 41, 43]. In adults, self-reported good adherence varied from 28% to 87% [37-39, 42]. Weight of the medication consumed showed a lower compliance than self-reported adherence [38]. Frequent reasons for non-adherence were forgetfulness, fear of side effects, belief that medication was no longer needed and belief that the medication was ineffective [42].

The certainty of evidence was very low, mostly due to indirectness: we assumed adherence/compliance was an indirect measure of the link between test result and clinical management. Also, most studies suffered from a high risk of bias.

Overall certainty of the evidence and overall result

The overall certainty of the evidence was very low. This implies that there is a very uncertain evidence base for the value of sIgE blood testing as an add-on test to history taking compared to history taking alone in patients suspected of having AR in primary care.

Challenging issues in applying GRADE and suggested solutions

Challenging issues and our suggested solutions are tabulated in *table 4*. A further explanation per element of the test-treatment strategy is stated below.

Lack of evidence

We noticed a lack of high certainty of evidence in all elements of the test-treatment strategy. A solution might be to conduct more high-quality research.

Table 4. Challenging issues and suggested solutions

Challenging issue	Suggested solution
Lack of high-quality evidence	Conduct more and/or better research, using the GRADE downgrading factors as guidance (e.g. studies with lower risk of bias, more direct studies, larger studies to decrease imprecision) [44, 45]
Definition of comparison test	Consider the use of the current test(s) in the clinical pathway and the proposed position of the index test when added to the clinical pathway
No evidence expected in specific test burden	Expand the scope and consider including grey literature or relying on patient advocates
Multiple systematic reviews available for treatment effect	Select reviews of at least moderate quality, taking into account PICO and search date
(Very) low-quality evidence when relying on published systematic reviews about treatment	Conduct a systematic review de novo if the (very) low-quality evidence is due to indirectness
Lack of information in natural course studies in which there was no treatment or treatment is very unlikely	Downgrade for indirectness
Literature search for link between test result and clinical management	▪ Include grey literature ▪ Focus (and pre-specify) on disease-specific details (e.g. treatment adherence and treatment difficulties) ▪ Discuss within guideline panel ▪ Include qualitative research or good practice statements
Definition of overall certainty in the evidence	Base overall quality on elements of the test-treatment strategy that are critical to decision-making
Time investment	Within the guideline panel, discuss the elements for which a systematic review is relevant (i.e. the main elements that drive the decision)

Diagnostic test accuracy

In the test-treatment strategy, we were interested in the role of the sIgE-test as add-on test for clinical history. The challenge was how to define the comparison test. We decided to assess the accuracy of the index test compared to nasal provocation (nasal challenge), which is considered a gold standard by clinicians.

Test burden

We assumed not to find evidence about test burden for the specific index tests in the index population. We therefore expanded the scope to any venipuncture for diagnostic or screening purposes in adults and/or children. However, this also resulted in no aggregated evidence. A solution might be to look for and rely on grey literature or ask patient advocates or representatives, since such information might be presented in sources like textbooks rather than in scientific literature.

Clinical management

The literature search yielded multiple systematic reviews. We decided to select the most recent systematic review with at least moderate-quality evidence about a specific intervention (avoidance measures, antihistamines, corticosteroids).

Relying on published systematic reviews resulted in (very) low-quality evidence for a limited number of selected outcome measures. De novo development of systematic reviews might help in retrieving evidence that fits more precisely with the PICO.

Natural course

We were interested in the follow-up of untreated AR patients. Treatment was often not specified in the selected studies. However, it is unlikely that cohorts of people did not receive any treatment over the years. We decided to downgrade for indirectness.

Link between test result and management

It was challenging to perform a literature search to identify evidence for the link between test result and clinical management. We decided to focus on treatment adherence and treatment difficulties. Another suggestion would be to rely on expert opinion in a CPG panel (including patient representatives) to formulate good practice statements.

Overall quality

Overall quality is defined as the weakest link in the chain of evidence [7]. However, not all elements of that chain could be explicitly assessed. We suggest determining the overall certainty of evidence by considering those elements that are critical to decision-making according to a CPG panel.

Time investment and expertise of the research team

Critically and systematically appraising the evidence of the full test-treatment strategy took substantially more time than only evaluating diagnostic test accuracy, since seven PICOs had to be answered instead of only one. We propose discussing within the CPG panel the elements of the evidence chain for which a systematic review of the literature has added value. Relying on other published guidelines (e.g. treatment guidelines) may also save time.

For this study we ensured that our author team included expertise in conducting systematic reviews (including DTA), applying GRADE, guideline development and clinical management of allergic rhinitis. Including different types of expertise is needed to efficiently collect, assess, summarize and interpret the different types of evidence. We strongly recommend involvement of experienced methodologists for the development of guidelines, in particular for diagnostic recommendations.

Discussion

Summary of the main results

By applying GRADE for diagnosis, we systematically evaluated elements of a diagnostic test (i.e. diagnostic accuracy, test burden, management effectiveness, natural course and the link between test result and clinical management). During this process, we faced challenging issues and suggested solutions to resolve them. This study can therefore serve as an illustrative example for evaluating a diagnostic test in the context of CPG development, considering effectiveness as well as efficiency.

For the sIgE-test, the results suggest that it is very uncertain whether it contributes to quality of life and to reducing AR symptoms and work/school absence. The diagnostic accuracy was quite high, but with very low certainty. And the downstream consequences were very uncertain as well. A CPG panel probably would not recommend the routine use of sIgE blood testing for the diagnosis of AR in primary care.

The main challenge in assessing the overall certainty of evidence was the lack of high-quality evidence for the various elements of the test-treatment strategy. For most outcome measures of the elements in the evidence chain, we found very low-quality evidence. Another challenge was the time needed to systematically evaluate the complete pathway. Consulting CPG panel members, including patient representatives, can help save time by selecting elements of the test-treatment strategy for which a systematic review of the evidence must be carried out and others for which one can rely on other CPGs or expert opinion. If this selection process is motivated and described explicitly, the certainty of the guideline will not be affected [46].

During CPG development, a guideline panel should determine which outcome measures should be included. Panel members can also advise on the methodological approach per element of the test-treatment strategy and about the necessity of performing systematic reviews. These discussions in the CPG panel are essential to good guideline development.

Gopalakrishna and colleagues applied the GRADE approach to three Cochrane reviews to evaluate the applicability of the GRADE approach for diagnosis [9]. They found challenges in formulating the question and applying the GRADE criteria. For example, assessors in this study experienced difficulties in judging the risk of bias in relation to the QUADAS criteria. The authors also identified issues with indirectness, inconsistency, imprecision and publication bias, all related to diagnostic accuracy.

However, the authors did not go beyond diagnostic accuracy to the downstream consequences.

A similar case study was reported by Hsu and colleagues, who applied the GRADE approach to make evidence-based recommendations within a CPG panel about the diagnosis of cow milk allergy [11]. This study showed that explicitly defining patient-important outcomes as true positives and false positives beyond sensitivity and specificity was helpful to panel members with limited experience in clinical epidemiology. However, Hsu et al. did not explicitly consider downstream consequences like test burden and management effectiveness. Our study fills this gap in the knowledge.

Strengths and limitations of this study

This study evaluated the feasibility of a systematic evaluation of a test-treatment strategy in the context of guideline development. Such an evaluation has the potential to deliver an in-depth assessment of the clinical value of diagnostic test results. We performed this critical exercise by applying GRADE from the perspective of the methodologist, apart from the dynamics of a real-time CPG panel. Consulting panel members (e.g. in determining outcome measures) and discussing for which elements of the test-treatment strategy a systematic review of the literature is required would probably improve the efficiency of the in-depth assessment.

The diagnostic test discussed in this case study may have some limitations for generalisability of the study findings. Although allergic rhinitis is a very common condition, the evidence-base for the sIgE-test was limited. Applying the GRADE for diagnosis approach to conditions and tests for which more research is available might reveal other challenges. However, a limited evidence-base occurs frequently in CPG development.

Implications for practice

We suggest that CPG developers prioritise the elements of a test-treatment strategy for which a systematic review of the literature is needed. This prioritisation should take place in the planning phase of a guideline development process, in collaboration with clinical professionals and patient representatives. The outcome of this process might have consequences for the resources that are needed to develop the guideline.

Implications for research

This study unveiled methodological and planning challenges in the process of evaluating the added value of a diagnostic test in a CPG setting, and proposed

solutions to cope with these challenges. A next step is to study whether these solutions are effective and feasible in the context of a CPG panel.

Conclusion

This study identified challenges by applying the GRADE approach for diagnostic tests and test strategies to assess the certainty in the evidence in all steps of the test-treatment strategy. Important challenges were identifying evidence, drawing conclusions in the absence of high-quality evidence and investing time. An important solution is to discuss the main elements of a test-treatment strategy that drive the decision within guideline panels and to identify for which elements a systematic review is relevant and for which elements the panel can rely on other sources of information, like expert opinion.

Acknowledgement
The authors thank Martijn Sijbom PhD, Leiden University Medical Center, for his cooperation in the determination of patient-important outcome measures, and Rikie Deurenberg, SIROSS, for the literature searches.

References

1. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001;323(7305):157-62. doi:10.1136/bmj.323.7305.157.
2. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jama. 1994;271(5):389-91. doi:10.1001/jama.271.5.389.
3. Noguchi Y, Matsui K, Imura H, Kiyota M, Fukui T. Quantitative evaluation of the diagnostic thinking process in medical students. J Gen Intern Med. 2002;17(11):839-44. doi:10.1046/j.1525-1497.2002.20139.x.
4. Elstein AS. Thinking about diagnostic thinking: a 30-year perspective. Adv Health Sci Educ Theory Pract. 2009;14 Suppl 1:7-18. doi:10.1007/s10459-009-9184-0.
5. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Clinical Practice Guidelines We Can Trust. In: Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US). Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011 isbn: doi:10.17226/13058.
6. Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
7. Schunemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE Guidelines: 22. The GRADE approach for tests and strategies - from test accuracy to patient important outcomes and recommendations. J Clin Epidemiol. 2019 doi:10.1016/j.jclinepi.2019.02.003.
8. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.
9. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014;67(7):760-8. doi:10.1016/j.jclinepi.2014.01.006.
10. Gopalakrishna G, Leeflang MM, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6(9):e010549. doi:10.1136/bmjopen-2015-010549.
11. Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implement Sci. 2011;6:62. doi:10.1186/1748-5908-6-62.
12. Brozek JL, Bousquet J, Agache I, Agarwal A, Bachert C, Bosnic-Anticevich S, et al. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines-2016 revision. The Journal of allergy and clinical immunology. 2017;140(4):950-8. doi:10.1016/j.jaci.2017.03.050.
13. McMaster University. Health Information Resarch Unit. Search Filters for MEDLINE in Ovid Syntax and the PubMed translation. 2019. Available from: https://hiru.mcmaster.ca/hiru/hiru_hedges_medline_strategies.aspx.
14. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36. doi:10.7326/0003-4819-155-8-201110180-00009.
15. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008. doi:10.1136/bmj.j4008.
16. Family Physicians Inquiries Network. Prognosis Search Filter (Ovid MEDLINE). 2010. Available from: https://c.ymcdn.com/sites/fpin.org/resource/collection/9DDBC599-1567-47DD-8CA9-4BAF49BA5176/Ovid%20Prognosis%20Filter.pdf.
17. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;158(4):280-6. doi:10.7326/0003-4819-158-4-201302190-00009.
18. Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc. 2015;13(3):147-53. doi:10.1097/XEB.0000000000000054.

19. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94. doi:10.1016/j.jclinepi.2010.04.026.
20. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2013;66(2):151-7. doi:10.1016/j.jclinepi.2012.01.006.
21. Garcia Robaina JC, Sanchez Machin I, Fernandez-Caldas E, Iraola Calvo V, Vazquez Moncholi C, Bonnet Moreno C, et al. Skin tests and conjunctival and bronchial challenges with extracts of Blomia tropicalis and Dermatophagoides pteronyssinus in patients with allergic asthma and/or rhinoconjunctivitis. International Archives of Allergy & Immunology. 2003;131(3):182-8.
22. Haxel BR, Huppertz T, Boessert P, Bast F, Fruth K. Correlation of skin test results and specific immunoglobulin E blood levels with nasal provocation testing for house-dust mite allergies. American journal of rhinology & allergy. 2016;30(1):60-4. doi:10.2500/ajra.2016.30.4262.
23. King MJ, Tamulis T, Lockey RF. Prick puncture skin tests and serum specific IgE as predictors of nasal challenge response to dermatophagoides pteronyssinus in older adults. Ann Allergy Asthma Immunol. 2008;101(1):12-7. doi:10.1016/s1081-1206(10)60828-9.
24. Sheikh A, Hurwitz B, Nurmatov U, van Schayck CP. House dust mite avoidance measures for perennial allergic rhinitis. Cochrane Database of Systematic Reviews. 2010(7):CD001563. doi:10.1002/14651858.CD001563.pub3.
25. Compalati E, Baena-Cagnani R, Penagos M, Badellino H, Braido F, Gomez RM, et al. Systematic review on the efficacy of fexofenadine in seasonal allergic rhinitis: a meta-analysis of randomized, double-blind, placebo-controlled clinical trials. International Archives of Allergy & Immunology. 2011;156(1):1-15. doi:10.1159/000321896.
26. Compalati E, Canonica GW. Efficacy and safety of rupatadine for allergic rhino-conjunctivitis: a systematic review of randomized, double-blind, placebo-controlled studies with meta-analysis. Curr Med Res Opin. 2013;29(11):1539-51. doi:10.1185/03007995.2013.822855.
27. Al Sayyad JJ, Fedorowicz Z, Alhashimi D, Jamal A. Topical nasal steroids for intermittent and persistent allergic rhinitis in children. Cochrane Database of Systematic Reviews. 2007(1):CD003163. doi:10.1002/14651858.CD003163.pub4.
28. Penagos M, Compalati E, Tarantini F, Baena-Cagnani CE, Passalacqua G, Canonica GW. Efficacy of mometasone furoate nasal spray in the treatment of allergic rhinitis. Meta-analysis of randomized, double-blind, placebo-controlled, clinical trials. Allergy. 2008;63(10):1280-91. doi:10.1111/j.1398-9995.2008.01808.x.
29. Rodrigo GJ, Neffen H. Efficacy of fluticasone furoate nasal spray vs. placebo for the treatment of ocular and nasal symptoms of allergic rhinitis: a systematic review. Clinical & Experimental Allergy. 2011;41(2):160-70. doi:10.1111/j.1365-2222.2010.03654.x.
30. Di Lorenzo G, Leto-Barone MS, La Piana S, Ditta V, Di Fede G, Rini GB. Clinical course of rhinitis and changes in vivo and in vitro of allergic parameters in elderly patients: a long-term follow-up study. Clin Exp Med. 2013;13(1):67-73. doi:10.1007/s10238-012-0175-8.
31. Greisner WA, 3rd, Settipane RJ, Settipane GA. Natural history of hay fever: a 23-year follow-up of college students. Allergy & Asthma Proceedings. 1998;19(5):271-5. doi:10.2500/108854198778557728.
32. Kellberger J, Dressel H, Vogelberg C, Leupold W, Windstetter D, Weinmayr G, et al. Prediction of the incidence and persistence of allergic rhinitis in adolescence: a prospective cohort study. Journal of Allergy & Clinical Immunology. 2012;129(2):397-402.e1-3. doi:10.1016/j.jaci.2011.08.016.
33. Kong W, Chen J, Wang Y, Xiang J, Zhang X, Wang J, et al. A population-based 5-year follow-up of allergic rhinitis in Chinese children. American journal of rhinology & allergy. 2012;26(4):315-20. doi:10.2500/ajra.2012.26.3790.
34. Lee SH, Choi JH, Suh JD, Chung S, Hong SC, Kim JK, et al. Natural Course of Allergic and Nonallergic Rhinitis After 2 Years in Korean Children. Clin. 2016;9(3):233-7. doi:10.21053/ceo.2015.01130.
35. Westman M, Stjarne P, Asarnoj A, Kull I, van Hage M, Wickman M, et al. Natural course and comorbidities of allergic and nonallergic rhinitis in children. Journal of Allergy & Clinical Immunology. 2012;129(2):403-8. doi:10.1016/j.jaci.2011.09.036.
36. Yonekura S, Okamoto Y, Horiguchi S, Sakurai D, Chazono H, Hanazawa T, et al. Effects of aging on the natural history of seasonal allergic rhinitis in middle-aged subjects in South chiba, Japan. International Archives of Allergy & Immunology. 2012;157(1):73-80. doi:10.1159/000324475.

37. Koberlein J, Kothe AC, Sieber J, Mosges R. Determining factors of patient compliance to treatment in allergic rhinitis. Asian Pacific Journal of Allergy and Immunology. 2013;31(2):148-56. doi:10.12932/AP0264.31.2.2013.
38. Loh CY, Chao SS, Chan YH, Wang DY. A clinical survey on compliance in the treatment of rhinitis using nasal steroids. Allergy: European Journal of Allergy and Clinical Immunology. 2004;59(11):1168-72. doi:10.1111/j.1398-9995.2004.00554.x.
39. Navarro A, Valero A, Rosales MJ, Mullol J. Clinical use of oral antihistamines and intranasal corticosteroids in patients with allergic rhinitis. Journal of Investigational Allergology and Clinical Immunology. 2011;21(5):363-9.
40. Ocak E, Kocaoz D, Acar B. How can we improve medical adherence to intranasal corticosteroids in children? International journal of pediatric otorhinolaryngology. 2017;100:194-7. doi:10.1016/j.ijporl.2017.07.010.
41. Pizzulli A, Perna S, Florack J, Pizzulli A, Giordani P, Tripodi S, et al. The impact of telemonitoring on adherence to nasal corticosteroid treatment in children with seasonal allergic rhinoconjunctivitis. Clinical & Experimental Allergy. 2014;44(10):1246-54. doi:10.1111/cea.12386.
42. Wang K, Wang C, Xi L, Zhang Y, Ouyang Y, Lou H, et al. A randomized controlled trial to assess adherence to allergic rhinitis treatment following a daily short message service (SMS) via the mobile phone. International archives of allergy and immunology. 2014;163(1):51-8. doi:10.1159/000356317.
43. Wong IYZ, Soh SE, Chng SY, Shek LPC, Goh DYT, Van Bever HPS, et al. Compliance with topical nasal medication - An evaluation in children with rhinitis. Pediatric Allergy and Immunology. 2010;21(8):1146-50. doi:10.1111/j.1399-3038.2010.01015.x.
44. Schunemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. Inconsistency, Imprecision, publication bias and other domains for rating the certainty of evidence for test accuracy and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. 2020 doi:10.1016/j.jclinepi.2019.12.021.
45. Schunemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020 doi:10.1016/j.jclinepi.2019.12.020.
46. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2010;182(18):E839-42. doi:10.1503/cmaj.090449.

Appendix 1. Search strategies

Medline search – 11 January 2019

1 rhinitis, allergic/ or rhinitis, allergic, perennial/ or rhinitis, allergic, seasonal/ (20322)
2 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).tw. (17242)
3 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).kf. (1796)
4 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).tw. (5602)
5 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).kf. (1472)
6 or/1-5 (29270)
7 Pollen/ (17040)
8 Mites/ (10648)
9 exp Pyroglyphidae/ (2618)
10 Fungi/ (40232)
11 (aero?allergen? or pollen or mite? or mould? or mold? or fungi or cat? or dog? or (animal adj dander)).tw. (457368)
12 (aero?allergen? or pollen or mite? or mould? or mold? or fungi or cat? or dog? or (animal adj dander)).kf. (36619)
13 or/7-12 (497955)
14 exp Immunoglobulin E/an, bl, im, ip [Analysis, Blood, Immunology, Isolation & Purification] (30553)
15 ((IgE or sIgE or (immunoglobulin? adj E)) adj3 (test* or analys* or diagn* or measur* or level?)).tw. (16449)
16 ((IgE or sIgE or (immunoglobulin? adj E)) adj3 (test* or analys* or diagn* or measur* or level?)).kf. (30)
17 exp Immunoglobulin E/ (39547)
18 exp HYPERSENSITIVITY/di [Diagnosis] (54391)
19 17 and 18 (6322)
20 14 or 15 or 16 or 19 (37821)
21 sensitiv*.mp. (1561259)
22 (predictive adj3 value?).mp. (254145)
23 accurac*.tw. (355165)
24 or/21-23 (1961533)
25 6 and 20 and 24 (715)
26 (dutch or english or german).la. (25518150)
27 25 and 26 (628)
28 limit 27 to yr="1998 -Current" (396)
29 13 and 28 (264)
30 Phlebotomy/ae [Adverse Effects] (588)
31 (phlebotom* or venesec* or ven?punct*).tw. (12356)
32 (phlebotom* or venesec* or ven?punct*).kf. (659)
33 ae.fs. (1620968)
34 (pain or advers* or h??matom* or vasovagal or nerve).tw. (1355170)
35 (pain or advers* or h??matom* or vasovagal or nerve).kf. (82580)
36 pain/ or acute pain/ or pain, procedural/ (128950)
37 Hematoma/ (22330)
38 Syncope, Vasovagal/ (1800)
39 (31 or 32) and (33 or 34 or 35 or 36 or 37 or 38) (2126)
40 30 or 39 (2313)
41 meta analysis.mp,pt. (154971)
42 meta?analysis.mp,pt. (96689)
43 review.pt. (2469826)
44 search*.tw. (397366)
45 or/41-44 (2757192)
46 40 and 45 (344)
47 26 and 46 (313)
48 47 (313)
49 limit 48 to yr="1998 -Current" (261)

50 *Phlebotomy/ae (331)
51 (phlebotom* or venesec* or ven?punct*).ti. (4139)
52 32 or 50 or 51 (4579)
53 49 and 52 (53)
54 26 and 40 and 45 and 52 (62)
55 from 47 keep 1-305 (305)
56 from 54 keep 1-59 (59)
57 exp Anti-Allergic Agents/ (28999)
58 Beclomethasone/ (2962)
59 Budesonide/ (4198)
60 FLUTICASONE/ (2684)
61 Mometasone Furoate/ (699)
62 Triamcinolone/ (3704)
63 exp Fluocinolone Acetonide/ (1462)
64 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).tw. (17798)
65 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).kf. (1860)
66 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).rn. (18058)
67 exp Adrenal Cortex Hormones/ (381633)
68 Administration, Intranasal/ (13670)
69 67 and 68 (1320)
70 or/58-66 (25067)
71 69 or 70 (25687)
72 exp Histamine H1 Antagonists/ (35345)
73 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).tw. (19937)
74 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).tw. (8798)
75 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).tw. (3406)
76 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).kf. (4190)
77 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).kf. (955)
78 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).kf. (236)
79 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).rn. (7082)
80 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).rn. (9790)
81 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).rn. (2741)
82 (azelastin* or levocabastin* or olopatadin*).tw. (1194)
83 (azelastin* or levocabastin* or olopatadin*).kf. (65)
84 (azelastin* or levocabastin* or olopatadin*).rn. (922)
85 or/72-84 (48356)
86 (prevent* or avoid* or develop* or reduc* or sanit*).tw. (7305135)
87 (prevent* or avoid* or develop* or reduc* or sanit*).kf. (170682)
88 86 or 87 (7363282)
89 13 and 88 (144166)
90 6 and 26 and 71 and 45 (274)
91 90 (274)
92 limit 91 to yr="1998 -Current" (228)
93 6 and 26 and 85 and 45 (780)
94 93 (780)
95 limit 94 to yr="1998 -Current" (593)
96 6 and 26 and 89 and 45 (351)

97 96 (351)
98 limit 97 to yr="1998 -Current" (312)
99 *rhinitis, allergic/ or *rhinitis, allergic, perennial/ or *rhinitis, allergic, seasonal/ (16022)
100 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).ti. (8398)
101 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).ti. (2570)
102 3 or 5 or 99 or 100 or 101 (18433)
103 92 and 102 (173)
104 95 and 102 (428)
105 98 and 102 (182)
106 (prognos$ or outcome$ or follow-up or predict$).tw,sh. (3682599)
107 exp Prognosis/ (1475607)
108 Disease Progression/ (144837)
109 ((clinical or natural$ or disease$) adj (progress$ or course$ or histor$)).tw,sh. (221264)
110 Time Factors/ (1140587)
111 106 or 107 or 108 or 109 or 110 (5219716)
112 exp Cohort Studies/ (1812919)
113 (cohort$ or compar$ or longitudinal$ or prospective$ or multivariate or reproducib$).tw,sh. (6934341)
114 112 or 113 (7498020)
115 111 and 114 (2626014)
116 6 and 26 and 115 (3473)
117 116 (3473)
118 limit 117 to yr="1998 -Current" (2979)
119 118 and 102 (1903)
120 119 and 45 (213)
121 (prognos$ or outcome$ or follow-up or predict$).ti. (754298)
122 exp *Prognosis/ (37741)
123 *Disease Progression/ (6228)
124 ((clinical or natural$ or disease$) adj (progress$ or course$ or histor$)).ti. (29727)
125 *Time Factors/ (2184)
126 121 or 122 or 123 or 124 or 125 (808389)
127 114 and 126 (515383)
128 119 and 127 (198)
129 128 (198)

Embase search – 11 January 2019

1 allergic rhinitis/ or perennial rhinitis/ or pollen allergy/ (36718)
2 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).tw. (26427)
3 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).kw. (6196)
4 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).tw. (7178)
5 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).kw. (1309)
6 or/1-5 (45786)
7 pollen/ or grass pollen/ (19487)
8 mite/ (9977)
9 exp pyroglyphidae/ (10784)
10 exp fungus/ (459268)
11 (aero?allergen? or pollen or mite? or mould? or mold? or fungi or cat? or dog? or (animal adj dander)).tw. (487040)
12 (aero?allergen? or pollen or mite? or mould? or mold? or fungi or cat? or dog? or (animal adj dander)).kw. (55047)
13 or/7-12 (887249)
14 immunoglobulin E/ (75528)
15 ((IgE or sIgE or (immunoglobulin? adj E)) adj3 (test* or analys* or diagn* or measur* or level?)).tw. (25978)
16 ((IgE or sIgE or (immunoglobulin? adj E)) adj3 (test* or analys* or diagn* or measur* or level?)).kw. (178)
17 exp hypersensitivity/di (52556)

18 14 and 17 (8141)
19 *immunoglobulin E/ (23783)
20 15 or 16 or 18 or 19 (46360)
21 sensitiv*.tw. (1564679)
22 diagnostic accuracy.sh. (233019)
23 diagnostic.tw. (878609)
24 21 or 22 or 23 (2393725)
25 6 and 20 and 24 (1250)
26 (dutch or english or german).la. (27888273)
27 25 and 26 (1116)
28 27 (1116)
29 limit 28 to yr="1998 -Current" (875)
30 phlebotomy/ (9713)
31 blood sampling/ (183249)
32 ae.fs. (1181754)
33 (30 or 31) and 32 (7522)
34 *phlebotomy/ (2107)
35 *blood sampling/ (6214)
36 (34 or 35) and 32 (363)
37 *blood sampling/ae (101)
38 *phlebotomy/ae (155)
39 37 or 38 (251)
40 (phlebotom* or venesec* or ven?punct*).tw. (16231)
41 (phlebotom* or venesec* or ven?punct*).kw. (1647)
42 (pain or advers* or h??matom* or vasovagal or nerve).tw. (1897306)
43 (pain or advers* or h??matom* or vasovagal or nerve).kw. (221055)
44 36 or 37 or 38 (363)
45 pain/ (283829)
46 hematoma/ (55288)
47 faintness/ (17055)
48 (40 or 41) and (32 or 42 or 43 or 45 or 46 or 47) (2626)
49 meta-analysis.mp. (237521)
50 search*.tw. (498274)
51 review.pt. (2394179)
52 49 or 50 or 51 (2855956)
53 39 or 48 (2767)
54 26 and 52 and 53 (333)
55 54 (333)
56 limit 55 to yr="1998 -Current" (299)
57 39 or 41 (1895)
58 (phlebotom* or venesec* or ven?punct*).ti. (3971)
59 48 and 58 (608)
60 57 or 59 (2276)
61 52 and 60 (214)
62 61 and 26 (196)
63 62 (196)
64 limit 63 to yr="1998 -Current" (180)
65 exp antiallergic agent/ (164000)
66 beclometasone/ (7175)
67 budesonide/ (19501)
68 fluticasone/ (7674)
69 mometasone furoate/ (4545)
70 Triamcinolone/ (13109)
71 fluocinolone acetonide/ (2482)
72 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).tw. (26064)

73 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).kw. (5436)
74 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).rn. (49143)
75 exp corticosteroid/ (856535)
76 intranasal drug administration/ (13407)
77 exp corticosteroid/na [Intranasal Drug Administration] (835)
78 75 and 76 (2043)
79 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 77 (63676)
80 exp histamine H1 receptor antagonist/ (143232)
81 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).tw. (25809)
82 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).tw. (11037)
83 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).tw. (5491)
84 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).kw. (3044)
85 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).kw. (1528)
86 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).kw. (1755)
87 (azelastin* or levocabastin* or olopatadin*).tw. (1789)
88 (azelastin* or levocabastin* or olopatadin*).kw. (531)
89 (azelastin* or levocabastin* or olopatadin*).kw. (531)
90 or/80-89 (153881)
91 (prevent* or avoid* or develop* or reduc* or sanit*).tw. (9207236)
92 (prevent* or avoid* or develop* or reduc* or sanit*).kw. (271937)
93 13 and (91 or 92) and 26 and 79 and 52 (176)
94 13 and (91 or 92) and 26 and 79 (1036)
95 *allergic rhinitis/ or *perennial rhinitis/ or *pollen allergy/ (19329)
96 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).ti. (11794)
97 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).ti. (2657)
98 3 or 5 or 95 or 96 (23243)
99 3 or 5 or 95 or 96 or 97 (24366)
100 99 and 13 and (91 or 92) and 26 (2678)
101 100 (2678)
102 limit 100 to yr="1998 -Current" (2227)
103 102 and 79 and 52 (33)
104 102 and 90 and 52 (67)
105 13 and (91 or 92) and 102 and 52 (293)
106 13 and (91 or 92) and 102 and 52 (293)
107 (prognos$ or outcome$ or follow-up or predict$).tw,sh. (5411531)
108 prognosis/ or disease course/ or "prediction and forecasting"/ (945583)
109 ((clinical or natural$ or disease$) adj (progress$ or course$ or histor$)).tw,sh. (352853)
110 time factor/ (24394)
111 107 or 108 or 109 or 110 (5809609)
112 cohort analysis/ (431437)
113 (cohort$ or compar$ or longitudinal$ or prospective$ or multivariate or reproducib$).tw,sh. (8244231)
114 112 or 113 (8244231)
115 (112 or 113) and 111 (2671922)
116 6 and 26 and 115 (4026)
117 116 (4026)
118 limit 116 to yr="1998 -Current" (3818)
119 99 and 116 (1865)
120 118 and 99 (1750)
121 (prognos$ or outcome$ or follow-up or predict$).ti. (1074486)

122 *prognosis/ or *disease course/ or *"prediction and forecasting"/ (54681)
123 ((clinical or natural$ or disease$) adj (progress$ or course$ or histor$)).ti. (36474)
124 *time factor/ (702)
125 121 or 122 or 123 or 124 (1119848)
126 115 and 125 (628766)
127 99 and 118 and 125 (219)
128 126 and 127 (219)
129 limit 127 to "prognosis (best balance of sensitivity and specificity)" (92)
130 129 (92)
131 limit 99 to "prognosis (best balance of sensitivity and specificity)" (3003)
132 limit 99 to ("prognosis (best balance of sensitivity and specificity)" and yr="1998 -Current") (2581)
133 114 and 131 (1177)
134 127 (219)

Medline search linked evidence – 18 March 2019

1 rhinitis, allergic/ or rhinitis, allergic, perennial/ or rhinitis, allergic, seasonal/ (20439)
2 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).tw. (17391)
3 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).kf. (1890)
4 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).tw. (5622)
5 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).kf. (1484)
6 or/1-5 (29470)
7 Pollen/ (17172)
8 Mites/ (10716)
9 exp Pyroglyphidae/ (2651)
10 Fungi/ (40606)
11 (aero?allergen? or pollen or mite? or mould? or mold? or fungi or cat? or dog? or (animal adj dander)).tw. (461410)
12 (aero?allergen? or pollen or mite? or mould? or mold? or fungi or cat? or dog? or (animal adj dander)).kf. (37322)
13 or/7-12 (502311)
14 exp Immunoglobulin E/an, bl, im, ip [Analysis, Blood, Immunology, Isolation & Purification] (30718)
15 ((IgE or sIgE or (immunoglobulin? adj E)) adj3 (test* or analys* or diagn* or measur* or level?)).tw. (16522)
16 ((IgE or sIgE or (immunoglobulin? adj E)) adj3 (test* or analys* or diagn* or measur* or level?)).kf. (32)
17 exp Immunoglobulin E/ (39752)
18 exp HYPERSENSITIVITY/di [Diagnosis] (54814)
19 17 and 18 (6366)
20 14 or 15 or 16 or 19 (38019)
21 sensitiv*.mp. (1576062)
22 (predictive adj3 value?).mp. (256951)
23 accurac*.tw. (360755)
24 or/21-23 (1982068)
25 6 and 20 and 24 (718)
26 (dutch or english or german).la. (25708668)talen
27 25 and 26 (630)
28 limit 27 to yr="1998 -Current" (398)
29 13 and 28 (265)
30 Phlebotomy/ae [Adverse Effects] (590)
31 (phlebotom* or venesec* or ven?punct*).tw. (12420)
32 (phlebotom* or venesec* or ven?punct*).kf. (663)
33 ae.fs. (1634142)
34 (pain or advers* or h??matom* or vasovagal or nerve).tw. (1363753)
35 (pain or advers* or h??matom* or vasovagal or nerve).kf. (84197)
36 pain/ or acute pain/ or pain, procedural/ (129639)
37 Hematoma/ (22429)
38 Syncope, Vasovagal/ (1820)

39 (31 or 32) and (33 or 34 or 35 or 36 or 37 or 38) (2140)
40 30 or 39 (2328)
41 meta analysis.mp,pt. (158898)
42 meta?analysis.mp,pt. (99029)
43 review.pt. (2491081)
44 search*.tw. (404598)
45 or/41-44 (2786102)
46 40 and 45 (345)
47 26 and 46 (314)
48 47 (314)
49 limit 48 to yr="1998 -Current" (261)
50 *Phlebotomy/ae (331)
51 (phlebotom* or venesec* or ven?punct*).ti. (4145)
52 32 or 50 or 51 (4588)
53 49 and 52 (52)
54 26 and 40 and 45 and 52 (61)
55 from 47 keep 1-305 (305)
56 from 54 keep 1-59 (59)
57 exp Anti-Allergic Agents/ (29148)
58 Beclomethasone/ (2970)
59 Budesonide/ (4219)
60 FLUTICASONE/ (2715)
61 Mometasone Furoate/ (703)
62 Triamcinolone/ (3721)
63 exp Fluocinolone Acetonide/ (1467)
64 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).tw. (17846)
65 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).kf. (1884)
66 (Beclomet?ason? or budesonide or fluticason? or momet?ason? or triamcinol* or ciclesonid*).rn. (18155)
67 exp Adrenal Cortex Hormones/ (383196)
68 Administration, Intranasal/ (13793)
69 67 and 68 (1329)
70 or/58-66 (25150)
71 69 or 70 (25775)
72 exp Histamine H1 Antagonists/ (35416)
73 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).tw. (19971)
74 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).tw. (8808)
75 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).tw. (3429)
76 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).kf. (4208)
77 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).kf. (961)
78 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).kf. (244)
79 (anti?histamin* or mepyramin* or pyrilamin* or antazolin* or di??enhydramin* or carbinoxamin* or doxylamin* or clemastin* or dimenhydrinat* or pheniramin* or chlorphenamin*).rn. (7094)
80 (chlorpheniramin* or brompheniramin* or triprolidin* or hydroxyzin* or promet?azin* or cyproheptadin* or azatadin* or ketotifen* or acrivastin*).rn. (9804)
81 (cetirizin* or loratadin* or mizolastin* or fexofenadin* or levocetirizin* or desloratadin*).rn. (2764)
82 (azelastin* or levocabastin* or olopatadin*).tw. (1196)
83 (azelastin* or levocabastin* or olopatadin*).kf. (67)
84 (azelastin* or levocabastin* or olopatadin*).rn. (926)
85 or/72-84 (48461)

86 (prevent* or avoid* or develop* or reduc* or sanit*).tw. (7382854)
87 (prevent* or avoid* or develop* or reduc* or sanit*).kf. (174607)
88 86 or 87 (7441526)
89 13 and 88 (146314)
90 6 and 26 and 71 and 45 (275)
91 90 (275)
92 limit 91 to yr="1998 -Current" (229)
93 6 and 26 and 85 and 45 (785)
94 93 (785)
95 limit 94 to yr="1998 -Current" (598)
96 6 and 26 and 89 and 45 (355)
97 96 (355)
98 limit 97 to yr="1998 -Current" (316)
99 *rhinitis, allergic/ or *rhinitis, allergic, perennial/ or *rhinitis, allergic, seasonal/ (16125)
100 (rhinitis adj3 (allerg* or perennial or nonseasonal or season*)).ti. (8472)
101 ((hay adj fever) or pollinosis or nasal allerg* or nose allerg*).ti. (2577)
102 3 or 5 or 99 or 100 or 101 (18587)=P focus
103 92 and 102 (174)
104 95 and 102 (431)
105 98 and 102 (183)
106 (prognos$ or outcome$ or follow-up or predict$).tw,sh. (3727416)
107 exp Prognosis/ (1492736)
108 Disease Progression/ (146844)
109 ((clinical or natural$ or disease$) adj (progress$ or course$ or histor$)).tw,sh. (222934)
110 Time Factors/ (1145974)
111 106 or 107 or 108 or 109 or 110 (5273351)
112 exp Cohort Studies/ (1834191)
113 (cohort$ or compar$ or longitudinal$ or prospective$ or multivariate or reproducib$).tw,sh. (7008865)
114 112 or 113 (7577375)
115 111 and 114 (2659760)
116 6 and 26 and 115 (3498)
117 116 (3498)
118 limit 117 to yr="1998 -Current" (3004)
119 118 and 102 (1920)
120 119 and 45 (213)
121 (prognos$ or outcome$ or follow-up or predict$).ti. (765484)
122 exp *Prognosis/ (38248)
123 *Disease Progression/ (6400)
124 ((clinical or natural$ or disease$) adj (progress$ or course$ or histor$)).ti. (29935)
125 *Time Factors/ (2217)
126 121 or 122 or 123 or 124 or 125 (820141)
127 114 and 126 (523602)
128 119 and 127 (201)
129 128 (201)
130 limit 129 to yr="2017 -Current" (33)
131 105 (183)
132 limit 131 to yr="2017 -Current" (30)
133 104 (431)
134 limit 133 to yr="2017 -Current" (37)
135 103 (174)
136 limit 135 to yr="2017 -Current" (13)
137 54 (61)
138 limit 137 to yr="2017 -Current" (4)
139 29 (265)
140 limit 139 to yr="2017 -Current" (21)

141 6 and 89 (2876)
142 "treatment adherence and compliance"/ or "patient acceptance of health care"/ or patient compliance/ (95712)
143 (treat* adj1 (diffic* or problem*)).ti. (402)
144 141 and 142 (22)
145 "treatment adherence and compliance"/ or "patient acceptance of health care"/ or patient compliance/ or medication adherence/ (110911)
146 ((treat* adj3 adher*) or (patient adj3 (compli* or adher*)) or (patient adj3 non-adherence) or (patient adj3 non-compliance)).tw. (42361)
147 ((treat* adj3 adher*) or (patient adj3 (compli* or adher*)) or (patient adj3 non-adherence) or (patient adj3 non-compliance)).kf. (1041)
148 or/145-147 (142855) compliance
149 102 and 148 and 26 (235)
150 149 (235)
151 limit 150 to yr="1998 -Current" (202)
152 102 and 26 (15026)
153 limit 152 to yr="1998 -Current" (10007)=P focus + talen vanaf 1998
154 "filter observational studies Medline".ti. (0)
155 epidemiologic studies/ (7902)
156 exp case-control studies/ (977645)
157 exp cohort studies/ (1834191)
158 cross-sectional studies/ (288468)
159 (case adj3 control).af. (306795)
160 (cohort adj5 (study or studies or analy$)).af. (382164)
161 (follow-up adj5 (study or studies)).af. (653926)
162 (longitudinal or retrospective or prospective or (cross adj5 sectional)).af. (2108805)
163 (observational adj5 (study or studies)).af. (153887)
164 or/155-163 (2911018)
165 "filter obs Medline".ti. (0)
166 153 and 164 (2543)
167 166 and 45 (164)
168 167 not 151 (158)
169 "filter mcmaster kwalitatief".ti. (0)
170 interview.mp. (148426)
171 interview*.mp. (354511)
172 experienc*.mp. (1013629)
173 qualitative.mp. (215023)
174 qualitative research/ (44547)
175 focus groups/ or interviews as topic/ (79217)
176 (focus adj3 group?).mp. (46368)
177 or/170-176 (1419965)
178 153 and 177 (629)
179 178 not 151 (603)
180 qualitative.ti,kf. (48510)
181 interview*.ti,kf. (36820)
182 experienc*.ti,kf. (239344)
183 174 or 175 or 180 or 181 or 182 (391207)
184 (focus adj3 group?).ti,kf. (4153)
185 183 or 184 (391943)
186 153 and 185 (57)
187 186 not 151 (54)

Appendix 2. Characteristics of included studies

Diagnostic accuracy

Garcia Robaina, 2003 [1]

First author	Garcia Robaina
Year of publication	2003
Journal	International Archives of Allergy and Immunology
Setting	Tertiary care, Spain
Study design	Diagnostic accuracy study
Study population	42 patients with positive skin tests to house dust mites, 31 females, mean age 21.7 years (SD: 7.02; range 11-38). 2 patients with asthma alone, 12 with rhinoconjunctivitis alone and 28 with rhinoconjunctivitis and asthma. Previously treated with antihistamines, corticosteroids and/or β-agonists. Clinical symptoms of house dust mite allergy.
Index test	Specific IgE to D.pteronyssinus
Reference test	Conjunctival challenge: 1 drop of increasing concentrations (0,1 HEP D.pteronyssinus/ml – 100 HEP/ml or positive reaction) of extract applied to the conjunctival sac. Saline solution as control in other eye. Positive test when erythema and pruritus of conjunctiva. Evaluation 20 min, 3 and 6 hrs after test.
Performance of the index test	
True positives	31/32=97%
False positives	0/32=0%
False negatives	0/32=0%
True negatives	1/32=3%
Sensitivity	100.00% [31/31] (95%CI: 88.78-100.00) Note: Sensitivity was calculated by reviewers.
Specificity	100.00% [1/1] (95%CI: 2.50-100.00) Note: Specificity was calculated by reviewers.
Pre-test probability	31/32=97%
Risk of bias (QUADAS-2)	
Domain 1: Patient selection	
A. Risk of bias	
Describe methods of patient selection	See above
Was a consecutive or random sample of patients enrolled?	~~Yes/no/~~unclear
Was a case-control design avoided?	Yes~~/no/unclear~~ (only patients with asthma and/or rhinoconjunctivitis included
Did the study avoid inappropriate exclusions?	Yes~~/no/unclear~~
Could the selection of patients have introduced bias?	RISK: ~~LOW/HIGH/~~UNCLEAR
B. Concerns regarding applicability	
Describe included patients (prior testing, intended use of index test and setting)	Only patients with confirmed rhinoconjunctivitis or asthma were included. The IgE-test in this systematic review is intended to be used in patients suspected of having allergic rhinitis
Is there concern that the included patients do not match the review question?	CONCERN: ~~LOW/~~HIGH~~/UNCLEAR~~
Domain 2: Index test	
A. Risk of bias	
Describe the index test and how it was conducted and interpreted	See above

Were the index test results interpreted without knowledge of the results of the reference standard?	~~Yes/no/~~unclear (probably not)
If a threshold was used, was it pre-specified?	~~Yes/~~no~~/unclear~~ (no threshold)
Could the conduct or interpretation of the index test have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~
B. Concerns regarding applicability	
Is there concern that the index test, its conduct, or interpretation, differ from the review question?	CONCERN: LOW~~/HIGH/UNCLEAR~~
Domain 3: Reference standard	
A. Risk of bias	
Describe the reference standard and how it was conducted and interpreted	See above
Is the reference standard likely to correctly classify the target condition?	Yes~~/no/unclear~~
Were the reference standard results interpreted without knowledge of the results of the index test?	~~Yes/no/~~unclear (probably not)
Could the reference standard, its conduct, or its interpretation have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~
B. Concerns regarding applicability	
Is there concern that the target condition as defined by the reference standard does not match the review question?	CONCERN: LOW~~/HIGH/UNCLEAR~~
Domain 4: Flow and timing	
A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table	32 patients (out of 40 with rhinoconjunctivitis) underwent 64 conjunctival challenges (for D.pteronyssinus and B.tropicalis). No specification of patients who did not receive the reference standard.
Describe the time interval and any interventions between index test(s) and reference standard	Not specified
Was there an appropriate interval between index test(s) and reference standard?	Not specified, probably no concerns
Did all patients receive a reference standard?	~~Yes/~~no~~/unclear~~
Did patients receive the same reference standard?	Yes~~/no/unclear~~ (two types, but patients received both)
Were all patients included in the analysis?	~~Yes/~~no~~/unclear~~ Patients who did not receive the reference standard, were not included in the analysis
Could the patient flow have introduced bias?	RISK: ~~LOW/~~HIGH~~/UNCLEAR~~

Haxel, 2016 [2]

First author	Haxel
Year of publication	2016
Journal	American Journal of Rhinology & Allergy
Setting	Tertiary care, Germany
Study design	Retrospective analysis
Study population	161 patients, 60% female, mean age 35.95 yrs (sd 16.66). Clinically assumed house dust mite allergy
Index test	Specific IgE for D.pteronyssinus and D.farinae: ImmunoCAP, in venous blood; concentrations converted into 7 classes (0, 0.1–0.35 kU/L; I, 0.35–0.70 kU/L; II, 0.70–3.50 kU/L; III, 3.50–17.5 kU/L; IV, 17.5–50.0 kU/L; V, 50.0–100 kU/L; and VI, >100 kU/L Note: for this systematic review dichotomised in positive or negative; threshold 0.35 kU/L.
Reference test	Nasal provocation: symptom scores (relevant acute nasal, ocular, cutaneous, bronchial and systemic symptoms), and nasal patency impairment (rhinomanometry) after allergen provocation. Allergen provocation: D.pteronyssinus and D.farinae 10.000 units/mL. Positive result: reduction of nasal airflow >40% at 150 Pa, symptom score > 3, combination of >20% reduction in airflow and symptom score >2.
Performance of the index test *D.pteronyssinus*	
True positives	54/114=47%
False positives	23/114=20%
False negatives	10/114=9%
True negatives	27/114=24%
Sensitivity	84.38% [54/64] (95%CI: 73.14-92.24)
Specificity	54.00% [27/50] (95%CI: 39.32-68.19)
Pre-test probability	64/114=56%
Performance of the index test *D.farinae*	
True positives	36/97=37%
False positives	26/97=27%
False negatives	7/97=7%
True negatives	28/97=29%
Sensitivity	D.farinae: 83.72% [36/43] (95%CI: 69.30-93.19)
Specificity	D.farinae: 83.72% [28/54] (95%CI: 37.84-65.66)
Pre-test probability	43/97=44%
Risk of bias (QUADAS-2)	
Domain 1: Patient selection	
A. Risk of bias	
Describe methods of patient selection	Retrospective analysis of patients presenting with clinically assumed house dust mite allergy
Was a consecutive or random sample of patients enrolled?	Yes~~/no/unclear~~ Probably
Was a case-control design avoided?	Yes~~/no/unclear~~
Did the study avoid inappropriate exclusions?	~~Yes/no/~~unclear Retrospective analysis, no information about exclusions
Could the selection of patients have introduced bias?	RISK: ~~LOW/~~HIGH~~/UNCLEAR~~
B. Concerns regarding applicability	

Describe included patients (prior testing, intended use of index test and setting)	Patients presented at university clinic with clinically assumed house dust mite allergy
Is there concern that the included patients do not match the review question?	CONCERN: ~~LOW/HIGH/~~UNCLEAR This systematic review focusses on patients with suspected allergic rhinitis in primary care. It is not clear whether this is similar to the patients in this study.
Domain 2: Index test	
A. Risk of bias	
Describe the index test and how it was conducted and interpreted	See above
Were the index test results interpreted without knowledge of the results of the reference standard?	~~Yes/no/~~unclear
If a threshold was used, was it pre-specified?	~~Yes/no/~~unclear
Could the conduct or interpretation of the index test have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~
B. Concerns regarding applicability	
Is there concern that the index test, its conduct, or interpretation, differ from the review question?	CONCERN: LOW~~/HIGH/UNCLEAR~~
Domain 3: Reference standard	
A. Risk of bias	
Describe the reference standard and how it was conducted and interpreted	See above
Is the reference standard likely to correctly classify the target condition?	Yes~~/no/unclear~~
Were the reference standard results interpreted without knowledge of the results of the index test?	~~Yes/no/~~unclear
Could the reference standard, its conduct, or its interpretation have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~
B. Concerns regarding applicability	
Is there concern that the target condition as defined by the reference standard does not match the review question?	CONCERN: LOW~~/HIGH/UNCLEAR~~
Domain 4: Flow and timing	
A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table	Retrospective analysis: excluded patients were not mentioned

Describe the time interval and any interventions between index test(s) and reference standard	No timeline was described in the manuscript; probably no concerns
Was there an appropriate interval between index test(s) and reference standard?	No timeline was described in the manuscript; probably no concerns
Did all patients receive a reference standard?	Yes~~/no/unclear~~
Did patients receive the same reference standard?	Yes~~/no/unclear~~
Were all patients included in the analysis?	Yes~~/no/unclear~~
Could the patient flow have introduced bias?	RISK: ~~LOW/HIGH/~~UNCLEAR Since excluded patients were not mentioned

King, 2008 [3]

First author	King
Year of publication	2008
Journal	Annals of Allergy, Asthma & Immunology
Setting	Secondary care, USA
Study design	Diagnostic accuracy study
Study population	28 older (≥60; mean: 67.8y, 11 female), 20 younger (20-59; mean: 34.1, 14 female) patients with suspected perennial allergic rhinitis with or without allergic asthma. Perennial rhinitis was defined as symptoms during most of the year for at least 1 year. Patients were referred to hospital or allergy clinics.
Index test	Specific IgE to D.pteronyssinus in kUA/L. Positive test: sIgE ≥ 0.35 kUA/L
Reference test	Baseline peak nasal inspiratory flow (PNIF) and symptom scores. Nasal challenge with increasing concentrations D.pteronyssinus (0.1, 10, 100, 1000 AU/mL) at 30 min. intervals. PNIF and symptoms (runny nose, nasal congestion, sneezing, itchy nose, itchy/gritty eyes, runny/watery eyes, red/burning eyes, ear/palate itch) scores (scores 0-3, max. total 24). Positive test: doubling of symptom scores to a minimum postchallenge score of 10, PNIF decline ≥50%, failure to return within 20% of baseline PNIF after challenge.
Performance of the index test	
True positives	11/43=26%
False positives	5/43=12%
False negatives	2/43=5%
True negatives	25/43=58%
Sensitivity	84.62% [11/13] (95%CI: 54.55-98.08)
Specificity	83.33% [25/30] (95%CI: 65.28-94.36)
Pre-test probability	13/43=30%
Risk of bias (QUADAS-2)	
Domain 1: Patient selection	
A. Risk of bias	
Describe methods of patient selection	Suspected clinical diagnosis of perennial allergic rhinitis, with or without asthma, age 20-59 or ≥60 yrs. Symptoms during most of the year ≥1 yr. Exclusion criteria: skin disease at place of skin tests, use of β-blockers, cigarette smoking, systemic diseases, pregnancy/lactation, nude sunbathing ≥30 days, immunotherapy in the last year, antihistamines/nasal

	medication use in the last week before nasal challenge, nasal polyps, sinusitis, upper airway abnormalities
Was a consecutive or random sample of patients enrolled?	~~Yes/no/~~unclear
Was a case-control design avoided?	Yes~~/no/unclear~~
Did the study avoid inappropriate exclusions?	Yes~~/no/unclear~~
Could the selection of patients have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~
B. Concerns regarding applicability	
Describe included patients (prior testing, intended use of index test and setting)	See above
Is there concern that the included patients do not match the review question?	CONCERN: LOW/HIGH/UNCLEAR (little concern, since the restriction of the participant group because of exclusion criteria, such as non-smokers, and absence of systemic diseases; this thus not fully represent the real target population for the test)
Domain 2: Index test	
A. Risk of bias	
Describe the index test and how it was conducted and interpreted	See above
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes~~/no/unclear~~
If a threshold was used, was it pre-specified?	Yes~~/no/unclear~~
Could the conduct or interpretation of the index test have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~
B. Concerns regarding applicability	
Is there concern that the index test, its conduct, or interpretation, differ from the review question?	CONCERN: LOW~~/HIGH/UNCLEAR~~
Domain 3: Reference standard	
A. Risk of bias	
Describe the reference standard and how it was conducted and interpreted	See above
Is the reference standard likely to correctly classify the target condition?	Yes~~/no/unclear~~
Were the reference standard results interpreted without knowledge of the results of the index test?	~~Yes/~~no~~/unclear~~
Could the reference standard, its conduct, or its interpretation have introduced bias?	RISK: ~~LOW/~~HIGH~~/UNCLEAR~~ (possible difference in approach depending on test result, although the authors state that personnel has been trained to avoid this)
B. Concerns regarding applicability	
Is there concern that the target condition as defined by the reference standard does not match the review question?	CONCERN: LOW~~/HIGH/UNCLEAR~~
Domain 4: Flow and timing	
A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table	4 patients declined or had unsuccessful venipuncture. All patients received the reference standard.

Describe the time interval and any interventions between index test(s) and reference standard	The sIgE-test was done before the nasal challenge, but no time frame has been described.
Was there an appropriate interval between index test(s) and reference standard?	Probably yes
Did all patients receive a reference standard?	Yes~~/no/unclear~~
Did patients receive the same reference standard?	Yes~~/no/unclear~~
Were all patients included in the analysis?	Yes~~/no/unclear~~
Could the patient flow have introduced bias?	RISK: LOW~~/HIGH/UNCLEAR~~

Burden

No studies.

Management – allergen avoidance measures

Sheikh, 2010 [4]

Author	Sheikh
Year of publication	2010
Journal	Cochrane Database of Systematic Reviews
Study design	Systematic review of RCT's
Study population	Patients with doctor-diagnosed allergic rhinitis, and confirmed house dust mite allergy by an objective test such as skin prick testing, allergen specific-IgE concentrations or provocation testing
Description of the intervention (including dosage and duration)	House dust mite control measures: High efficiency particulate air (HEPA) filters, acaricides, barrier bedding (=allergy control bedding), barrier bedding and acaricide
Description of control group	Placebo or different house dust mite control measures
Outcomes	Primary: quality of life, sick leave, nasal symptom scores, adverse outcome. Secondary: nasal peak inspiratory flow, nasal provocation test, rhinomanometry, medication usage, compliance with treatment, drop-outs, change in house dust mite level achieved
Effect on outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	High risk of bias in included studies, due to lack of information about randomisation procedures, lack of blinding in studies, absence of intention-to-treat design, generally small numbers, high drop-out. No meta-analyses due to few trials uncovered and clinical heterogeneity, thus narrative review of results of different interventions: *HEPA filters:* Study Reisman 1990: 32 from 40 patients evaluated: aggregated rhinitis and asthma symptom scores/medication scores: lower after active filtration vs placebo: day 8.79 vs 10.38, night 8.28 vs 9.90 (no statistical testing for total score). Nasal congestion, discharge, eye irritation, and upper airway itching reduced statistically significant, whereas cough, asthma and medication use did not. *Acaricides:* Study Kniest 1991: 20 patients: symptom scores 9-12 months vs 0-3 months lower in acaricide group vs control group; no absolute symptom scores. Study Bernstein 1995: 32 children, no disaggregated symptom scores for asthma and rhinitis. Barrier bedding (=allergy control bedding): Study Moon 1999: 29 from 30 patients evaluated: Mean daily symptom scores: decreased after 4 weeks in experimental group with 2.9 vs 0.3 in control group, statistically significant. Study Terreehorst 2003: 232 patients from 279 evaluated). No significant differences in symptom scores. Study Ghazala 2004: 26 from 30

	patients completed the study: no differences in symptom scores reported between intervention and placebo. Study Brehler 2006: 21 from 32 patients completed the study. No significant reduction in symptom scores between intervention and control. *Barrier bedding and acaricides:* Study Incorvaia 2008: 25 from 29 patients evaluated: unclear difference between intervention and placebo
Risk of bias (AMSTAR-2)	
Did the research questions and inclusion criteria for the review include the components of PICO?	For yes: ☑ Population ☑ Intervention ☑ Comparator group ☑ Outcome Optional (recommended) ☐ Timeframe for follow-up ☑ Yes ☐ No
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol	For partial yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment For yes: As for partial yes, plus the protocol should be registered and should also have specified ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol ☑ Yes ☐ Partial yes ☐ No
Did the review authors explain their selection of the study designs for inclusion in the review?	For yes, the review should satisfy ONE of the following: ☐ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI ☐ Yes ☑ No
Did the review authors use a comprehensive literature search strategy?	For partial yes (all the following): ☑searched at least 2 databases (relevant to research question) ☑ provided key word and/or search strategy ☑ justified publication restrictions (e.g. language) For yes, should also have (all the following): ☑ searched the reference lists/bibliographies of included studies ☐ searched trial/study registries ☑ included/consulted content experts in the field ☐ where relevant, searched for grey literature ☑ conducted search within 24 months of completion of the review ☐ Yes ☑ Partial yes ☐ No
Did the review authors perform study selection in duplicate?	For yes, either one of the following: ☑ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include

	☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80%), with the remainder selected by one reviewer ☑ Yes ☐ No
Did the review authors perform data extraction in duplicate?	For yes, either ONE of the following: ☐ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieve good agreement (at least 80 percent), with the remainder extracted by one reviewer ☐ Yes ☑ No
Did the review authors provide a list of excluded studies and justify the exclusions?	For Partial Yes: ☑ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review For Yes, must also have: ☑ Justified the exclusion from the review of each potentially relevant study ☑ Yes ☐ Partial yes ☐ No
Did the review authors describe the included studies in adequate detail?	For Partial Yes (ALL the following): ☑ described populations ☑ described interventions ☑ described comparators ☑ described outcomes ☑ described research designs For Yes, should also have ALL the following: ☑ described population in detail ☑ described intervention in detail (including doses where relevant) ☑ described comparator in detail (including doses where relevant) ☑ described study's setting ☑ timeframe for follow-up ☑ Yes ☐ Partial yes ☐ No
Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RCTs For Partial Yes, must have assessed RoB from ☑ unconcealed allocation, and ☑ lack of blinding of patients and assessors when assessing outcomes (unnecessary for objective outcomes such as allcause mortality) For Yes, must also have assessed RoB from: ☑ allocation sequence that was not truly random, and ☑ selection of the reported result from among multiple measurements or analyses of a specified outcome ☑ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI
	NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome

	☐ Yes ☐ Partial Yes ☐ No ☑ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?	For Yes ☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information but it was not reported by study authors also qualifies ☐ Yes ☑ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	RCTs For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☐ AND investigated the causes of any heterogeneity ☐ Yes ☐ No ☑ No meta-analysis conducted
	For NRSI For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review ☐ Yes ☐ No ☑ No meta-analysis conducted
If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	For Yes: ☐ included only low risk of bias RCTs ☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect. ☐ Yes ☐ No ☑ No meta-analysis conducted
Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	For Yes: ☐ included only low risk of bias RCTs ☑ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results ☑ Yes ☐ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	For Yes: ☐ There was no significant heterogeneity in the results ☑ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review ☑ Yes ☐ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	For Yes: ☐ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias ☐ Yes ☐ No ☑ No meta-analysis conducted

Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	For Yes: ☐ The authors reported no competing interests OR ☑ The authors described their funding sources and how they managed potential conflicts of interest ☑ Yes ☐ No

Management – antihistamines

Compalati, 2011 [5]

Author	Compalati
Year of publication	2011
Journal	Int Arch Allerg Immunol
Study design	Systematic review
Study population	8 double-blind placebo-controlled randomised clinical trials including 3.532 participants with seasonal allergic rhinitis (children and adults)
Description of the intervention (including dosage and duration)	Fexofenadine in different dosages (30 mg b.i.d., 120 mg b.i.d., 120 mg o.d., 180 mg o.d.), 14 or 15 days
Description of control group	Placebo
Outcome measures	12- or 24-hour reflective total symptom scores (TSS; sum of sneezing, rhinorrhea, itchy nose/palate, itchy/watery/red eyes, excluding nasal congestion), morning instantaneous TSS, reflective individual nasal symptom scores (rhinorrhea, sneezing, itching, nasal obstruction), adverse events.
Effect on outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	12-hour reflective total symptom score: SMD: -0.42 (95%CI: -0.51 to -0.34) (5 studies, 2098 patients). 24-hour reflective total symptom score: SMD: -0.42 (95%CI: -0.49 to -0.35) (3 studies, 1434 patients). Sneezing: SMD: -0.37 (95%CI: -0.44 to -0.30) (7 studies, 3307 patients). Rhinorrhea: SMD: -0.24 (95%CI: -0.31 to -0.17) (7 studies, 3307 patients). Nasal congestion: SMD: -0.17 (95%CI: -0.24 tot -0.10) (7 studies, 3307 patients). Nasal itching: SMD: -0.27 (95%CI: -0.31 to -0.24) (7 studies, 3307 patients)
Risk of bias (AMSTAR-2)	
Did the research questions and inclusion criteria for the review include the components of PICO?	For yes: ☑ Population ☑ Intervention ☑ Comparator group ☑ Outcome Optional (recommended) ☐ Timeframe for follow-up ☑ Yes ☐ No
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol	For partial yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment For yes: As for partial yes, plus the protocol should be registered and should also have specified ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol ☐ Yes ☐ Partial yes ☑ No

Did the review authors explain their selection of the study designs for inclusion in the review?	For yes, the review should satisfy ONE of the following: ☑ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI ☑ Yes ☐ No
Did the review authors use a comprehensive literature search strategy?	For partial yes (all the following): ☑ searched at least 2 databases (relevant to research question) ☑ provided key word and/or search strategy ☑ justified publication restrictions (e.g. language) For yes, should also have (all the following): ☐ searched the reference lists/bibliographies of included studies ☐ searched trial/study registries ☐ included/consulted content experts in the field ☐ where relevant, searched for grey literature ☑ conducted search within 24 months of completion of the review ☐ Yes ☑ Partial yes ☐ No
Did the review authors perform study selection in duplicate?	For yes, either one of the following: ☑ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include ☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80%), with the remainder selected by one reviewer ☑ Yes ☐ No
Did the review authors perform data extraction in duplicate?	For yes, either ONE of the following: ☑ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieve good agreement (at least 80 percent), with the remainder extracted by one reviewer ☑ Yes ☐ No
Did the review authors provide a list of excluded studies and justify the exclusions?	For Partial Yes: ☑ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review For Yes, must also have: ☐ Justified the exclusion from the review of each potentially relevant study ☐ Yes ☑Partial yes ☐ No
Did the review authors describe the included studies in adequate detail?	For Partial Yes (ALL the following): ☑ described populations ☑ described interventions ☑ described comparators ☑ described outcomes ☑ described research designs For Yes, should also have ALL the following: ☐ described population in detail ☑ described intervention in detail (including doses where relevant) ☑ described comparator in detail (including doses where relevant) ☐ described study's setting ☑ timeframe for follow-up ☐ Yes ☑ Partial yes ☐ No

Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RCTs For Partial Yes, must have assessed RoB from ☑ unconcealed allocation, and ☑ lack of blinding of patients and assessors when assessing outcomes (unnecessary for objective outcomes such as allcause mortality) For Yes, must also have assessed RoB from: ☑ allocation sequence that was not truly random, and ☑ selection of the reported result from among multiple measurements or analyses of a specified outcome ☑ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI
	NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome ☐ Yes ☐ Partial Yes ☐ No ☑ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?	For Yes ☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information but it was not reported by study authors also qualifies ☐ Yes ☑ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	RCTs For Yes: ☑ The authors justified combining the data in a meta-analysis ☑ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☑ AND investigated the causes of any heterogeneity ☑ Yes ☑ No ☐ No meta-analysis conducted
	For NRSI For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review ☐ Yes ☐ No ☑ No meta-analysis conducted
If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	For Yes: ☐ included only low risk of bias RCTs ☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect. ☐ Yes ☑ No ☐ No meta-analysis conducted

Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	For Yes: ☐ included only low risk of bias RCTs ☑ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results ☑ Yes ☐ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	For Yes: ☑ There was no significant heterogeneity in the results ☐ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review ☑ Yes ☐ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	For Yes: ☐ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias ☐ Yes ☑ No ☐ No meta-analysis conducted
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	For Yes: ☐The authors reported no competing interests OR ☐ The authors described their funding sources and how they managed potential conflicts of interest ☐ Yes ☑ No

Compalati, 2013 [6]

Author	**Compalati**
Year of publication	**2013**
Journal	Current Medical Research & Opinion
Study design	Systematic review
Study population	10 double-blind placebo-controlled randomised controlled trials involving 2.418 children and adults with allergic rhinitis (persistent/intermittent, seasonal/perennial)
Description of the intervention (including dosage and duration)	Rupatidine 10 mg, Rupatidine 20 mg, Rupatidine oral solution 2,5-5 mg
Description of control group	Placebo
Outcome measures	Overall allergy symptoms score (sum of sneezing, rhinorrhea, itchy nose/palate, itchy/watery/red eyes, nasal congestion), total nasal symptoms, individual nasal and ocular symptoms, patient's satisfaction, frequency of adverse events
Effect on outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Overall allergy symptoms score reduction: SMD: -0.37 (95%CI: -0.46 to -0.27) (8 studies, 1650 patients). Total nasal symptom reduction: SMD: -0.36 (95%CI: -0.48 to -0.25) (7 studies, 1178 patients). Rhinorrhea: SMD: -0.30 (95%CI: -0.41 to -0.19) (7 studies, 1282 patients). Sneezing: SMD: -0.39 (95%CI: -0.52 to -0.26) (6 studies, 932 patients). Nasal obstruction: SMD: -0.25 (95%CI: -0.37 to -0.13) (5 studies, 982 patients). Nasal itching: SMD: -0.21 (95%CI: -0.33 to -0.10) (6 studies, 1178 patients). Itchy eyes: SMD: -0.29 (95%CI: -0.45 to -0.14) (4 studies, 683 patients). Watery eyes: SMD: -0.25 (95%CI: -0.45 to -0.06) (2 studies, 399 patients)
Risk of bias (AMSTAR-2)	
Did the research questions and inclusion criteria for the review include the components of PICO?	For yes: ☑ Population ☑ Intervention

	☑ Comparator group ☑ Outcome Optional (recommended) ☐ Timeframe for follow-up ☑ Yes ☐ No
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol	For partial yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment For yes: As for partial yes, plus the protocol should be registered and should also have specified ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol ☐ Yes ☐ Partial yes ☑ No
Did the review authors explain their selection of the study designs for inclusion in the review?	For yes, the review should satisfy ONE of the following: ☑ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI ☑ Yes ☐ No
Did the review authors use a comprehensive literature search strategy?	For partial yes (all the following): ☑ searched at least 2 databases (relevant to research question) ☑ provided key word and/or search strategy ☑ justified publication restrictions (e.g. language) For yes, should also have (all the following): ☑ searched the reference lists/bibliographies of included studies ☑ searched trial/study registries ☐ included/consulted content experts in the field ☐ where relevant, searched for grey literature ☑ conducted search within 24 months of completion of the review ☐ Yes ☑ Partial yes ☐ No
Did the review authors perform study selection in duplicate?	For yes, either one of the following: ☑ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include ☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80%), with the remainder selected by one reviewer ☑ Yes ☐ No
Did the review authors perform data extraction in duplicate?	For yes, either ONE of the following: ☑ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieve good agreement (at least 80 percent), with the remainder extracted by one reviewer ☑ Yes ☐ No

Did the review authors provide a list of excluded studies and justify the exclusions?	For Partial Yes: ☑ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review For Yes, must also have: ☐ Justified the exclusion from the review of each potentially relevant study ☐ Yes ☑Partial yes ☐ No
Did the review authors describe the included studies in adequate detail?	For Partial Yes (ALL the following): ☑ described populations ☑ described interventions ☑ described comparators ☑ described outcomes ☑ described research designs For Yes, should also have ALL the following: ☑ described population in detail ☑ described intervention in detail (including doses where relevant) ☑ described comparator in detail (including doses where relevant) ☐ described study's setting ☑ timeframe for follow-up ☐ Yes ☑ Partial yes ☐ No
Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RCTs For Partial Yes, must have assessed RoB from ☑ unconcealed allocation, and ☑ lack of blinding of patients and assessors when assessing outcomes (unnecessary for objective outcomes such as allcause mortality) For Yes, must also have assessed RoB from: ☑ allocation sequence that was not truly random, and ☑ selection of the reported result from among multiple measurements or analyses of a specified outcome ☑ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI
	NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome ☐ Yes ☐ Partial Yes ☐ No ☑ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?	For Yes ☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information but it was not reported by study authors also qualifies ☐ Yes ☑ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	RCTs For Yes: ☑ The authors justified combining the data in a meta-analysis ☑ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☑ AND investigated the causes of any heterogeneity ☑ Yes ☐ No ☐ No meta-analysis conducted

	For NRSI For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review ☐ Yes ☐ No ☑ No meta-analysis conducted
If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	For Yes: ☐ included only low risk of bias RCTs ☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect. ☐ Yes ☑No ☐ No meta-analysis conducted
Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	For Yes: ☐ included only low risk of bias RCTs ☑ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results ☑ Yes ☐ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	For Yes: ☐ There was no significant heterogeneity in the results ☑ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review ☑ Yes ☐ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	For Yes: ☑ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias ☑ Yes ☑ No ☐ No meta-analysis conducted
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	For Yes: ☑The authors reported no competing interests OR ☐ The authors described their funding sources and how they managed potential conflicts of interest ☑ Yes ☐ No

Management – corticosteroids

Al Sayyad, 2007 [7]

Author	**Al Sayyad**
Year of publication	**2007**
Journal	Cochrane Database of Systematic Reviews
Study design	Cochrane systematic review
Study population	3 studies including 79 children with perennial allergic rhinitis

Description of the intervention (including dosage and duration)	Different interventions: beclomethasone dipropionate aerosol spray 300 mg/day, beclomethasone dipropionate inhaled 50 mg in each nostril four times a day, flunisolide nasal spray in aqueous propylene glycol solution 0.025% 0.1ml per actuation. One spray in each nostril 3 times a day
Description of control group	Placebo
Outcome measures	Improvement of global symptoms, individual symptom scores which included any appropriate measures of nasal obstruction, runny nose, sneezing, itching, eye symptoms (including parent rated symptom scores). Nasal assessment scores of inspiratory peak flow levels. Assessment of allergen sensitivity in either the eye or nose. Measurement of IgE antibodies. Quality of life instruments to measure: performance at school, absenteeism, social behaviour, emotional well-being, social relationships. Adverse effects
Effect on outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	No meta-analysis because of the scarcity and poor quality of the data. No quantitative results.
Risk of bias (AMSTAR-2)	
Did the research questions and inclusion criteria for the review include the components of PICO?	For yes: ☑ Population ☑ Intervention ☑ Comparator group ☑ Outcome Optional (recommended) ☐ Timeframe for follow-up ☑ Yes ☐ No
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol	For partial yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment For yes: As for partial yes, plus the protocol should be registered and should also have specified ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol ☐ Yes ☐ Partial yes ☑ No
Did the review authors explain their selection of the study designs for inclusion in the review?	For yes, the review should satisfy ONE of the following: ☐ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI ☐ Yes ☑ No
Did the review authors use a comprehensive literature search strategy?	For partial yes (all the following): ☑ searched at least 2 databases (relevant to research question) ☑ provided key word and/or search strategy

	☑ justified publication restrictions (e.g. language) For yes, should also have (all the following): ☑ searched the reference lists/bibliographies of included studies ☑ searched trial/study registries ☐ included/consulted content experts in the field ☐ where relevant, searched for grey literature ☑ conducted search within 24 months of completion of the review ☐ Yes ☑ Partial yes ☐ No
Did the review authors perform study selection in duplicate?	For yes, either one of the following: ☑ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include ☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80%), with the remainder selected by one reviewer ☑ Yes ☐ No
Did the review authors perform data extraction in duplicate?	For yes, either ONE of the following: ☑ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieve good agreement (at least 80 percent), with the remainder extracted by one reviewer ☑ Yes ☐ No
Did the review authors provide a list of excluded studies and justify the exclusions?	For Partial Yes: ☑ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review For Yes, must also have: ☑ Justified the exclusion from the review of each potentially relevant study ☑ Yes ☐ Partial yes ☐ No
Did the review authors describe the included studies in adequate detail?	For Partial Yes (ALL the following): ☑ described populations ☑ described interventions ☑ described comparators ☑ described outcomes ☑ described research designs For Yes, should also have ALL the following: ☑ described population in detail ☑ described intervention in detail (including doses where relevant) ☑ described comparator in detail (including doses where relevant) ☐ described study's setting ☑ timeframe for follow-up ☐ Yes ☑ Partial yes ☐ No
Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RCTs For Partial Yes, must have assessed RoB from ☑ unconcealed allocation, and ☑ lack of blinding of patients and assessors when

	assessing outcomes (unnecessary for objective outcomes such as allcause mortality) For Yes, must also have assessed RoB from: ☑ allocation sequence that was not truly random, and ☑ selection of the reported result from among multiple measurements or analyses of a specified outcome ☑ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI
	NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome ☐ Yes ☐ Partial Yes ☐ No ☑ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?	For Yes ☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information but it was not reported by study authors also qualifies ☐ Yes ☑ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	RCTs For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☐ AND investigated the causes of any heterogeneity ☐ Yes ☐ No ☑ No meta-analysis conducted
	For NRSI For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review ☐ Yes ☐ No ☑ No meta-analysis conducted

If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	For Yes: ☑ included only low risk of bias RCTs ☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect. ☐ Yes ☐ No ☑ No meta-analysis conducted
Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	For Yes: ☑ included only low risk of bias RCTs ☐ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results ☐ Yes ☑ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	For Yes: ☐ There was no significant heterogeneity in the results ☐ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review ☐ Yes ☑ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	For Yes: ☐ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias ☐ Yes ☐ No ☑ No meta-analysis conducted
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	For Yes: ☑The authors reported no competing interests OR ☐ The authors described their funding sources and how they managed potential conflicts of interest ☑ Yes ☐ No

Penagos, 2008 [8]

Author	Penagos
Year of publication	2008
Journal	Allergy
Study design	Systematic review
Study population	16 double-blind RCTs including 2998 children and adults with seasonal or persistent allergic rhinitis
Description of the intervention (including dosage and duration)	Mometaosone furoate nasal spray 100 or 200 µg
Description of control group	Compared to placebo for this review
Outcome measures	Total nasal symptom score (TNSS), nasal individual symptom scores (congestion, rhinorrhoea, sneezing, nasal itching), non-nasal symptom scores (ocular, otic, palate and throat complaints, cough, etc.), nasal airflow, adverse events frequency
Effect on outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Total nasal symptom score (TNSS): SMD: -0.56 (95%CI: -0.71 to -0.41) (10 studies in adults, 1878 patients); sub-analysis due to heterogeneity, of studies that assessed the post-challenge effect on TNSS: SMD: -0.33 (95%CI: -0.50 to -0.17). 1 study in children: SMD: -0.41 (95%CI: -0.65 to -0.17) (n=271). Individual nasal symptom scores:

	Nasal stuffiness/congestion: SMD: -0.41 (95%CI: -0.56 to -0.27) (7 studies, 1582 patients); significant heterogeneity. Rhinorrhoea: SMD: -0.44 (95%CI: -0.66 to -0.21) (7 studies, 1582 patients); significant heterogeneity. Sneezing: SMD: -0.40 (95%CI: -0.57 to 0.23) (7 studies, 1582 patients); significant heterogeneity. Nasal itching: SMD: -0.39 (95%CI: -0.53 to -0.25) (7 studies, 1582 patients). Non-nasal symptom scores: SMD: -0.30 (95%CI: -0.43 to -0.18) (4 studies, 1009 patients)
Risk of bias (AMSTAR-2)	
Did the research questions and inclusion criteria for the review include the components of PICO?	For yes: ☑ Population ☑ Intervention ☐ Comparator group ☐ Outcome Optional (recommended) ☐ Timeframe for follow-up ☐ Yes ☑ No
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol	For partial yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment For yes: As for partial yes, plus the protocol should be registered and should also have specified ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol ☐ Yes ☐ Partial yes ☑ No
Did the review authors explain their selection of the study designs for inclusion in the review?	For yes, the review should satisfy ONE of the following: ☐ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI ☐ Yes ☑ No
Did the review authors use a comprehensive literature search strategy?	For partial yes (all the following): ☑ searched at least 2 databases (relevant to research question) ☑ provided key word and/or search strategy ☑ justified publication restrictions (e.g. language) For yes, should also have (all the following): ☑ searched the reference lists/bibliographies of included studies ☐ searched trial/study registries ☐ included/consulted content experts in the field ☐ where relevant, searched for grey literature ☑ conducted search within 24 months of completion of the review ☐ Yes ☑ Partial yes ☐ No
Did the review authors perform study selection in duplicate?	For yes, either one of the following: ☑ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include ☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80%), with the remainder selected by one reviewer

	☑ Yes ☐ No
Did the review authors perform data extraction in duplicate?	For yes, either ONE of the following: ☑ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieve good agreement (at least 80 percent), with the remainder extracted by one reviewer ☑ Yes ☐ No
Did the review authors provide a list of excluded studies and justify the exclusions?	For Partial Yes: ☐ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review For Yes, must also have: ☐ Justified the exclusion from the review of each potentially relevant study ☐ Yes ☐ Partial yes ☑ No
Did the review authors describe the included studies in adequate detail?	For Partial Yes (ALL the following): ☑ described populations ☑ described interventions ☑ described comparators ☐ described outcomes ☑ described research designs For Yes, should also have ALL the following: ☐ described population in detail ☑ described intervention in detail (including doses where relevant) ☑ described comparator in detail (including doses where relevant) ☐ described study's setting ☑ timeframe for follow-up ☐ Yes ☐ Partial yes ☑ No
Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RCTs For Partial Yes, must have assessed RoB from ☑ unconcealed allocation, and ☑ lack of blinding of patients and assessors when assessing outcomes (unnecessary for objective outcomes such as allcause mortality) For Yes, must also have assessed RoB from: ☑ allocation sequence that was not truly random, and ☑ selection of the reported result from among multiple measurements or analyses of a specified outcome ☑ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI
	NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome ☐ Yes ☐ Partial Yes ☐ No ☑ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?	For Yes

	☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information but it was not reported by study authors also qualifies ☐ Yes ☑ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	RCTs For Yes: ☑ The authors justified combining the data in a meta-analysis ☑ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☑ AND investigated the causes of any heterogeneity ☑ Yes ☐ No ☐ No meta-analysis conducted
	For NRSI For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review ☐ Yes ☐ No ☑ No meta-analysis conducted
If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	For Yes: ☑ included only low risk of bias RCTs ☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect. ☑ Yes ☐ No ☐ No meta-analysis conducted
Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	For Yes: ☑ included only low risk of bias RCTs ☐ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results ☑ Yes ☐ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	For Yes: ☐ There was no significant heterogeneity in the results ☑ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review ☑ Yes ☐ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	For Yes: ☑ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias ☑ Yes ☐ No ☐ No meta-analysis conducted
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	For Yes: ☐The authors reported no competing interests OR ☐ The authors described their funding sources and how they managed potential conflicts of interest ☐ Yes ☑ No

Rodrigo, 2011 [9]

Author	**Rodrigo**
Year of publication	**2011**
Journal	Clinical & Experimental Allergy
Study design	Systematic review
Study population	16 double-blind placebo-controlled RCTs involving 5348 patients with allergic rhinitis. Of them, 7 studies with 2589 patients with seasonal allergic rhinitis, and 9 studies with 2759 patients with perennial allergic rhinitis. 13 studies were carried out in adults and children > 12 years. 3 studies were carried out in children.
Description of the intervention (including dosage and duration)	Fluticasone furoate nasal spray 110 µg once daily
Description of control group	Placebo
Outcome measures	Mean change in daily reflective total ocular symptom score (rTOSS), mean change in AM pre-dose instantaneous total ocular symptom score (iTOSS), mean change in daily reflective nasal symptom score (rTNSS), mean change in AM pre-dose instantaneous total nasal symptom score (iTNSS), individual nasal and ocular symptoms, overall evaluation of response to therapy, quality of life (rhinoconjunctivitis quality of life questionnaire (RQLQ)), adverse events
Effect on outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Mean change in daily reflective total ocular symptom score (rTOSS): Seasonal allergic rhinitis: MD: -0.54 (95%CI: -0.70 to -0.37) (6 studies, 2219 patients). Perennial allergic rhinitis: MD: -0.33 (95%CI: -0.61 to -0.05) (3 studies, 919 patients). Eye itching/burning: Seasonal allergic rhinitis: MD: -0.20 (95%CI: -0.29 to -0.11) (3 studies, 886 patients). Perennial allergic rhinitis: MD: -0.14 (95%CI: -0.27 to -0.01) (2 studies, 604 patients). Eye tearing/watering: Seasonal allergic rhinitis: MD: -0.22 (95%CI: -0.31 to -0.13) (3 studies, 886 patients). Perennial allergic rhinitis: MD: -0.11 (95%CI: -0.21 to -0.01) (2 studies, 604 patients). Eye redness: Seasonal allergic rhinitis: MD: -0.21 (95%CI: -0.30 to -0.12) (3 studies, 886 patients). Perennial allergic rhinitis: MD: -0.11 (95%CI: -0.19 to -0.09) (2 studies, 604 patients). Mean change in daily reflective total nasal symptom score (rTNSS): Seasonal allergic rhinitis: MD: -1.14 (95%CI: -1.57 to -0.72) (6 studies, 2589 patients). Perennial allergic rhinitis: MD: -0.83 (95%CI: -1.08 to -0.59) (7 studies, 2539 patients). Rhinorrhea: Seasonal allergic rhinitis: MD: -0.35 (95%CI: -0.48 to -0.22) (4 studies, 1141 patients). Perennial allergic rhinitis: MD: -0.20 (95%CI: -0.32 to -0.07) (4 studies, 1054 patients). Nasal congestion: Seasonal allergic rhinitis: MD: -0.31 (95%CI: -0.40 to -0.23) (4 studies, 1141 patients). Perennial allergic rhinitis: MD: -0.16 (95%CI: -0.24 to -0.09) (4 studies, 1054 patients). Nasal itching: Seasonal allergic rhinitis: MD: -0.31 (95%CI: -0.39 to -0.22) (4 studies, 1141 patients). Perennial allergic rhinitis: MD: -0.22 (95%CI: -0.30 to -0.14) (4 studies, 1054 patients). Sneezing: Seasonal allergic rhinitis: MD: -0.39 (95%CI: -0.48 to -0.31) (4 studies, 1141 patients). Perennial allergic rhinitis: MD: -0.25 (95%CI: -0.32 to -0.18) (4 studies, 1054 patients)
Risk of bias (AMSTAR-2)	
Did the research questions and inclusion criteria for the review include the components of PICO?	For yes: ☑ Population ☑ Intervention ☑ Comparator group ☑ Outcome Optional (recommended)

	☐ Timeframe for follow-up ☑ Yes ☐ No
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol	For partial yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment For yes: As for partial yes, plus the protocol should be registered and should also have specified ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol ☐ Yes ☐ Partial yes ☑ No
Did the review authors explain their selection of the study designs for inclusion in the review?	For yes, the review should satisfy ONE of the following: ☐ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI ☐ Yes ☑ No
Did the review authors use a comprehensive literature search strategy?	For partial yes (all the following): ☑ searched at least 2 databases (relevant to research question) ☑ provided key word and/or search strategy ☑ justified publication restrictions (e.g. language) For yes, should also have (all the following): ☐ searched the reference lists/bibliographies of included studies ☑ searched trial/study registries ☐ included/consulted content experts in the field ☐ where relevant, searched for grey literature ☑ conducted search within 24 months of completion of the review ☐ Yes ☑ Partial yes ☐ No
Did the review authors perform study selection in duplicate?	For yes, either one of the following: ☑ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include ☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80%), with the remainder selected by one reviewer ☑ Yes ☐ No
Did the review authors perform data extraction in duplicate?	For yes, either ONE of the following: ☑ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieve good agreement (at least 80 percent), with the remainder extracted by one reviewer ☑ Yes ☐ No
Did the review authors provide a list of excluded studies and justify the exclusions?	For Partial Yes: ☐ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review For Yes, must also have:

	☐ Justified the exclusion from the review of each potentially relevant study ☐ Yes ☐ Partial yes ☑ No
Did the review authors describe the included studies in adequate detail?	For Partial Yes (ALL the following): ☑ described populations ☑ described interventions ☑ described comparators ☑ described outcomes ☑ described research designs For Yes, should also have ALL the following: ☑ described population in detail ☑ described intervention in detail (including doses where relevant) ☑ described comparator in detail (including doses where relevant) ☐ described study's setting ☑ timeframe for follow-up ☐ Yes ☑ Partial yes ☐ No
Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RCTs For Partial Yes, must have assessed RoB from ☑ unconcealed allocation, and ☑ lack of blinding of patients and assessors when assessing outcomes (unnecessary for objective outcomes such as allcause mortality) For Yes, must also have assessed RoB from: ☑ allocation sequence that was not truly random, and ☑ selection of the reported result from among multiple measurements or analyses of a specified outcome ☑ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI
	NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome ☐ Yes ☐ Partial Yes ☐ No ☑ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?	For Yes ☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information but it was not reported by study authors also qualifies ☐ Yes ☑ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	RCTs For Yes: ☑ The authors justified combining the data in a meta-analysis ☑ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☑ AND investigated the causes of any heterogeneity ☑ Yes ☐ No ☐ No meta-analysis conducted
	For NRSI For Yes:

	☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review ☐ Yes ☐ No ☑ No meta-analysis conducted
If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	For Yes: ☐ included only low risk of bias RCTs ☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect. ☐ Yes ☑ No ☐ No meta-analysis conducted
Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	For Yes: ☐ included only low risk of bias RCTs ☐ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results ☐ Yes ☑ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	For Yes: ☐ There was no significant heterogeneity in the results ☑ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review ☑ Yes ☐ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	For Yes: ☑ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias ☑ Yes ☐ No ☐ No meta-analysis conducted
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	For Yes: ☐The authors reported no competing interests OR ☐ The authors described their funding sources and how they managed potential conflicts of interest ☐ Yes ☑ No

Natural course

Di Lorenzo, 2013 [10]

Author	Di Lorenzo
Year of publication	2013
Journal	Clinical and Experimental Medicine
Study design	Follow-up study
Setting	Tertiary care, Italy
Study population	Baseline (1990-1995) 313 patients with clinical diagnosis of allergic rhinitis without asthma, 57% female, age range 44-56 yrs (mean 48.2, sd 2.9).

	Number at follow-up: 108 patients, 64% female, age range 59-69 yrs (mean 62.1).
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Nasal symptoms on VAS scale; diagnosis: mean 68.0, follow-up: 53.5 (p<0.0001). Complete remission: 27/118: 22.9%; less severe symptoms: 28/118: 23.7%; no change: 51/118: 42.2%; more severe symptoms: 12/118: 10.2%. Average follow-up 16.3 yrs (range 14-19)
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	Yes~~/no/unclear~~
Was follow-up sufficiently long and complete?	~~Yes~~/no~~/unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	Yes~~/no/unclear~~
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	~~Yes~~/no~~/unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes/no/~~unclear

Greisner, 1998 [11]

Author	Greisner
Year of publication	1998
Journal	Allergy and Asthma Proceedings
Study design	Follow-up study
Setting	Population based, USA
Study population	Baseline 1962/1963: 1836 college freshmen, 30% female. Number at follow-up: 1021 persons (56%), 31% female, mean age 40 yrs (range: 38-64y; 97% 40-42 yrs); at follow-up: 306 individuals with hay fever from a sample of 738 that had previously been skin tested
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Symptoms, at follow-up: 22.9% symptom free, 32.0% better (but not symptom free), unchanged 33.3%, worsened:9.2%, unknown: 2.6%. Follow-up: 23 years
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	Yes~~/no/unclear~~
Was follow-up sufficiently long and complete?	Yes~~/no/unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	~~Yes/no/~~unclear
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	~~Yes~~/no~~/unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes~~/no~~/unclear~~

Kellberger, 2012 [12]

Author	Kellberger
Year of publication	2012
Journal	Journal of Allergy and Clinical Immunology
Study design	Prospective cohort study

Setting	Population-based, Germany
Study population	Baseline 1995-1996: 6399 fourth grade class children, population-based sample. 85% answered questionnaire. Number at follow-up 2002-2003: 4893 adolescents aged 15-18 yrs; 77% answered questionnaire. Only persons of German descent were analysed.
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Symptoms (sneezing/runny nose/blocked nose, itchy/watery eyes): 64% retained symptoms at follow-up. Follow-up:7 yrs
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	Yes~~/no/unclear~~
Was follow-up sufficiently long and complete?	~~Yes~~/no~~/unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	~~Yes/no~~/unclear
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	Yes~~/no/unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes~~/no~~/unclear~~

Kong, 2012 [13]

Author	Kong
Year of publication	2012
Journal	American Journal of Rhinology & Allergy
Study design	Follow-up study
Setting	Population-based, China
Study population	Baseline: 1211 3-6 yrs old children in Wuhan, China, of whom 328 symptom positive (27.1%). Number at follow-up: 870 children (71.8%), of whom 256 symptom positive (29.4%).
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	From children symptom positive at baseline (n=328), 149 remained symptom positive (45%). Follow-up: 5 yrs
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	Yes~~/no/unclear~~
Was follow-up sufficiently long and complete?	~~Yes~~/no~~/unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	Yes~~/no/unclear~~
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	~~Yes~~/no~~/unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes~~/no~~/unclear~~

Lee, 2016 [14]

Author	Lee
Year of publication	2016
Journal	Clinical and Experimental Otorhinolaryngology
Study design	Prospective cohort study
Setting	Public school, Seoul, South-Korea

Study population	Baseline: 178 7 yrs (range, 6.5 to 7.5 years) old Korean children with allergic rhinitis. 107 boys and 71 girls.
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Number at follow-up: 122, of whom 18 had allergic rhinitis. Allergic rhinitis: 28% [5/18] remained allergic rhinitis, 72% [13/18] symptom free. Follow-up: 2 yrs.
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	~~Yes~~/no/~~unclear~~
Was follow-up sufficiently long and complete?	~~Yes~~/no/~~unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	Yes/~~no/unclear~~
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	~~Yes~~/no/~~unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes~~/no/~~unclear~~

Westman, 2012 [15]

Author	Westman
Year of publication	2012
Journal	Journal of Allergy and Clinical Immunology
Study design	Prospective cohort study
Setting	Population-based, Sweden
Study population	4089 children born from 1994 to 1996 enrolled at median age of 3 months. Analysed: 2024 children, who were tested for Phadiatop at both 4 and 8 years of age and had completed the questionnaires at age 0, 4, and 8 years
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	Of 4 yrs olds with allergic rhinitis, 87% continued having allergic rhinitis at the age of 8 yrs; 12% underwent remission. Follow-up: 4 yrs.
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	Yes/~~no/unclear~~
Was follow-up sufficiently long and complete?	~~Yes~~/no/~~unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	Yes/~~no/unclear~~
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	Yes/~~no/unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes~~/no/~~unclear~~

Yonekura, 2012 [16]

Author	Yonekura
Year of publication	2012

Journal	International Archives of Allergy and Immunology
Study design	Follow-up study
Setting	Population-based, Japan
Study population	Baseline: 1560 people >39 yrs from a small rural town in Japan, population-based sample, 58% female, 19% sensitization to cedar pollen, 5% seasonal allergic rhinitis due to cedar pollen, 16% sensitization to mite, 1.5% mite perennial allergic rhinitis. Number at follow-up: 703 persons, of whom 52 with seasonal allergic rhinitis in 1995.
Outcome measures (nasal / ocular symptoms, concentration, sleep problems, absenteeism from school / work, quality of life)	In 10/52 (19.2%) persons the symptoms of seasonal allergic rhinitis disappeared during follow-up. Follow-up: 10 yrs.
Risk of bias (tailormade criteria)	
Primary	
Was there a representative sample of patient (who were at a similar, identifiable, common, and possible early point in the course of the disease)?	Yes~~/no/unclear~~
Was follow-up sufficiently long and complete?	~~Yes~~/no~~/unclear~~
Secondary	
Were objective and unbiased outcome criteria used?	~~Yes/no~~/unclear
Were all characteristics of patients known or suspected to affect the outcome recorded (e.g. comorbidity)	Yes~~/no/unclear~~
Was there adjustment for important prognostic factors (e.g. information about treatment)?	~~Yes~~/no~~/unclear~~

Linked evidence

Köberlein, 2013 [17]

Author	Köberlein
Year of publication	2013
Journal	Asian Pacific Journal of Allergy and Immunology
Study design	Retrospective study
Setting	Data of post-marketing surveillance study in German medical practices
Study population	42,111 patients with allergic rhinitis, using antihistamine desloratadine 4-6 weeks, mean age 38.1 years (sd 14.9, range 11-101), 42.9% male, mean duration of disease 7.7 years (sd 66)
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	Physician reported compliance by asking patients whether the medication had been taken as instructed: excellent 74.5%, good 23.6%, 1.6% moderate, 0.3% poor
Risk of bias (JBI Critical Appraisal Checklist for Studies Reporting Prevalence Data)	Recall bias
Was the sample frame appropriate to address the target population?	~~Yes/No/~~Unclear~~/Not applicable~~
Were study participants sampled in an appropriate way?	~~Yes/No/~~Unclear~~/Not applicable~~
Was the sample size adequate?	Yes~~/No/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	~~Yes~~/No~~/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	~~Yes/No/~~Unclear~~/Not applicable~~

Were valid methods used for the identification of the condition?	~~Yes/~~No~~/Unclear/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	Yes~~/No/Unclear/Not applicable~~
Was there appropriate statistical analysis?	Yes~~/No/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	Yes~~/No/Unclear/Not applicable~~

Loh, 2004 [18]

Author	Loh
Year of publication	2004
Journal	Allergy
Study design	Prospective study, follow-up 30 days
Setting	Ear-nose-throat outpatient clinic, university hospital, Singapore
Study population	63 patients with allergic rhinitis, of whom 84.1% allergic, treated with nasal steroids (triamcinolone acetonide 1 dd 2 puffs). Mean age 29 years (range 15-68), 78% male, 79.4% had rhinitis longer than 1 year
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	≥ 75 compliance: 87% reported by patients, 65% by weighing the medication. 50-75 compliance: 11% reported by patients, 24% by weighing the medication. <50% compliance: 2% reported by patients, 11% by weighing the medication. 77.8% of the patients reported forgetfulness of using the medication for few times
Risk of bias	
Was the sample frame appropriate to address the target population?	Yes~~/No/Unclear/Not applicable~~
Were study participants sampled in an appropriate way?	Yes~~/No/Unclear/Not applicable~~
Was the sample size adequate?	~~Yes/~~No~~/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	Yes~~/No/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	Yes~~/No/Unclear/Not applicable~~
Were valid methods used for the identification of the condition?	Yes~~/No/Unclear/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	Yes~~/No/Unclear/Not applicable~~
Was there appropriate statistical analysis?	Yes~~/No/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	Yes~~/No/Unclear/Not applicable~~

Navarro, 2011 [19]

Author	Navarro
Year of publication	2011
Journal	Journal of Investigation of Allergology and Clinical Immunology
Study design	Observational study
Setting	Public health, primary care, ear-nose-throat clinics, allergy clinics, Spain
Study population	4040 patients with allergic rhinitis (confirmation not mentioned), 48% male, mean age 34 years (sd 14), duration of disease: mean 9 years (sd 8). Causes:

	pollen 68%, dust mite 52%, animal epithelia 21%, fungi 9%
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	Self report patients: 77% has taken the recommended doses (antihistamines and/or corticosteroids) for all or most of the indicated period/time, 20% reported adherence for only a short period of time, or when having symptoms, 1% never took treatment
Risk of bias	
Was the sample frame appropriate to address the target population?	Yes~~/No/Unclear/Not applicable~~
Were study participants sampled in an appropriate way?	~~Yes/No/~~Unclear~~/Not applicable~~
Was the sample size adequate?	Yes~~/No/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	Yes~~/No/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	Yes~~/No/Unclear/Not applicable~~
Were valid methods used for the identification of the condition?	~~Yes/No/~~Unclear~~/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	~~Yes/No/~~Unclear~~/Not applicable~~
Was there appropriate statistical analysis?	Yes~~/No/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	Yes~~/No/Unclear/Not applicable~~

Ocak, 2017 [20]

Author	Ocak
Year of publication	2017
Journal	International Journal of Pediatric Otorhinolaryngology
Study design	Questionnaire survey
Setting	Tertiary referral hospital, Turkey
Study population	76 children with allergic rhinitis, confirmed by skin-prick test and/or sIgE, on mometasone 1 puff/day therapy, mean age 7.82 years (range 3-16), 54% male
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	MMAS-8 score (Morisky medication adherence questionnaire) was used to measure adherence: 71.51% had 'good adherence'
Risk of bias	Limited description of methods and results
Was the sample frame appropriate to address the target population?	~~Yes/~~No~~/Unclear/Not applicable~~
Were study participants sampled in an appropriate way?	Yes~~/No/Unclear/Not applicable~~
Was the sample size adequate?	~~Yes/~~No~~/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	~~Yes/~~No~~/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	~~Yes/No/~~Unclear~~/Not applicable~~
Were valid methods used for the identification of the condition?	Yes~~/No/Unclear/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	Yes~~/No/Unclear/Not applicable~~
Was there appropriate statistical analysis?	~~Yes/~~No~~/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	Yes~~/No/Unclear/Not applicable~~

Pizzulli, 2014 [21]

Author	Pizzulli
Year of publication	2014
Journal	Clinical and Experimental Allergy
Study design	RCT, evaluating the effect of telemonitoring on medication adherence
Setting	Specialized care unit, Germany
Study population	63 patients with seasonal allergic rhinitis due to grass pollen, mean age 11.8 years (sd 2.7), 62% male, all on mometasone treatment, 66% with asthma, mean duration of disease 3.7 years (sd 2.6). Results of control group are reported in this evidence review: 32 patients, mean age 11.0 years (sd 2.9), 63% male, 69% on specific immunotherapy, 66% with asthma, 19% with atopic dermatitis, mean duration of disease 3.3 years (sd 3.0)
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	Optimal treatment is defined as at least puffs mometasone per day. Optimal treatment in control group: 12.5%
Risk of bias	
Was the sample frame appropriate to address the target population?	~~Yes~~/No/~~Unclear/Not applicable~~
Were study participants sampled in an appropriate way?	Yes/~~No/Unclear/Not applicable~~
Was the sample size adequate?	Yes/~~No/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	Yes/~~No/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	~~Yes~~/No/~~Unclear/Not applicable~~
Were valid methods used for the identification of the condition?	Yes/~~No/Unclear/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	Yes/~~No/Unclear/Not applicable~~
Was there appropriate statistical analysis?	Yes/~~No/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	Yes/~~No/Unclear/Not applicable~~

Wang, 2013 [22]

Author	Wang
Year of publication	2013
Journal	International Archives of Allergy and Immunology
Study design	RCT, single blind, evaluating the effect of a daily SMS on treatment adherence
Setting	University Hospital, China
Study population	50 patients with allergic rhinitis (confirmed with positive skin prick test and/or positive sIgE), of whom 39 completed the study (follow-up 30 days). Results of the control group are reported in this evidence review: 19 patients, 42% male, mean age 31.0 years (sd 10.88) mean duration of disease 2.07 years (sd 1.09)
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	Non-adherence: 18/25 patients. Reasons for non-adherence: 63.2% forgot to take the medication, 31.5% fear of side effects,

	36.8% belief that the medication was no longer needed, 10.5% believe that the medication was not effective, 5.26% inconvenience, 10.5% other reasons
Risk of bias	Inconsistencies in the report of numbers
Was the sample frame appropriate to address the target population?	Yes~~/No/Unclear/Not applicable~~
Were study participants sampled in an appropriate way?	Yes~~/No/Unclear/Not applicable~~
Was the sample size adequate?	~~Yes~~/No~~/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	~~Yes~~/No~~/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	~~Yes~~/No~~/Unclear/Not applicable~~
Were valid methods used for the identification of the condition?	~~Yes~~/No~~/Unclear/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	Yes~~/No/Unclear/Not applicable~~
Was there appropriate statistical analysis?	Yes~~/No/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	~~Yes~~/No~~/Unclear/Not applicable~~

Wong, 2010 [23]

Author	Wong
Year of publication	2010
Journal	Pediatric Allergy Journal
Study design	Questionnaire survey
Setting	National University Hospital Children's Specialist Clinic, Singapore
Study population	194 children (mean age 7.54 years, range 1-15) with rhinitis, of whom 79 with allergic rhinitis (clinically diagnosed, 61.9% male). All children had experience using topical nasal sprays and/or drops
Outcome measures (allergen avoidance, use of corticosteroids, use of antihistamines, compliance, treatment difficulties)	24.7% found the use of topical nasal medications unacceptable. 50% of all children had treatment difficulties/ unpleasantness: 31% itch/pain/ discomfort, 23% medication flowing down throat/nose, 20% struggle away, 16% experiences fear/anxiety/cries, 16% flat refusal, 15% unpleasant aftertaste, 10% only allowing application of 1 nostril, 9% nostril dryness, 0% nosebleed
Risk of bias	
Was the sample frame appropriate to address the target population?	Yes~~/No/Unclear/Not applicable~~
Were study participants sampled in an appropriate way?	Yes~~/No/Unclear/Not applicable~~
Was the sample size adequate?	Yes~~/No/Unclear/Not applicable~~
Were the study subjects and the setting described in detail?	Yes~~/No/Unclear/Not applicable~~
Was the data analysis conducted with sufficient coverage of the identified sample?	Yes~~/No/Unclear/Not applicable~~
Were valid methods used for the identification of the condition?	~~Yes/No~~/Unclear~~/Not applicable~~
Was the condition measured in a standard, reliable way for all participants?	~~Yes~~/No/Unclear~~/Not applicable~~
Was there appropriate statistical analysis?	Yes~~/No/Unclear/Not applicable~~
Was the response rate adequate, and if not, was the low response rate managed appropriately?	Yes~~/No/Unclear/Not applicable~~

References

1. Garcia Robaina JC, Sanchez Machin I, Fernandez-Caldas E, Iraola Calvo V, Vazquez Moncholi C, Bonnet Moreno C, et al. Skin tests and conjunctival and bronchial challenges with extracts of Blomia tropicalis and Dermatophagoides pteronyssinus in patients with allergic asthma and/or rhinoconjunctivitis. International Archives of Allergy & Immunology. 2003;131(3):182-8.
2. Haxel BR, Huppertz T, Boessert P, Bast F, Fruth K. Correlation of skin test results and specific immunoglobulin E blood levels with nasal provocation testing for house-dust mite allergies. American journal of rhinology & allergy. 2016;30(1):60-4. doi:10.2500/ajra.2016.30.4262.
3. King MJ, Tamulis T, Lockey RF. Prick puncture skin tests and serum specific IgE as predictors of nasal challenge response to dermatophagoides pteronyssinus in older adults. Ann Allergy Asthma Immunol. 2008;101(1):12-7. doi:10.1016/s1081-1206(10)60828-9.
4. Sheikh A, Hurwitz B, Nurmatov U, van Schayck CP. House dust mite avoidance measures for perennial allergic rhinitis. Cochrane Database of Systematic Reviews. 2010(7):CD001563. doi:10.1002/14651858.CD001563.pub3.
5. Compalati E, Baena-Cagnani R, Penagos M, Badellino H, Braido F, Gomez RM, et al. Systematic review on the efficacy of fexofenadine in seasonal allergic rhinitis: a meta-analysis of randomized, double-blind, placebo-controlled clinical trials. International Archives of Allergy & Immunology. 2011;156(1):1-15. doi:10.1159/000321896.
6. Compalati E, Canonica GW. Efficacy and safety of rupatadine for allergic rhino-conjunctivitis: a systematic review of randomized, double-blind, placebo-controlled studies with meta-analysis. Curr Med Res Opin. 2013;29(11):1539-51. doi:10.1185/03007995.2013.822855.
7. Al Sayyad JJ, Fedorowicz Z, Alhashimi D, Jamal A. Topical nasal steroids for intermittent and persistent allergic rhinitis in children. Cochrane Database of Systematic Reviews. 2007(1):CD003163. doi:10.1002/14651858.CD003163.pub4.
8. Penagos M, Compalati E, Tarantini F, Baena-Cagnani CE, Passalacqua G, Canonica GW. Efficacy of mometasone furoate nasal spray in the treatment of allergic rhinitis. Meta-analysis of randomized, double-blind, placebo-controlled, clinical trials. Allergy. 2008;63(10):1280-91. doi:10.1111/j.1398-9995.2008.01808.x.
9. Rodrigo GJ, Neffen H. Efficacy of fluticasone furoate nasal spray vs. placebo for the treatment of ocular and nasal symptoms of allergic rhinitis: a systematic review. Clinical & Experimental Allergy. 2011;41(2):160-70. doi:10.1111/j.1365-2222.2010.03654.x.
10. Di Lorenzo G, Leto-Barone MS, La Piana S, Ditta V, Di Fede G, Rini GB. Clinical course of rhinitis and changes in vivo and in vitro of allergic parameters in elderly patients: a long-term follow-up study. Clin Exp Med. 2013;13(1):67-73. doi:10.1007/s10238-012-0175-8.
11. Greisner WA, 3rd, Settipane RJ, Settipane GA. Natural history of hay fever: a 23-year follow-up of college students. Allergy & Asthma Proceedings. 1998;19(5):271-5. doi:10.2500/108854198778557728.
12. Kellberger J, Dressel H, Vogelberg C, Leupold W, Windstetter D, Weinmayr G, et al. Prediction of the incidence and persistence of allergic rhinitis in adolescence: a prospective cohort study. Journal of Allergy & Clinical Immunology. 2012;129(2):397-402.e1-3. doi:10.1016/j.jaci.2011.08.016.
13. Kong W, Chen J, Wang Y, Xiang J, Zhang X, Wang J, et al. A population-based 5-year follow-up of allergic rhinitis in Chinese children. American journal of rhinology & allergy. 2012;26(4):315-20. doi:10.2500/ajra.2012.26.3790.
14. Lee SH, Choi JH, Suh JD, Chung S, Hong SC, Kim JK, et al. Natural Course of Allergic and Nonallergic Rhinitis After 2 Years in Korean Children. Clin. 2016;9(3):233-7. doi:10.21053/ceo.2015.01130.
15. Westman M, Stjarne P, Asarnoj A, Kull I, van Hage M, Wickman M, et al. Natural course and comorbidities of allergic and nonallergic rhinitis in children. Journal of Allergy & Clinical Immunology. 2012;129(2):403-8. doi:10.1016/j.jaci.2011.09.036.
16. Yonekura S, Okamoto Y, Horiguchi S, Sakurai D, Chazono H, Hanazawa T, et al. Effects of aging on the natural history of seasonal allergic rhinitis in middle-aged subjects in South chiba, Japan. International Archives of Allergy & Immunology. 2012;157(1):73-80. doi:10.1159/000324475.
17. Koberlein J, Kothe AC, Sieber J, Mosges R. Determining factors of patient compliance to treatment in allergic rhinitis. Asian Pacific Journal of Allergy and Immunology. 2013;31(2):148-56. doi:10.12932/AP0264.31.2.2013.

18. Loh CY, Chao SS, Chan YH, Wang DY. A clinical survey on compliance in the treatment of rhinitis using nasal steroids. Allergy: European Journal of Allergy and Clinical Immunology. 2004;59(11):1168-72. doi:10.1111/j.1398-9995.2004.00554.x.
19. Navarro A, Valero A, Rosales MJ, Mullol J. Clinical use of oral antihistamines and intranasal corticosteroids in patients with allergic rhinitis. Journal of Investigational Allergology and Clinical Immunology. 2011;21(5):363-9.
20. Ocak E, Kocaoz D, Acar B. How can we improve medical adherence to intranasal corticosteroids in children? International journal of pediatric otorhinolaryngology. 2017;100:194-7. doi:10.1016/j.ijporl.2017.07.010.
21. Pizzulli A, Perna S, Florack J, Pizzulli A, Giordani P, Tripodi S, et al. The impact of telemonitoring on adherence to nasal corticosteroid treatment in children with seasonal allergic rhinoconjunctivitis. Clinical & Experimental Allergy. 2014;44(10):1246-54. doi:10.1111/cea.12386.
22. Wang K, Wang C, Xi L, Zhang Y, Ouyang Y, Lou H, et al. A randomized controlled trial to assess adherence to allergic rhinitis treatment following a daily short message service (SMS) via the mobile phone. International archives of allergy and immunology. 2014;163(1):51-8. doi:10.1159/000356317.
23. Wong IYZ, Soh SE, Chng SY, Shek LPC, Goh DYT, Van Bever HPS, et al. Compliance with topical nasal medication - An evaluation in children with rhinitis. Pediatric Allergy and Immunology. 2010;21(8):1146-50. doi:10.1111/j.1399-3038.2010.01015.x.

Appendix 3. GRADE Evidence Profiles

Diagnostic accuracy sIgE (D.pteronyssinus)

Question: Should sIgE (D.pteronyssinus) be used to diagnose allergic rhinitis in patients highly suspected of having allergic rhinitis?
Bibliography: Garcia Robaina, 2003; Haxel, 2016; King, 2008 [1-3]
Sensitivity: 0.84 to 1.00
Specificity: 0.54-1.00
Prevalences 30%*

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1.000 patients tested	Test accuracy Certainty of evidence
			Risk of bias	Indirect-ness	Inconsis-tency	Impreci-sion	Publication bias	pre-test probability of 30%	
True positives (patients with allergic rhinitis)	3 studies 189 patients	cross-sectional (cohort type accuracy study)	serious[a]	serious[b]	not serious	serious[c]	none	252 to 300	⨁◯◯◯ VERY LOW
False negatives (patients incorrectly classified as not having allergic rhinitis)								0 to 48	
True negatives (patients without allergic rhinitis)	3 studies 189 patients	cross-sectional (cohort type accuracy study)	serious[a]	serious[b]	not serious	serious[d]	none	378 to 700	⨁◯◯◯ VERY LOW
False positives (patients incorrectly classified as having allergic rhinitis)								0 to 322	

** This prevalence was chosen, based on the Dutch clinical guideline on allergic and non-allergic rhinitis for general practitioners [4], and was confirmed in the study of King et al [3]*
a. By far the largest study (Haxel 2016) has high risk of selection bias (QUADAS domain patient selection). One study (Garcia Robaina, 2003) has high risk of bias in flow & timing, one study (King, 2008) has high risk of bias in interpreting results of the reference test.
b. Two studies have been performed in tertiary (university) care, one study in secondary care. This review focusses on primary care. Diagnostic accuracy might vary between primary, secondary and tertiary care, because of variation in pre-test probabilities.
c. Only 108 patients with TP or FN results
d. Only 81 patients with TN or FP results

Diagnostic accuracy sIgE (D.farinae)

Question: Should sIgE (D.farinae) be used to diagnose allergic rhinitis in patients highly suspected of having allergic rhinitis?
Bibliography: Haxel, 2016 [2]
Sensitivity: 0.84 (95% CI: 0.69 to 0.93)
Specificity: 0.52 (95% CI: 0.37 to 0.66)
Prevalences 30%*

Outcome	№ of studies (№ of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1.000 patients tested	Test accuracy Certainty of evidence
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of 30%	
True positives (patients with allergic rhinitis)	1 studies 97 patients	cross-sectional (cohort type accuracy study)	serious[a]	serious[b]	not serious	serious[c]	none	251 (208 to 280)	⨁◯◯◯ VERY LOW
False negatives (patients incorrectly classified as not having allergic rhinitis)								49 (20 to 92)	
True negatives (patients without allergic rhinitis)	1 studies 97 patients	cross-sectional (cohort type accuracy study)	serious[a]	serious[b]	not serious	serious[c]	none	363 (262 to 460)	⨁◯◯◯ VERY LOW
False positives (patients incorrectly classified as having allergic rhinitis)								337 (240 to 438)	

** This prevalence was chosen, based on the Dutch clinical guideline on allergic and non-allergic rhinitis for general practitioners [4], and was confirmed in the study of King et al [3]*
a. Concerns about patient selection and flow & timing due to lack of information
b. The included study has been performed in tertiary care. This review focusses on primary care
c. Only 64 patients with TP or FN results

Avoidance measures

Question: Avoidance measures compared to no avoidance measures for patients with allergic rhinitis
Setting: Primary care
Bibliography: Sheikh, 2010 [5]

Certainty assessment							Impact	Certainty
Nº of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations		
Symptoms								
8	randomised trials	very serious [a]	not serious	not serious	serious[b]	none	*HEPA filters:* Study Reisman 1990: 32 from 40 patients evaluated: aggregated rhinitis and asthma symptom scores/medication scores: lower after active filtration vs placebo: day 8.79 vs 10.38, night 8.28 vs 9.90 (no statistical testing for total score). Nasal congestion, discharge, eye irritation, and upper airway itching reduced statistically significant, whereas cough, asthma and medication use did not. *Acaricides:* Study Kniest 1991: 20 patients: symptom scores 9-12 months vs 0-3 months lower in acaricide group vs control group; no absolute symptom scores. Study Bernstein 1995: 32 children, no disaggregated symptom scores for asthma and rhinitis. *Barrier bedding (=allergy control bedding):* Study Moon 1999: 29 from 30 patients evaluated: Mean daily symptom scores: decreased after 4 weeks in experimental group with 2.9 vs 0.3 in control group, statistically significant. Study Terreehorst 2003: 232 patients from 279 evaluated). No significant differences in symptom scores. Study Ghazala 2004: 26 from 30 patients completed the study: no differences in symptom scores reported between intervention and placebo. Study Brehler 2006: 21 from 32 patients completed the study. No significant reduction in symptom scores between intervention and control. *Barrier bedding and acaricides:* Study Incorvaia 2008: 25 from 29 patients evaluated: unclear difference between intervention and placebo	⊕◯◯◯ VERY LOW

a. Lack of information about randomisation procedures, lack of blinding in studies, absence of intention-to-treat design, large lost-to-follow-up
b. Few patients per specific intervention

Corticosteroids versus placebo

Question: Corticosteroids compared to placebo for patients with allergic rhinitis
Bibliography: Al Sayyad, 2007; Penagos, 2008; Rodrigo, 2010 [6-8]

Certainty assessment							Summary of findings				
Nº of participants (studies)	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Overall certainty of evidence	Study event rates (%)		Relative effect (95% CI)	Anticipated absolute effects	
							With placebo	With corticosteroids		Risk with placebo	Risk difference with corticosteroids
Nasal symptoms (mometasone) (assessed with: total nasal symptom score)											
1878 (10 RCTs)	not serious	serious[a]	not serious	not serious	none	⨁⨁⨁◯ MODERATE	911	967	-	-	SMD 0.56 SD lower (0.71 lower to 0.41 lower)
Non-nasal symptoms (mometasone)											
1009 (4 RCTs)	not serious	not serious	serious[b]	not serious	none	⨁⨁⨁◯ MODERATE	502	507	-	-	SMD 0.3 SD lower (0.43 lower to 0.18 lower)
Ocular symptoms (fluticasone, seasonal allergic rhinitis) (assessed with: Mean change in daily reflective total ocular symptom score)											
2219 (6 RCTs)	very serious[c]	not serious	not serious	not serious	none	⨁⨁◯◯ LOW	1112	1107	-		MD 0.54 lower (0.7 lower to 0.37 lower)
Ocular symptoms (fluticasone, perennial allergic rhinitis) (assessed with: Mean change in daily reflective total ocular symptom score)											
919 (3 RCTs)	very serious[c]	not serious	not serious	not serious	none	⨁⨁◯◯ LOW	455	464	-		MD 0.33 lower (0.61 lower to 0.05 lower)

CI: Confidence interval; SMD: Standardised mean difference; MD: Mean difference; SD: standard deviation
a. Moderate inconsistency between trial results; I2=58%
b. The outcome reported is 'non-nasal symptom scores', including ocular symptoms, otic, palate and throat complaints, cough, etc. For this outcome, we are interested in ocular symptoms only, therefore we downgraded for indirectness.
c. Unclear randomisation and allocation in almost all included studies; all studies were sponsored by pharmaceutical industries

Antihistamines versus placebo

Question: Antihistamines compared to placebo for patients with allergic rhinitis
Bibliography: Compalati, 2011; Compalati, 2013 [9, 10]

Certainty assessment							Summary of findings				
Nº of participants (studies)	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Overall certainty of evidence	Study event rates (%)		Relative effect (95% CI)	Anticipated absolute effects	
							With placebo	With antihistamines		Risk with placebo	Risk difference with antihistamines
Symptom score (fexofenadine) (follow up: range 24 hours to 24 hours; assessed with: 24-hour reflective total symptom score; Scale from: -1.00 to 1.00)											
1434 (3 RCTs)	serious[a]	not serious	not serious	not serious	none	⨁⨁⨁◯ MODERATE	718	716	-	-	SMD 0.42 SD lower (0.49 lower to 0.35 lower)
Total nasal symptom reduction (rupatidine) (Scale from: -1.00 to 1.00)											
1178 (7 RCTs)	serious[b]	not serious	not serious	not serious	none	⨁⨁⨁◯ MODERATE	595	583	-	-	SMD 0.36 SD lower (0.48 lower to 0.25 lower)
Ocular symptoms (rupatidine) (assessed with: itchy eyes; Scale from: -1.00 to 1.00)											
683 (4 RCTs)	serious[b]	not serious	not serious	not serious	none	⨁⨁⨁◯ MODERATE	345	338	-	-	SMD 0.29 SD lower (0.45 lower to 0.14 lower)

CI: Confidence interval; SMD: Standardised mean difference; SD: standard deviation
a. Medium risk of bias in the included studies, based on judgement of review authors
b. Unclear risk of selection bias and possible selective reporting, based on judgement of review authors

Natural course

Question: Course of nasal and ocular symptoms in patient with confirmed allergic rhinitis over the years
Bibliography: Di Lorenzo, 2013; Greisner, 1998; Kellberger, 2012; Kong 2012; Lee, 2016; Westman, 2012; Yonekura, 2012 [11-17]

Certainty assessment							**Effect**			**Certainty**
№ of studies	Study design	Risk of bias	Inconsis-tency	Indirect-ness	Impre-cision	Other considerations	№ of events	№ of individuals	Rate (95% CI)	
Remission (follow up: range 2 years to 23 years)										
7	Observa-tional studies	very serious[a]	serious[b]	serious[c]	serious	none	12-72%	8986		⨁◯◯◯ VERY LOW
Fewer symptoms or remission (follow up: range 16 years to 23 years)										
2	Observa-tional studies	serious[a]	not serious	serious[c]	serious	none	46-55%	1129		⨁◯◯◯ VERY LOW

a. Some of the included studies had a quite short follow-up, some of the studies had a large proportion lost-to-follow-up, most of the studies did not report comorbidity and medication use (and it is unlikely that participants did not use medication at all)
b. Large heterogeneity, which cannot be explained by differences in patient population, intervention or follow-up
c. It is assumed that the participants in the study use medication

References

1. Garcia Robaina JC, Sanchez Machin I, Fernandez-Caldas E, Iraola Calvo V, Vazquez Moncholi C, Bonnet Moreno C, et al. Skin tests and conjunctival and bronchial challenges with extracts of Blomia tropicalis and Dermatophagoides pteronyssinus in patients with allergic asthma and/or rhinoconjunctivitis. International Archives of Allergy & Immunology. 2003;131(3):182-8.
2. Haxel BR, Huppertz T, Boessert P, Bast F, Fruth K. Correlation of skin test results and specific immunoglobulin E blood levels with nasal provocation testing for house-dust mite allergies. American journal of rhinology & allergy. 2016;30(1):60-4. doi:10.2500/ajra.2016.30.4262.
3. King MJ, Tamulis T, Lockey RF. Prick puncture skin tests and serum specific IgE as predictors of nasal challenge response to dermatophagoides pteronyssinus in older adults. Ann Allergy Asthma Immunol. 2008;101(1):12-7. doi:10.1016/s1081-1206(10)60828-9.
4. Aalberse J, Fokkens W, Lucassen P, Sijbom M, Van Sleeuwen D, Wiersma TJ, et al. NHG-Standaard Allergische en niet-allergische rinitis. Utrecht: NHG - Dutch College of General Practitioners; 2018.
5. Sheikh A, Hurwitz B, Nurmatov U, van Schayck CP. House dust mite avoidance measures for perennial allergic rhinitis. Cochrane Database of Systematic Reviews. 2010(7):CD001563. doi:10.1002/14651858.CD001563.pub3.
6. Al Sayyad JJ, Fedorowicz Z, Alhashimi D, Jamal A. Topical nasal steroids for intermittent and persistent allergic rhinitis in children. Cochrane Database of Systematic Reviews. 2007(1):CD003163. doi:10.1002/14651858.CD003163.pub4.
7. Penagos M, Compalati E, Tarantini F, Baena-Cagnani CE, Passalacqua G, Canonica GW. Efficacy of mometasone furoate nasal spray in the treatment of allergic rhinitis. Meta-analysis of randomized, double-blind, placebo-controlled, clinical trials. Allergy. 2008;63(10):1280-91. doi:10.1111/j.1398-9995.2008.01808.x.
8. Rodrigo GJ, Neffen H. Efficacy of fluticasone furoate nasal spray vs. placebo for the treatment of ocular and nasal symptoms of allergic rhinitis: a systematic review. Clinical & Experimental Allergy. 2011;41(2):160-70. doi:10.1111/j.1365-2222.2010.03654.x.
9. Compalati E, Canonica GW. Efficacy and safety of rupatadine for allergic rhino-conjunctivitis: a systematic review of randomized, double-blind, placebo-controlled studies with meta-analysis. Curr Med Res Opin. 2013;29(11):1539-51. doi:10.1185/03007995.2013.822855.
10. Compalati E, Baena-Cagnani R, Penagos M, Badellino H, Braido F, Gomez RM, et al. Systematic review on the efficacy of fexofenadine in seasonal allergic rhinitis: a meta-analysis of randomized, double-blind, placebo-controlled clinical trials. International Archives of Allergy & Immunology. 2011;156(1):1-15. doi:10.1159/000321896.
11. Di Lorenzo G, Leto-Barone MS, La Piana S, Ditta V, Di Fede G, Rini GB. Clinical course of rhinitis and changes in vivo and in vitro of allergic parameters in elderly patients: a long-term follow-up study. Clin Exp Med. 2013;13(1):67-73. doi:10.1007/s10238-012-0175-8.
12. Greisner WA, 3rd, Settipane RJ, Settipane GA. Natural history of hay fever: a 23-year follow-up of college students. Allergy & Asthma Proceedings. 1998;19(5):271-5. doi:10.2500/108854198778557728.
13. Kellberger J, Dressel H, Vogelberg C, Leupold W, Windstetter D, Weinmayr G, et al. Prediction of the incidence and persistence of allergic rhinitis in adolescence: a prospective cohort study. Journal of Allergy & Clinical Immunology. 2012;129(2):397-402.e1-3. doi:10.1016/j.jaci.2011.08.016.
14. Kong W, Chen J, Wang Y, Xiang J, Zhang X, Wang J, et al. A population-based 5-year follow-up of allergic rhinitis in Chinese children. American journal of rhinology & allergy. 2012;26(4):315-20. doi:10.2500/ajra.2012.26.3790.
15. Lee SH, Choi JH, Suh JD, Chung S, Hong SC, Kim JK, et al. Natural Course of Allergic and Nonallergic Rhinitis After 2 Years in Korean Children. Clin. 2016;9(3):233-7. doi:10.21053/ceo.2015.01130.
16. Westman M, Stjarne P, Asarnoj A, Kull I, van Hage M, Wickman M, et al. Natural course and comorbidities of allergic and nonallergic rhinitis in children. Journal of Allergy & Clinical Immunology. 2012;129(2):403-8. doi:10.1016/j.jaci.2011.09.036.
17. Yonekura S, Okamoto Y, Horiguchi S, Sakurai D, Chazono H, Hanazawa T, et al. Effects of aging on the natural history of seasonal allergic rhinitis in middle-aged subjects in South chiba, Japan. International Archives of Allergy & Immunology. 2012;157(1):73-80. doi:10.1159/000324475.

Chapter 3.

Do clinical practice guidelines consider evidence about diagnostic test consequences on patient-relevant outcomes? A critical document analysis

Mariska Tuut
Jako Burgers
Trudy van der Weijden
Miranda Langendam

Journal of Evaluation in Clinical Practice 2022; https://doi.org/10.1111/jep.13619

Abstract

Rationale, aims and objectives: Supporting evidence for diagnostic test recommendations in clinical practice guidelines (CPGs) should not only include diagnostic accuracy, but also downstream consequences of the test result on patient-relevant outcomes. The aim of this study is to assess the extent to which evidence-based CPGs about diagnostic tests cover all relevant test-treatment pathway components.

Methods: We performed a systematic document analysis and quality assessment of publicly accessible CPGs about three common diagnostic tests: C-reactive protein (CRP), colonoscopy and fractional exhaled nitric oxide (FeNO). Evaluation of the impact of the full test-treatment pathway (diagnostic accuracy, burden of the test, natural course of target condition, treatment effectiveness, and link between test result and administration of treatment) on patient relevant outcomes was considered best practice for developing medical test recommendations.

Results: We retrieved 15 recommendations in 15 CPGs. The methodological quality of the CPGs varied from poor to excellent. Ten recommendations considered diagnostic accuracy. Four of these were funded on a systematic review and rating of the certainty in the evidence. None of the CPGs evaluated all steps of the test-treatment pathway. Burden of the test was considered in three CPGs, but without systematically reviewing the evidence. Natural course was considered in two CPGs, without a systematic review of the evidence. In three recommendations, treatment effectiveness was considered, supported with a systematic review and rating of the certainty in the evidence in one CPG. The link between test result and treatment administration was not considered in any CPG.

Conclusions: The included CPGs hardly seem to consider evidence about test consequences on patient-relevant outcomes. This might be explained by reporting issues and challenging methodology. Future research is needed to investigate how to facilitate guideline developers in explicit reliable consideration of all steps of a test-treatment pathway when developing diagnostic test recommendations.

Keywords: diagnosis, clinical guidelines, systematic reviews, evidence-based medicine

Introduction

Clinicians use medical tests to confirm or exclude a clinical diagnosis (e.g. PCR-test to diagnose COVID-19), to test the likelihood of a certain clinical diagnosis (e.g. PSA-test to screen for risk on prostate cancer) or for follow-up of patients to monitor recovery (e.g. rehabilitation checklists) [1]. Test results guide (treatment) decisions. The clinical value of a medical test depends on various elements: the patient population characteristics (e.g. prevalence of the disease), test characteristics (e.g. sensitivity and specificity) and its downstream consequences (e.g. benefits and harms of treatment) on patient-important outcomes [2].

Clinical practice guidelines (CPGs) provide recommendations to support professionals and patients in clinical decision-making, with the ultimate goal of improving or maintaining patients' health. In the development of CPGs, the benefits and harms of the interventions of interest are systematically assessed with regard to patient-relevant outcomes. The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach is designed to facilitate this process [3].

Diagnostic CPGs provide recommendations about the use of a certain test (or test strategy). Supporting evidence for these recommendations consists of studies about diagnostic accuracy [4].However, acceptable test characteristics (sensitivity and specificity) are not enough to improve patients' health. CPG developers should also consider downstream consequences (e.g. burden of the test and the proportion of patients with a certain test result who receive the recommended treatment) on patient-relevant outcomes (e.g. mortality, morbidity and quality of life)(see *figure 1*) [5, 6].

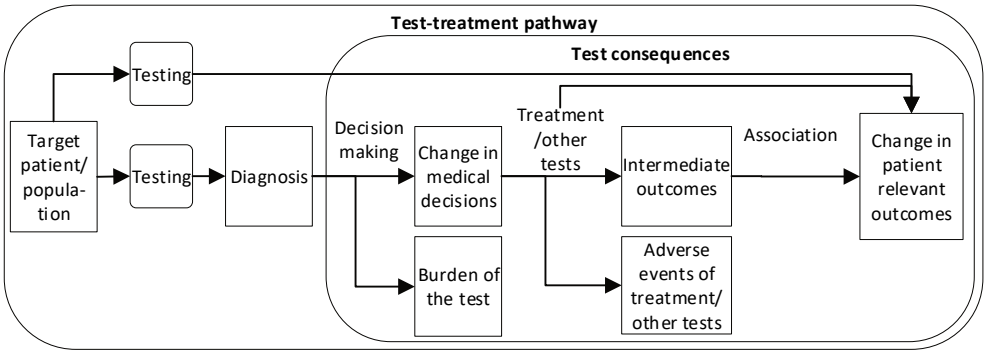


Figure 1.Test-treatment pathway (adapted from Harris et al, 2001) [7]

The interpretation of evidence about the value of therapeutic interventions is complex, and there is room for improvement [8]. This applies even more to evidence about diagnostic tests and its translation into CPG recommendations [9-11]. There have been

few randomised controlled trials on the value of test-treatment pathways for patient-relevant outcomes [9]. Evaluating the value of diagnostic tests on patient-relevant outcomes in CPGs is thus complex since it requires integration of various pieces of evidence for the different links in a chain (see *figure 1*).

In the GRADE approach for diagnostic tests and test strategies, the first step is to formulate the clinical question, including definition of patient-important outcomes and description of the aim of the test (add-on, replacement or triage). The next step is to assess diagnostic accuracy and downstream consequences of testing. These include the burden of the test, clinical management, natural course of the target condition (to estimate the outcomes of patients with a false negative test result), and the link between test result and management (proportion of patients with a certain test result who receive the recommended treatment). Ideally, each evidence component is based on a systematic review of the literature and the certainty in the evidence for each component is determined separately [9].Finally, the evidence components are integrated and the overall certainty in the evidence is assessed [12, 13]. To move from evidence to recommendation, guideline developers use the GRADE evidence-to-decision framework [12].

The aim of this study is to assess the extent to which evidence-based CPGs about diagnostic tests cover all relevant test-treatment pathway components.

Specific objectives are to assess the types of supporting evidence used for CPG recommendations about diagnostic tests, and to explore determinants of best practices. In the context of CPG development about the value of a diagnostic test, we formulated the following research questions:

1. Which types of evidence (diagnostic accuracy, burden of the test, natural course, treatment effectiveness, link between test result and administration of treatment) are used to support the recommendations?
2. Which factors (e.g. composition of the guideline panel, use of the GRADE approach, methodological quality according to AGREE II's domain methodology) contribute to completeness of the evidence?
3. To what extent can differences between CPG recommendations be explained by including different types of evidence?

Answers to these questions elucidate gaps in the implementation of good CPG development methods when developing recommendations about diagnostic tests and test strategies and can help guideline methodologists in developing strategies to facilitate this process.

Methods

Design

In order to assess the types of supporting evidence used for CPG recommendations about diagnostic tests, and to identify factors related to the extent of the supporting evidence, we performed a systematic document analysis of recent versions of publicly accessible CPGs concerning 3 diagnostic topics.

Topics

We chose tests that are frequently used to diagnose three common diseases. We considered tests with different characteristics (primary vs. secondary care, negligible vs. reasonable risk of serious burden of the test, low vs. high costs) to identify possible factors related to differences in methodological approach in the development of the CPGs. We therefore selected the following tests:

- C-reactive protein test (CRP) to increase the likelihood of pneumonia (annual incidence estimated at 0.5-1.1%) in primary care patients with cough (excluding diagnostic procedures in patients suspected of having a COVID-19 infection) [14]
- Colonoscopy to detect colon cancer (annual incidence 1,148,515 new cases) in secondary care patients suspected of having (primary) colon cancer (excluding screening and tests in patients at risk of hereditary types of colon cancer) [15]
- Fractional exhaled nitric oxide (FeNO) to diagnose (severe) asthma (prevalence 3.6%) in children and adults in primary and secondary care (excluding monitoring of asthma) [16]

Search and selection of relevant CPGs

Current, publicly accessible, recent (publication date 2016-2020) CPGs were eligible if they included recommendations about the tests mentioned above, were CPGs at a national or international level, and were published in English, German or Dutch.

To identify relevant CPGs, one author (MT) performed the search and selected the CPGs. The selection was checked for accuracy by a second author (JB). In February 2021, we searched the International Guideline Library from Guidelines International Network (GIN, (https://guidelines.ebmportal.com/), including around 3000 CPGs, mostly developed by organizational GIN members), databases from organizational GIN members active in CPG development (n=103), the TRIP database (Turning Research Into Practice (https://www.tripdatabase.com/), containing around 10.000 English-language CPGs) and Medline (see *Appendix1* for full search details).

Identification of recommendations

We analysed the content of the selected CPGs to identify relevant recommendations, including supporting evidence available online (e.g. tables with study characteristics, evidence documents, GRADE Evidence Profiles), as well as information about the methods of CPG development of the developing organisation (e.g. methodology manuals).

Data extraction

In the preparatory phase of this study, we piloted data extraction on two recommendations with four authors (MT, ML, JB, TvdW) to refine the data extraction form and define the variables for which we needed data extraction in duplicate. One author (MT) extracted the initial characteristics of each recommendation and CPG (CPG title (including English translation if relevant), initial developing organisation, country, publication year, recommendation text (including English translation if relevant)).

Detailed information about each recommendation and CPG was extracted by one author (MT) and critically reviewed by another author (ML, JB or TvdW) using a predefined and piloted data extraction form (see *Appendix 2* for the data extraction form and the categorisation of the variables). The form consisted of questions about scope and target audience of the CPG and composition of the CPG panel, involvement of methodologist(s), methodological quality of the CPG (using AGREE II, domain methodology, items 7-12) [17, 18], patient involvement (using AGREE II item 5) [17, 18], the types and extent of supporting evidence for the recommendation (consideration and inclusion of systematic evaluation with assessment of the certainty in the evidence about diagnostic accuracy, burden of the test, natural course, treatment effectiveness and link between test result and administration of treatment), grading of the recommendation and use of the GRADE approach, direction of the recommendation, and characteristics of the test and target condition. Disagreements between the reviewers were discussed until consensus was reached.

Analysis

We tabulated basic and detailed characteristics of the included recommendations and CPGs. Systematic evaluation (with a systematic review of the literature and assessment of the certainty in the evidence) of the impact of the full test-treatment pathway (diagnostic accuracy, burden of the test, natural course of target condition, treatment effectiveness, and link between test result and administration of treatment) on patient relevant outcomes was considered best practice for developing medical test recommendations.

We planned to analyse a possible relation between differences in evidence base and methodological factors (e.g. composition of the CPG panel, involvement of patients and methodologists, development approach). However, because the data about the evidence base were quite homogenous we were not able to perform these analyses.

Results

Search and selection of relevant CPGs

Full details of the search and selection process are described in *Appendix 1*. In short, the search identified 15 unique relevant recommendations in 15 CPGs: four about CRP related to the diagnosis pneumonia in primary care [19-22], five about colonoscopy in secondary care patients suspected of having colon cancer [23-27], and six about the use of FeNO to diagnose (severe) asthma [28-33]. The search and selection process is illustrated in *figure 2*.

In *table 1*, we present the included CPGs with information about the developing organisation, the country of publication and the publication year. All guidelines originated from high-income countries.

Quality of the guidelines and use of the GRADE approach

Table 2 presents detailed information about the composition of the CPG panel, the methodological quality of the included CPGs, the direction and grading of the recommendation and the reported and actual use of the GRADE approach. Nine out of 15 CPGs included a methodologist in the development process, in the CPG panel and/or at bureau level [23, 24, 27-33]. In all CPGs about FeNO a methodologist was involved, and in none of the CPGs about CRP. Patient involvement and inclusion of patient perspective varied a lot between the CPGs. AGREE II methodology domain scores varied from 8 to 42 (possible range from worst to best: 6-42), with the highest scores for the CPGs about FeNO. Thirteen of the included recommendations were in favour of the test of interest, only one recommendation about CRP [22], and one recommendation about FeNO [28], advised against the use of the test. Eleven recommendations were graded, which included all recommendations about CRP [19-22], two out of five recommendations about colonoscopy [25, 26], and five out of six recommendations about FeNO [28-30, 32, 33]. Seven CPGs reported to have used the GRADE approach [20, 21, 24, 28-30]; in four of these elements of the GRADE approach (such as a GRADE evidence profile) were recognized [24, 28-30]. No clear differences between the topics were identified in the (reported) use of the GRADE approach.

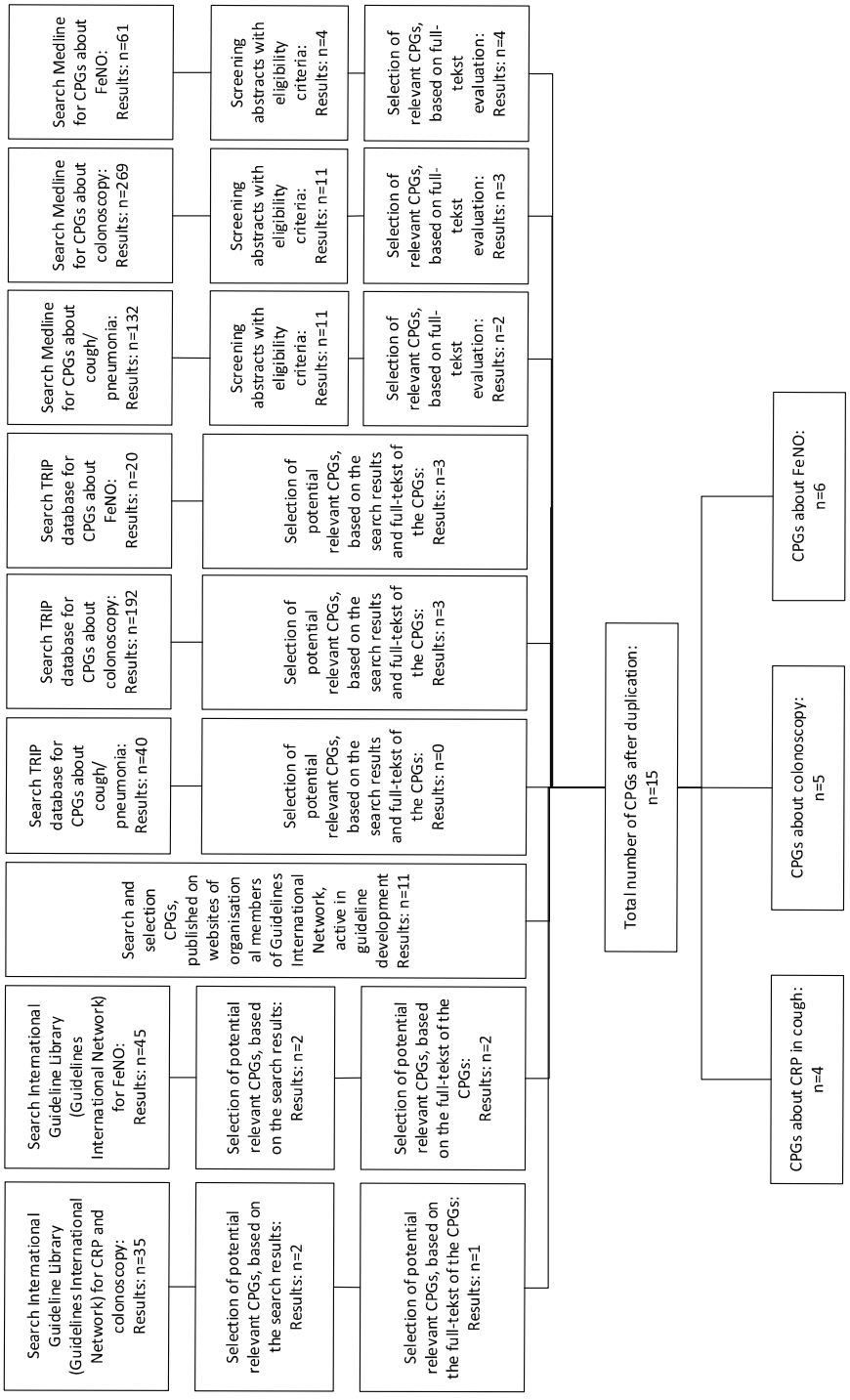


Figure 2. Search and selection of relevant CPGs

Table 1. Basic characteristics of the included CPGs

Organisation	Year	Country	Title (original language)	English-translated title in case of non-English language CPG
CRP				
Deutschen Gesellschaft für Pneumologie und Beatmungsmedizin (DGPB) [20]	2016	Germany	Behandlung von erwachsenen Patienten mit ambulant erworbener Pneumonie und Prävention	Prevention and management of adult patients with community acquired pneumonia
American College of Chest Physicians (ACP) [21]	2019	United States	Adult Outpatients with acute cough due to suspected pneumonia or influenza	
Ministry of Public Health, Qatar (MoPH) [19]	2019	Qatar	The diagnosis & management of community acquired pneumonia	
Deutschen Gesellschaft für Pädiatrische Infektiologie (DGPI) [22]	2017	Germany	Management der ambulant erworbenen Pneumonie bei Kindern und Jugendlichen (pädiatrische ambulant erworbene Pneumonie, pCAP)	Management of community acquired pneumonia in children and adolescents
Colonoscopy				
European Society for Medical Oncology (ESMO) [25]	2020	Europe	Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up	
Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. (AWMF) [23]	2019	Germany	Kolorektales Karzinom	Colorectal cancer
Association of Coloproctology of Great Britain & Ireland (ACPGBI) [26]	2017	Great Britain and Ireland	Guidelines for the Management of Cancer of the Colon, Rectum and Anus (2017) – Diagnosis, Investigations and Screening	
Federatie Medisch Specialisten (FMS) [24][24]	2019	The Netherlands	Colorectaal carcinoom	Colorectal cancer
Nederlands Huisartsen Genootschap (NHG) [27]	2017	The Netherlands	Rectaal bloedverlies	Rectal bleeding
FeNO				
Arbeitsgemeinschaft der Wissenschaftlichen	2020	Germany	Nationale VersorgungsLeitlinie Asthma	National Guideline on asthma

Organisation	Year	Country	Title (original language)	English-translated title in case of non-English language CPG
Medizinischen Fachgesellschaften e.V. (AWMF) [28]				
Ministry of Public Health (MoPH_A) [31]	2019	Qatar	The diagnosis & management of asthma in adults	
Ministry of Public Health (MoPH_C) [33]	2019	Qatar	The diagnosis & management of asthma in children	
National Asthma Education and Prevention Program (NAEPP) [29]	2020	USA	Managing Asthma in Adolescents and Adults	
National Institute for Health and Care Excellence (NICE) [30]	2020	UK	Asthma: diagnosis, monitoring and chronic asthma management	
Scottish Intercollegiate Guidelines Network (SIGN) [32]	2019	UK	British guideline on the management of asthma	

Supporting evidence for the recommendations

Detailed information about the supporting evidence for the included recommendations is presented in *table 3*. Ten CPGs out of 15 considered diagnostic accuracy [20-22, 24, 26-30, 32], of which four underpinned these considerations with a systematic review of the literature and a judgement of the certainty in the evidence [21, 28-30]. Burden of the test was considered in three CPGs [24, 27, 29], and two CPGs considered the natural course of the disease [19, 32], all without systematically reviewing the literature. Three CPGs considered treatment effectiveness [19, 25, 28], of which one performed a systematic review of the literature with judgement of the certainty in the evidence [28]. Not any CPG considered the link between the test result and administration of treatment. As a consequence, there were no CPGs that considered all test consequences of the test-treatment pathway.

Since no CPG systematically evaluated all steps of the test-treatment pathway, we were not able to identify a best practice, nor could we study possible relationships between clarifying factors and supporting evidence for a recommendation.

Table 2. Detailed characteristics of the CPG quality, the recommendation and the (reported) use of GRADE

	CPG panel		Methodology (AGREE II scores)							Recom-mendation		GRADE	
Guideline	Methodologist involvement	Patient involvement (AGREE II score)	Systematic evidence search methods	Clear criteria for evidence selection	Clear description of the strengths and limitations of the body of evidence	Clear description methods for formulating recommendations	Health benefits, side effects and risks have been considered	Explicit link between recommendation and supporting evidence	SUM score methodology domain†	Direction	Graded	Reported use of GRADE approach	Elements of GRADE approach in CPG
CRP													
DGPB, 2016 [20]	-	2	4	2	3	6	3	5	23	+	+	+	-
ACP, 2019 [21]	-	5	6	5	2	6	2	6	27	+	+	+	-
MoPH, 2019 [19]	-	1	2	1	2	1	1	1	8	+	+	-	-
DGPI, 2017 [22]	-	2	3	1	1	7	5	2	19	-	+	-	-
Colonoscopy													
ESMO, 2020 [25]	-	1	1	1	1	3	1	2	9	+	+	-	-
AWMF, 2019 [23]	+	3	7	7	5	7	3	6	35	+	-	-	-
ACPGBI, 2017 [26]	-	1	2	1	3	2	1	4	13	+	+	-	-
FMS, 2019 [24]	+	5	2	1	2	4	6	7	22	+	-	+	+
NHG, 2017 [27]	+	1	5	1	1	5	6	7	25	+	-	-	-
FeNO													
AWMF, 2020 [28]	+	7	7	7	7	7	6	7	41	-	+	+	+
MoPH_A, 2019 [31]	+	3	4	4	2	2	2	2	16	+	-	-	-
MoPH_C, 2019 [33]	+	3	3	3	2	2	2	2	14	+	+	-	-
NAEPP, 2020 [29]	+	7	7	7	7	7	7	7	42	+	+	+	+
NICE, 2020 [30]‡	+	7	7	7	7	6	7	7	41	+	+	+	+
SIGN, 2019 [32]	+	7	7	3	3	4	5	7	29	+	+	+	-

+: yes; +/-: unclear; -: no; †possible range: 6-42; ‡This CPG contains two separate recommendations concerning the use of FeNO in the diagnosis of childhood respectively adult onset asthma; the scores are identical

Table 3. Detailed information about the supporting evidence for the recommendations

Guideline	Diagnostic accuracy			Burden of the test			Natural course			Treatment effectiveness			Considered Link between test result and administration of treatment
	Considered	Considered with systematic review of the literature	Considered with systematic review of the literature and certainty in the evidence	Considered	Considered with systematic review of the literature	Considered with systematic review of the literature and certainty in the evidence	Considered	Considered with systematic review of the literature	Considered with systematic review of the literature and certainty in the evidence	Considered	Considered with systematic review of the literature	Considered with systematic review of the literature and certainty in the evidence	
CRP													
DGPB, 2016 [20]	+	-	-	-	-	-	-	-	-	-	-	-	-
ACP, 2019 [21]	+	+	+	-	-	-	-	-	-	-	-	-	-
MoPH, 2019 [19]	-	-	-	-	-	-	+	-	-	+	-	-	-
DGPI, 2017 [22]	+	-	-	-	-	-	-	-	-	-	-	-	-
Colonoscopy													
ESMO, 2020 [25]	-	-	-	-	-	-	-	-	-	+	-	-	-
AWMF, 2019 [23]	-	-	-	-	-	-	-	-	-	-	-	-	-
ACPGBI, 2017 [26]	+	-	-	-	-	-	-	-	-	-	-	-	-
FMS, 2019 [24]	+	-	-	+	-	-	-	-	-	-	-	-	-
NHG, 2017 [27]	+	-	-	+	-	-	-	-	-	-	-	-	-
FeNO													
AWMF, 2020 [28]	+	+	+	-	-	-	-	-	-	+	+	+	-
MoPH_A, 2019 [31]	-	-	-		-	-	-	-	-	-	-	-	-
MoPH_C, 2019 [33]	-	-	-	-	-	-	-	-	-	-	-	-	-
NAEPP, 2020 [29]	+	+	+	+	-	-	-	-	-	-	-	-	-
NICE, 2020 [30][†]	+	+	+	-	-	-	-	-	-	-	-	-	-
SIGN, 2019 [32]	+	-	-	-	-	-	+	-	-	-	-	-	-

+: yes; +/-: unclear; -: no; [†]This CPG contains two separate recommendations concerning the use of FeNO in the diagnosis of childhood respectively adult onset asthma; the scores are identical

Discussion

Our document analysis on a sample of 15 CPGs about CRP, colonoscopy and FeNO diagnostic tests revealed that none of these CPGs reported evidence on all components of the test-treatment pathway. Consideration of any test consequences on patient-relevant outcomes was described in only six CPGs (three CPGs considered burden of the test, two considered natural course of the disease of interest, and four considered treatment effectiveness). Systematic review of the literature, including a judgement of the certainty in the supporting evidence was only reported for four recommendations and covered diagnostic accuracy in all four cases and treatment effectiveness in one case.

The importance of systematically evaluating test consequences for the purpose of developing CPGs has been recognised [5, 12, 13]. For instance, one could imagine that a certain diagnostic test might have limited value when it has no treatment consequences (e.g. no treatment available). Or, when comparing two tests with the same diagnostic accuracy to ascertain the same disease, one could recognize that differences in test burden may play an important role.

This study suggests that implementation of the systematic evaluation of the value of a test is lagging behind. This also applies to CPGs that claim to use the GRADE approach. There seems to be a gap between following a methodologically robust approach and developing CPGs in practice.

Two issues may explain that gap. First, guideline developers may have considered the downstream consequences of a diagnostic test but did not explicitly report these. It may not be strictly necessary to systematically evaluate all evidence components. However, we still recommend transparent documentation of choices made in the guideline development process. A guideline user should be able to read which elements of a test-treatment pathway were considered and how, and which were not considered and why.

Second, performing systematic literature reviews of the complete test-treatment pathway – including assessment of the certainty in the evidence of test accuracy and downstream consequences – is complex and time-consuming. The use of the GRADE approach for the evaluation of diagnostic tests and test strategies is considered challenging [10, 11]. Strategies to facilitate the use of this approach, such as training of CPG panel members, may improve the application. Unfortunately, we could not determine factors that contribute to successful use of the GRADE approach, because we could not identify a ‘best practice’.

A lack of transparency in combination with the use of state-of-the-art methods was also described by Arevalo-Rodriguez and colleagues, who studied the methods and reports of 191 rapid reviews of medical tests [34]. In the majority of those reviews, the study selection method was not reported. Although almost 20% of the reviews claimed to have applied the GRADE approach, few actually reported the data extraction and quality appraisal methods.

This finding is consistent with a recent report on the application of GRADE in U.S. guidelines [35]. Although guideline developers indicated that they used the GRADE approach, only 10% of the included CPGs reported on all 8 criteria for assessing the certainty in the evidence (e.g. indirectness and dose-response gradient), and around half of these included an evidence profile or summary of findings table.

Gopalakrishna et al. studied barriers in the development of recommendations about medical tests in a qualitative study among European CPG developers [36]. They also reported challenges in the development of recommendations about medical tests, e.g. in the definition of key questions, the types of evidence and outcomes included in the CPG, and synthesizing and appraising the evidence. Awareness and education were reported as the most important ways to solve these challenges.

Our study emphasises the need for more knowledge and expertise among CPG developers when evaluating diagnostic tests. Currently available competency-based frameworks for CPG developers do not include a special focus on diagnostic test evaluation [37, 38]. This also applies for current training programs of CPG panel members, e.g. INGUIDE [39]. Facilitating the implementation of GRADE for diagnosis by defining competencies and training needs may improve the quality of CPGs about diagnostic tests.

Strengths and limitations

This study evaluated the supporting evidence of recommendations in CPGs on three medical tests. The selection of only three topics is a limitation of this study. However, we chose three diagnostic tests with divergent characteristics (e.g. invasiveness, possible burden of the test, disease of interest, costs) allowing comparison of many CPGs. The homogenous results in all three clusters of CPGs strengthens the external validity of our findings. Additionally, we found large variance in methodological quality of the included CPGs. However, high scoring CPGs on the AGREE II domain methodology did not reflect a better or more transparent underpinning of the recommendations than lower scoring CPGs.

Due to the document analysis design we could not retrieve information about the dynamics in the CPG panels that could explain their decisions and reasons for lack of transparency in the CPG documents. We did not contact the CPG developers, since in our opinion CPG users should be able to find the considerations of the panel beyond the recommendations in the published documents of the CPG.

Implications for practice

We suggest that developers of CPGs about diagnostic tests clearly describe which elements of a test-treatment pathway were or were not considered and why. In addition, CPG developers should indicate the presence or absence of systematic reviews of the evidence, including determination of the certainty in that evidence, for all evaluated parts of the test-treatment pathway, which is also usual in recommendations about therapy. Facilitating the implementation of GRADE for diagnosis will be useful to improve the clinical content of CPGs.

Implications for research

This study highlighted the lack of (transparency about) supporting evidence for diagnostic test recommendations in CPGs. A next step could be to study why CPG developers do not report all elements of the test-treatment pathway, including a review of the evidence and its quality. Furthermore, it is worthwhile to research how to facilitate CPG developers in explicitly and reliably considering all relevant steps of a test-treatment pathway when developing diagnostic test recommendations.

Conclusion

Diagnostic test recommendations in the included CPGs are mainly based on evidence and considerations on diagnostic accuracy. Other steps of the test-treatment strategy, such as burden of the test, natural course of the disease of interest, effectiveness of treatment of the disease of interest and the link between the test result and the administration of treatment should receive more attention in CPGs in order to consider evidence about test consequences on patient-relevant outcomes.

References

1. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001;323(7305):157-62. doi:10.1136/bmj.323.7305.157.
2. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jama. 1994;271(5):389-91. doi:10.1001/jama.271.5.389.
3. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94. doi:10.1016/j.jclinepi.2010.04.026.
4. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ. 2002;324(7336):539-41. doi:10.1136/bmj.324.7336.539.
5. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.
6. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. Evidence-based medicine. 2008;13(6):162-3. doi:10.1136/ebm.13.6.162-a.
7. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21-35. doi:10.1016/s0749-3797(01)00261-6.
8. Florez ID, Brouwers MC, Kerkvliet K, Spithoff K, Alonso-Coello P, Burgers J, et al. Assessment of the quality of recommendations from 161 clinical practice guidelines using the Appraisal of Guidelines for Research and Evaluation-Recommendations Excellence (AGREE-REX) instrument shows there is room for improvement. Implement Sci. 2020;15(1):79. doi:10.1186/s13012-020-01036-5.
9. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol. 2012;65(3):282-7. doi:10.1016/j.jclinepi.2011.07.003.
10. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014;67(7):760-8. doi:10.1016/j.jclinepi.2014.01.006.
11. Tuut MK, de Beer JJA, Burgers JS, van de Griendt EJ, van der Weijden T, Langendam MW. Applying GRADE for diagnosis revealed methodological challenges: an illustrative example for guideline developers. J Clin Epidemiol. 2020;131:123-32. doi:10.1016/j.jclinepi.2020.11.021.
12. Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
13. Schunemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE Guidelines: 22. The GRADE approach for tests and strategies - from test accuracy to patient important outcomes and recommendations. J Clin Epidemiol. 2019 doi:10.1016/j.jclinepi.2019.02.003.
14. Lim WS, Baudouin SV, George RC, Hill AT, Jamieson C, Le Jeune I, et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. Thorax. 2009;64 Suppl 3:iii1-55. doi:10.1136/thx.2009.121434.
15. International Agency for Research on Cancer. Cancer Today: Colon. Available from: https://gco.iarc.fr/today/data/factsheets/cancers/8-Colon-fact-sheet.pdf.
16. Soriano JB, Kendrick PJ, Paulson KR. revalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Respir Med. 2020;8(6):585-96. doi:10.1093/clinids/23.3.58.
17. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2010;182(18):E839-42. doi:10.1503/cmaj.090449.
18. AGREE Next Steps Consortium. The AGREE II Instrument [Electronic version]. 2017. Available from: http://www.agreetrust.org.

19. Ministry of Public Health. The diagnosis & management of community acquired pneumonia. Doha, Qatar Ministry of Public Health; 2019. Available from: https://www.moph.gov.qa/english/OurServices/eservices/Pages/Clinical-Guidelines.aspx.
20. Ewig S, Höffken G, Kern WV, Rohde G, Flick H, Krause R, et al. Behandlung von erwachsenen Patienten mit ambulant erworbener Pneumonie und Prävention. Bochum Deutschen Gesellschaft für Pneumologie und Beatmungsmedizin; 2016. Available from: https://www.awmf.org/uploads/tx_szleitlinien/020-020l_S3_ambulant_erworbene_Pneumonie_Behandlung_Praevention_2016-02-2.pdf.
21. Hill AT, Gold PM, El Solh AA, Metlay JP, Ireland B, Irwin RS, et al. Adult Outpatients With Acute Cough Due to Suspected Pneumonia or Influenza: CHEST Guideline and Expert Panel Report. Chest. 2019;155(1):155-67. doi:10.1016/j.chest.2018.09.016.
22. Rose MA, Barker M, Liese J, Adams O, Ankermann T, Baumann U, et al. [Guidelines for the Management of Community Acquired Pneumonia in Children and Adolescents (Pediatric Community Acquired Pneumonia, pCAP) - Issued under the Responsibility of the German Society for Pediatric Infectious Diseases (DGPI) and the German Society for Pediatric Pulmonology (GPP)]. Pneumologie. 2020;74(8):515-44. doi:10.1055/a-1139-5132.
23. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. S3-Leitlinie Kolorektales Karzinom. Berlin Leitlinienprogramm Onkologie der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF), Deutschen Krebsgesellschaft e.V. und Deutschen Krebshilfe; 2019. Available from: https://www.awmf.org/uploads/tx_szleitlinien/021-007OLl_S3_Kolorektales-Karzinom-KRK_2019-01.pdf.
24. Federatie Medisch Specialisten. Colorectaal carcinoom. Utrecht, the Netherlands Federatie Medisch Specialisten; 2019. Available from: https://richtlijnendatabase.nl/richtlijn/colorectaal_carcinoom_crc/startpagina_-_crc.html.
25. Argiles G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2020;31(10):1291-305. doi:10.1016/j.annonc.2020.06.022.
26. Cunningham C, Leong K, Clark S, Plumb A, Taylor S, Geh I, et al. Association of Coloproctology of Great Britain & Ireland (ACPGBI): Guidelines for the Management of Cancer of the Colon, Rectum and Anus (2017) - Diagnosis, Investigations and Screening. Colorectal Dis. 2017;19 Suppl 1:9-17. doi:10.1111/codi.13703.
27. Nederlands Huisartsen Genootschap - Dutch College of General Practitioners. Rectaal bloedverlies (M89). 2017. Available from: https://www.nhg.org/standaarden/samenvatting/rectaal-bloedverlies.
28. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. Nationale VersorgungsLeitlinie Asthma - Langfassung. 4. Auflage. 2020. Available from: https://register.awmf.org/assets/guidelines/nvl-002l_S3_Asthma_2020-09.pdf. 10.6101/AZQ/000469.
29. Cloutier MM, Dixon AE, Krishnan JA, Lemanske RF, Jr., Pace W, Schatz M. Managing Asthma in Adolescents and Adults: 2020 Asthma Guideline Update From the National Asthma Education and Prevention Program. Jama. 2020;324(22):2301-17. doi:10.1001/jama.2020.21974.
30. National Institute for Health and Care Excellence. Asthma: diagnosis and monitoring of asthma in adults, children and young people. 2017. Available from: https://www.nice.org.uk/guidance/ng80.
31. Ministry of Public Health Qatar. National Clinical Guideline: The diagnosis and management of asthma in adults. 2019. Available from: https://www.moph.gov.qa/_layouts/download.aspx?SourceUrl=/Admin/Lists/ClinicalGuidelinesAttachments/Attachments/18/The%20Diagnosis%20and%20Management%20of%20Asthma%20in%20Adults.pdf.
32. Scottish Intercollegiate Guidelines Network. SIGN 158. British guideline on the management of asthma. 2019. Available from: https://www.sign.ac.uk/media/1773/sign158-updated.pdf.
33. Ministry of Public Health Qatar. National Clinical Guideline: The diagnosis and management of asthma in children. 2019. Available from: https://www.moph.gov.qa/_layouts/download.aspx?SourceUrl=/Admin/Lists/ClinicalGuidelinesAttachments/Attachments/19/The%20Diagnosis%20and%20Management%20of%20Asthma%20in%20Children.pdf.
34. Arevalo-Rodriguez I, Moreno-Nunez P, Nussbaumer-Streit B, Steingart KR, Gonzalez Pena LDM, Buitrago-Garcia D, et al. Rapid reviews of medical tests used many similar methods to systematic

reviews but key items were rarely reported: a scoping review. J Clin Epidemiol. 2019;116:98-105. doi:10.1016/j.jclinepi.2019.09.004.
35. Dixon C, Dixon PE, Sultan S, Mustafa R, Morgan RL, Murad MH, et al. Guideline Developers in the United States were Inconsistent in Applying Criteria for Appropriate GRADE Use. J Clin Epidemiol. 2020 doi:10.1016/j.jclinepi.2020.01.026.
36. Gopalakrishna G, Leeflang MM, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6(9):e010549. doi:10.1136/bmjopen-2015-010549.
37. Norris SL, Meerpohl JJ, Akl EA, Schunemann HJ, Gartlehner G, Chen Y, et al. The skills and experience of GRADE methodologists can be assessed with a simple tool. J Clin Epidemiol. 2016;79:150-8.e1. doi:10.1016/j.jclinepi.2016.07.001.
38. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schünemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
39. INGUIDE. International Guideline Training and Certification Program. Available from: www.inguide.org.

Appendix 1. Search for CPGs

Search for CPGs with recommendations about CRP (in primary care patients with cough) or colonoscopy (in secondary care patients suspected of having colon cancer)

International Guideline Library (Guidelines International Network)

We searched the International Guideline Library, hosted by Guidelines International Network, on January 15th, 2021. We applied the following criteria:

- Publication scope: Diagnosis
- Countries of application: no restriction
- Guideline publication status: living guideline OR published
- Languages: English OR de OR en OR nl
- Authors: no restriction
- Publication year: 2016 OR 2017 OR 2018 OR 2019 OR 2020
- Willingness to collaborate: no restriction
- Name of endorsing member organisation: no restriction

Result: n=35 hits

We studied the retrieved results for information about:

- C-reactive protein test (CRP) to diagnose pneumonia in primary care patients with cough (excluding diagnostic procedures in patients suspected of having a COVID-19 infection)
- Colonoscopy to diagnose colon cancer in secondary care patients suspected of having (primary) colon cancer (excluding screening and tests in patients at risk of hereditary types of colon cancer)

This led to the inclusion of 2 possible relevant CPGs. After studying the full-text of the CPGs, we included one CPG in the final analysis [1].

Databases from the members of Guideline International Network

We searched the databases of all organisational GIN-members that stated to work in the field of guideline development on January 15th, 2021. Results after search and selection (in full-text of the CPG) are presented in the table below:

Table 1. CPGs about the use of CRP or colonoscopy, as retrieved by searching the websites of GIN-members active in guideline development

Member Organisation	CRP guidelines	Colonoscopy guidelines
(MOPH QA) Ministry of Public Health/Qatar	The diagnosis and management of community acquired pneumonia (2019) [2]	
(AACAP, US) American Academy of Child and Adolescent Psychiatry		
(ACIS) (ES) Scientific and Technical Advice Unit, avalia-t Galician Agency for Health Knowledge Management		
(Cochrane UK) Cochrane		
(DK) Danish Center for Clinical Practice Guidelines | Cancer		
(PHCC QA) Primary Health Care Corporation Qatar		
(WHO CH) World Health Organization		
AAFP (US) - American Academy of Family Physicians		
AAN (US) - American Academy of Neurology		
AAO HNSF (US) - American Academy of Otolaryngology - Head and Neck Surgery Foundation		
AAOS (USA) - American Academy of Orthopaedic Surgeons		
ACP (US) - American College of Physicians		
ACSQHC (AU) - Australian Commission on Safety and Quality in Health Care		Colonoscopy Clinical Care Standard (2020) [3]
AHRQ (US) - Agency for Healthcare Research and Quality		
AHTA (AU) - Adelaide Health Technology Assessment		
AMB (BR) - Brazilian Medical Association		
American Cancer Society		
American Society of Plastic Surgeons		
AND (US) - Academy of Nutrition and Dietetics		
APA (US) - American Psychological Association		
APTA (US) - American Physical Therapy Association		
AQuMed/AEZQ (DE) - German Agency for Quality in Medicine		
ASCO (US) - American Society of Clinical Oncology		
ASH (US) American Society of Hematology		

Member Organisation	CRP guidelines	Colonoscopy guidelines
AWMF (DE) - Association of Scientific Medical Societies	Behandlung von erwachsenen Patienten mit ambulant erworbener Pneumonie and Prävention (2016) [4]	Kolorektales Karzinom (2019) [5]
CADTH (CA)		
CAP (US) - College of American Pathologists		
Care Beyond Diagnosis		
CC (FI) - Current Care Guidelines / the Finnish Medical Society Duodecim		
CEBAM (BE) - Belgian Centre for Evidence-Based Medicin		
CEM (LU) - Cellule d'expertise médicale		
Center for Healthcare Quality Assessment and Control under the MoH of the Russian Federation (RU)		
Centers for Disease Control and Prevention		
Centre for Effective Practice		
CGS (DE) - User Group		
CHEST (US) - American College of Chest Physicians		
CISTERN		
Clinical Epidemiology and Evidence-Based Medicine (CEEBM) Unit, Cipto Mangunkusumo Hospital, Jakarta, Indonesia		
Cochrane Iberoamerica - INPECS		
CONITEC (BR) - National Committee for Health Technology Incorporation		
Covidence		
CSDS (LU) Conseil scientifique du domaine de la santé		
Czech health research council		
Department of Standardization of Chinese Medicine of Guangdong Provincial Hospital of Chinese Medicine		
DKG (DE) - German Cancer Society		See AWMF
DOH (IE) Department of Health		Diagnosis, staging and treatment of patients with colon cancer (2020) [6]
EBPracticenet (BE) Working Group		
EBSCO Health (Dynamed) (USA)		
ECRI Institute		

Member Organisation	CRP guidelines	Colonoscopy guidelines
Effective Basic Services Africa		
Endocrine Society		
ERWCPT (BE) - European Region of the World Confederation of Physical Therapy		
ESC (FR) - European Society of Cardiology		
ESCMID		
European Academy of Neurology		
Federal Institute for Quality Assurance and Transparency in Healthcare		
FMS (NL) - Federation of Medical Specialists		Colorectaal carcinoom (2019) [7]
G-I-N		
GOEG (AT) - Health Austria / Federal Institute for Quality in Health Care		
GRADE Working Group		
HAS (FR) - French National Authority for Health		
Hdir (NO) - Norwegian Directorate for Health		
HTA DoH (MY) - HTA Unit, Ministry of Health, Malaysia		
IACS (ES) - GuíaSalud-Aragon Institute of Health Sciences		
IETS (CO) - Institute of Technology Assessment in Health		
IKNL (NL) - Comprehensive Cancer Organisation, the Netherlands		
INEAS the national authority for assessment and accreditation in healthcare		
INESSS (CA) - Institut national d'excellence en santé et en services sociaux		
Institute of Health Data Science, Lanzhou University		
International Guidelines Center, Inc. (dba Guideline Central)		
JBI (AU) - Joanna Briggs Institute		
Kaiser Permanente, Care Management Institute		
KAMS (KR) - Korean Academy of Medical Science		
KCE (BE) - Belgian Healthcare Knowledge Centre		
KNGF (NL) - Royal Dutch Society for Physical Therapy		
MAGIC - Making GRADE the Irresistible Choice		
McMaster University (CA)		

Member Organisation	CRP guidelines	Colonoscopy guidelines
Minds Center (JP) - Medical Information Network Distribution Service Center, Japan Council for Quality Health Care		
MoH (UA) - The State Expert Center, Ministry of Health, Ukraine		
National Blood Authority		
National Evidence based Healthcare Collaborating Agency (NECA)		
NBHW (SE) - The National Board of Health and Welfare		
Netherlands Society of Occupational Medicine (NVAB)		
NHFA (AU) - National Heart Foundation of Australia		
NHG (NL) - Dutch College of General Practitioners		Rectaal bloedverlies (2017) [8]
NHMRC (AU) - National Health and Medical Research Council		
NICE (UK) - National Institute for Health and Care Excellence		
NIPH (NO) - Norwegian Institute of Public Health		
OSTEBA (ES) - Basque Office for Health Technology Assessment		
Public Health Agency of Canada		
RER Assr (IT) - Regional Health and Social Care Agency Emilia Romagna		
RNAO (CA) - Registered Nurses' Association of Ontario		
SIGN (GB) - Scottish Intercollegiate Guidelines Network		
Society for Cardiovascular Angiography and Interventions		
SST (DK) - Danish Health Authority		
Stiftung Gesundheitswissen		
TGL (AU) - Therapeutic Guidelines Ltd		
The National Center for Evidence Based Health Practice, The Saudi Health Council		
Think Pink: Bahrain Breast Cancer Society		
University of South Australia		
Verpleegkundigen & Verzorgenden Nederland		
ZINL (NL) - National Health Care Institute		
ZZQ (DE) - Agency for Quality in Dentistry		

TRIP database

We searched the TRIP database on January 15th, 2021, with the following search strings:

- (cough or pneumonia [anywhere in the document]) and (CRP or protein or biomarker [anywhere in the document]), from: 2016 to: 2020, area: 'Primary care'; filter on guidelines. This resulted in 40 hits. Application of the eligibility criteria on this result, led to the exclusion of all 40 hits.
- (colonoscopy [anywhere in the document]), from: 2016 to: 2020; filter on guidelines. This resulted in 192 hits. Application of the eligibility criteria on this result, led to the exclusion of 187 hits. We included five CPGs in the final analysis [1, 5, 9-11].

Medline

We searched the Medline database, using Ovid Silverplatter on January 16th, 2021, with the following search strings:

- (cough OR pneumonia [Title]) AND (guideline* OR recommendation* [Title]). We limited the search results to publication years 2016-2020, and publications in English, German, or Dutch language. This resulted in 132 hits.
- (colorectal or (colon adj3 cancer) or (colon adj3 carcinoma) or colonoscopy [Title] AND (guideline* OR recommendation* [Title]. We limited the search results to publication years 2016-2020, and publications in English, German, or Dutch language. This resulted in 269 hits.

Screening of the results in the cough/pneumonia search resulted in the selection of 11 abstracts about cough/pneumonia and 11 abstracts about colonoscopy. These abstracts were studied in full-text CPGs. This led to the inclusion of three CPGs about cough/pneumonia [12-14], and three CPGs about colonoscopy [1, 5, 15].
The reasons for exclusion are stated in *tables 2* and *3*.

Table 2. List of excluded studies about CRP

First author, year of publication	Reason for exclusion
Cao, 2018 [16]	No recommendation about CRP testing
Correa, 2018 [12]	No recommendation about CRP testing
Grief, 2018 [17]	No recommendation about CRP testing
Jones, 2020 [18]	No recommendation about CRP testing
Kardos, 2019 [19]	No recommendation about CRP testing
Metlay, 2019 [20]	No recommendation about CRP testing
Moore, 2019 [21]	No recommendation about CRP testing
Wiersinga, 2018 [22]	No recommendation about CRP testing as a diagnostic test
Wilkes, 2016 [23]	No recommendation about CRP testing

Table 3. List of excluded studies about colonoscopy

First author, year of publication	Reason for exclusion
Benson, 2018 [24]	No recommendation about colonoscopy
Benton, 2016 [25]	The clinical practice guideline that is published about dates from 2015
Bisschops, 2019 [26]	No recommendation about the value of colonoscopy
Cubiella, 2018 [27]	Full text of the guideline is not published in English, German or Dutch
Ramage, 2016 [28]	The guideline is not about colon cancer
Read, 2016 [29]	This is not a guideline
Zehnbauer, 2017 [30]	No recommendation about colonoscopy

Search for CPGs with recommendations about FeNO (fraction nitric oxide in exhaled air) to diagnose (severe) asthma

International Guideline Library (Guidelines International Network)

We searched the International Guideline Library, hosted by Guidelines International Network, on February 23th, 2021. We applied the following criteria:

- Publication scope: Diagnosis
- Countries of application: no restriction
- Guideline publication status: living guideline OR published
- Languages: English OR de OR en OR nl
- Authors: no restriction
- Publication year: 2016 OR 2017 OR 2018 OR 2019 OR 2020
- Willingness to collaborate: no restriction
- Name of endorsing member organisation: no restriction

Result: n=45 hits

We studied the retrieved results for information about FeNO (fraction nitric oxide in exhaled air) as a marker of airway inflammation as a diagnostic tool in the diagnosis of severe asthma. This led to the inclusion of 2 possible relevant CPGs [31, 32].

Databases from the members of Guideline International Network

We searched the databases of all organisational GIN-members that stated to work in the field of guideline development on February 25th 2021. Results after search and selection (in full-text of the CPG) are presented in the table below:

Table 4. CPGs about the use of FeNO, as retrieved by searching the websites of GIN-members active in guideline development

Member Organisation	FeNO in asthma
(MOPH QA) Ministry of Public Health/Qatar	The diagnosis and management of asthma in adults (2019) (also in GIN database)[31] Diagnosis and management of asthma in children (2019) (also in GIN database) [32]
(AACAP, US) American Academy of Child and Adolescent Psychiatry	
(ACIS) (ES) Scientific and Technical Advice Unit, avalia-t. Galician Agency for Health Knowledge Management	
(Cochrane UK) Cochrane	
(DK) Danish Center for Clinical Practice Guidelines | Cancer	
(PHCC QA) Primary Health Care Corporation Qatar	
(WHO CH) World Health Organization	
AAFP (US) - American Academy of Family Physicians	
AAN (US) - American Academy of Neurology	
AAO HNSF (US) - American Academy of Otolaryngology - Head and Neck Surgery Foundation	
AAOS (USA) - American Academy of Orthopaedic Surgeons	
ACP (US) - American College of Physicians	
ACSQHC (AU) - Australian Commission on Safety and Quality in Health Care	
AHRQ (US) - Agency for Healthcare Research and Quality	
AHTA (AU) - Adelaide Health Technology Assessment	
AMB (BR) - Brazilian Medical Association	
American Cancer Society	
American Society of Plastic Surgeons	
AND (US) - Academy of Nutrition and Dietetics	
APA (US) - American Psychological Association	
APTA (US) - American Physical Therapy Association	
AQuMed/AEZQ (DE) - German Agency for Quality in Medicine	Nationale Versorgungsleitlinie Asthma, 2020 [33]
ASCO (US) - American Society of Clinical Oncology	
ASH (US) American Society of Hematology	
AWMF (DE) - Association of Scientific Medical Societies	Nationale Versorgungsleitlinie Asthma, 2020 (same as AQuMed/AEZQ) [33]
CADTH (CA)	

Member Organisation	FeNO in asthma
CAP (US) - College of American Pathologists	
Care Beyond Diagnosis	
CC (FI) - Current Care Guidelines / the Finnish Medical Society Duodecim	
CEBAM (BE) - Belgian Centre for Evidence-Based Medicine	
CEM (LU) - Cellule d'expertise médicale	
Center for Healthcare Quality Assessment and Control under the MoH of the Russian Federation (RU)	
Centers for Disease Control and Prevention	
Centre for Effective Practice	
CGS (DE) - User Group	
CHEST (US) - American College of Chest Physicians	
CISTERN	
Clinical Epidemiology and Evidence-Based Medicine (CEEBM) Unit, Cipto Mangunkusumo Hospital, Jakarta, Indonesia	
Cochrane Iberoamerica - INPECS	
CONITEC (BR) - National Committee for Health Technology Incorporation	
Covidence	
CSDS (LU) Conseil scientifique du domaine de la santé	
Czech health research council	
Department of Standardization of Chinese Medicine of Guangdong Provincial Hospital of Chinese Medicine	
DKG (DE) - German Cancer Society	
DOH (IE) Department of Health	
Ebpracticenet (BE) Working Group Development of Primary Care Guidelines	
EBSCO Health (DynaMed) (USA)	
ECRI	
Effective Basic Services Africa	
Endocrine Society	
ERWCPT (BE) - European Region of the World Confederation of Physical Therapy	
ESC (FR) - European Society of Cardiology	
ESCMID	
European Academy of Neurology	

Member Organisation	FeNO in asthma
Federal Institute for Quality Assurance and Transparency in Healthcare	
FMS (NL) - Federation of Medical Specialists	
GIN	
GOEG (AT) - Health Austria / Federal Institute for Quality in Health Care	
GRADE Working Group	
HAS (FR) - French National Authority for Health	
Hdir (NO) - Norwegian Directorate for Health	
HTA DoH (MY) - HTA Unit, Ministry of Health, Malaysia	
IACS (ES) - GuíaSalud-Aragon Institute of Health Sciences	
IETS (CO) - Institute of Technology Assessment in Health	
IKNL (NL) - Comprehensive Cancer Organisation, the Netherlands	
INEAS the national authority for assessment and accreditation in healthcare	
INESSS (CA) - Institut national d'excellence en santé et en services sociaux	
Institute of Health Data Science, Lanzhou University	
International Guidelines Center, Inc. (dba Guideline Central)	Asthma: diagnosis, monitoring and chronic asthma management (National Guideline Centre), 2017 [34]
JBI (AU) - Joanna Briggs Institute	
Kaiser Permanente, Care Management Institute	
KAMS (KR) - Korean Academy of Medical Science	
KCE (BE) - Belgian Healthcare Knowledge Centre	
KNGF (NL) - Royal Dutch Society for Physical Therapy	
MAGIC [Making GRADE the Irresistible Choice] Evidence Ecosystem Foundation	
McMaster University (CA)	
Minds Center (JP) - Medical Information Network Distribution Service Center, Japan Council for Quality Health Care	
MoH (UA) - The State Expert Center, Ministry of Health, Ukraine	
National Blood Authority	
National Evidence based Healthcare Collaborating Agency (NECA)	
NBHW (SE) - The National Board of Health and Welfare	
Netherlands Society of Occupational Medicine (NVAB)	

Member Organisation	FeNO in asthma
NHFA (AU) - National Heart Foundation of Australia	
NHG (NL) - Dutch College of General Practitioners	Astma bij volwassenen, 2020 [35]
NHMRC (AU) - National Health and Medical Research Council	
NICE (UK) - National Institute for Health and Care Excellence	Asthma: diagnosis, monitoring and chronic asthma management, 2020 [34]
NIPH (NO) - Norwegian Institute of Public Health	
OSTEBA (ES) - Basque Office for Health Technology Assessment	
Public Health Agency of Canada	
RER Assr (IT) - Regional Health and Social Care Agency Emilia Romagna	
RNAO (CA) - Registered Nurses' Association of Ontario	
SIGN (GB) - Scottish Intercollegiate Guidelines Network	British guideline on the management of asthma, 2019 [36]
Society for Cardiovascular Angiography and Interventions	
SST (DK) - Danish Health Authority	
Stiftung Gesundheitswissen	
TGL (AU) - Therapeutic Guidelines Ltd	
The National Center for Evidence Based Health Practice, The Saudi Health Council	
Think Pink: Bahrain Breast Cancer Society	
University of South Australia	
Verpleegkundigen & Verzorgenden Nederland	
ZINL (NL) - National Health Care Institute	
ZZQ (DE) - Agency for Quality in Dentistry	

TRIP database

We searched the TRIP database on February 25th, 2021, with the following search strings:

- FeNO [anywhere in the document]), from: 2016 to: 2020, filter on guidelines. This resulted in 20 hits. Application of the eligibility criteria on this result, led to the exclusion of 18 hits. We included two CPGs [34, 36].

Medline

We searched the Medline database, using Ovid Silverplatter on February 26th, 2021, with the following search strings:

- (asthma [Title]) AND (guideline* OR recommendation* [Title]) AND (diagnos*[all fields]). We limited the search results to publication years 2016-2020, and publications in English, German, or Dutch language. This resulted in 61 hits. Four of these hits lead to relevant CPGs [37-40].

References

1. Argiles G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2020;31(10):1291-305. doi:10.1016/j.annonc.2020.06.022.
2. Ministry of Public Health. The diagnosis & management of community acquired pneumonia. Doha, Qatar Ministry of Public Health; 2019. Available from: https://www.moph.gov.qa/english/OurServices/eservices/Pages/Clinical-Guidelines.aspx.
3. Australian Commission on Safety and Quality in Health Care. Colonoscopy Clinical Care Standard. 2020. Available from: https://www.safetyandquality.gov.au/sites/default/files/2020-04/colonoscopy_clinical_care_standard_updated_2020.pdf.
4. Ewig S, Höffken G, Kern WV, Rohde G, Flick H, Krause R, et al. Behandlung von erwachsenen Patienten mit ambulant erworbener Pneumonie und Prävention. Bochum Deutschen Gesellschaft für Pneumologie und Beatmungsmedizin; 2016. Available from: https://www.awmf.org/uploads/tx_szleitlinien/020-020l_S3_ambulant_erworbene_Pneumonie_Behandlung_Praevention_2016-02-2.pdf.
5. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. S3-Leitlinie Kolorektales Karzinom. Berlin Leitlinienprogramm Onkologie der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF), Deutschen Krebsgesellschaft e.V. und Deutschen Krebshilfe; 2019. Available from: https://www.awmf.org/uploads/tx_szleitlinien/021-007OLl_S3_Kolorektales-Karzinom-KRK_2019-01.pdf.
6. Department of Health Ireland. Diagnosis, staging and treatment of patients with colon cancer. Dublin, Ireland 2020. Available from: https://www.gov.ie/en/publication/9cf17-diagnosis-staging-and-treatment-of-patients-with-colon-cancer/.
7. Federatie Medisch Specialisten. Colorectaal carcinoom. Utrecht, the Netherlands Federatie Medisch Specialisten; 2019. Available from: https://richtlijnendatabase.nl/richtlijn/colorectaal_carcinoom_crc/startpagina_-_crc.html.
8. Nederlands Huisartsen Genootschap - Dutch College of General Practitioners. Rectaal bloedverlies (M89). 2017. Available from: https://www.nhg.org/standaarden/samenvatting/rectaal-bloedverlies.
9. Oakland K, Chadwick G, East JE, Guy R, Humphries A, Jairath V, et al. Diagnosis and management of acute lower gastrointestinal bleeding: guidelines from the British Society of Gastroenterology. Gut. 2019;68(5):776-89. doi:10.1136/gutjnl-2018-317807.
10. Malaysia Health Technology Assessment Section (MaHTAS). Management of colorectal carcinoma, 2017. 2017. Available from: https://www.moh.gov.my/moh/resources/penerbitan/CPG/CPG%20Management%20of%20Colorectal%20%20Carcinoma.pdf.
11. ASGE Standards of Practice Committee, Gurudu SR, Bruining DH, Acosta RD, Eloubeidi MA, Faulx AL, et al. The role of endoscopy in the management of suspected small-bowel bleeding. Gastrointest Endosc. 2017;85(1):22-31. doi:10.1016/j.gie.2016.06.013.
12. Correa RA, Costa AN, Lundgren F, Michelin L, Figueiredo MR, Holanda M, et al. 2018 recommendations for the management of community acquired pneumonia. J Bras Pneumol. 2018;44(5):405-23. doi:10.1590/S1806-37562018000000130.
13. Hill AT, Gold PM, El Solh AA, Metlay JP, Ireland B, Irwin RS, et al. Adult Outpatients With Acute Cough Due to Suspected Pneumonia or Influenza: CHEST Guideline and Expert Panel Report. Chest. 2019;155(1):155-67. doi:10.1016/j.chest.2018.09.016.
14. Rose MA, Barker M, Liese J, Adams O, Ankermann T, Baumann U, et al. [Guidelines for the Management of Community Acquired Pneumonia in Children and Adolescents (Pediatric Community Acquired Pneumonia, pCAP) - Issued under the Responsibility of the German Society for Pediatric Infectious Diseases (DGPI) and the German Society for Pediatric Pulmonology (GPP)]. Pneumologie. 2020;74(8):515-44. doi:10.1055/a-1139-5132.
15. Cunningham C, Leong K, Clark S, Plumb A, Taylor S, Geh I, et al. Association of Coloproctology of Great Britain & Ireland (ACPGBI): Guidelines for the Management of Cancer of the Colon, Rectum and Anus (2017) - Diagnosis, Investigations and Screening. Colorectal Dis. 2017;19 Suppl 1:9-17. doi:10.1111/codi.13703.

16. Cao B, Huang Y, She DY, Cheng QJ, Fan H, Tian XL, et al. Diagnosis and treatment of community-acquired pneumonia in adults: 2016 clinical practice guidelines by the Chinese Thoracic Society, Chinese Medical Association. The clinical respiratory journal. 2018;12(4):1320-60. doi:10.1111/crj.12674.
17. Grief SN, Loza JK. Guidelines for the Evaluation and Treatment of Pneumonia. Prim Care. 2018;45(3):485-503. doi:10.1016/j.pop.2018.04.001.
18. Jones BE, Herman DD, Dela Cruz CS, Waterer GW, Metlay JP, Ruminjo JK, et al. Summary for Clinicians: Clinical Practice Guideline for the Diagnosis and Treatment of Community-acquired Pneumonia. Ann Am Thorac Soc. 2020;17(2):133-8. doi:10.1513/AnnalsATS.201909-704CME.
19. Kardos P, Dinh QT, Fuchs KH, Gillissen A, Klimek L, Koehler M, et al. [Guidelines of the German Respiratory Society for Diagnosis and Treatment of Adults Suffering from Acute, Subacute and Chronic Cough]. Pneumologie. 2019;73(3):143-80. doi:10.1055/a-0808-7409.
20. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. Am J Resp Crit Care Med. 2019;200(7):e45-e67. doi:10.1164/rccm.201908-1581ST.
21. Moore A, Harnden A, Grant CC, Patel S, Irwin RS, Panel CEC. Clinically Diagnosing Pertussis-associated Cough in Adults and Children: CHEST Guideline and Expert Panel Report. Chest. 2019;155(1):147-54. doi:10.1016/j.chest.2018.09.027.
22. Wiersinga WJ, Bonten MJ, Boersma WG, Jonkers RE, Aleva RM, Kullberg BJ, et al. Management of community-acquired pneumonia in adults: 2016 guideline update from the Dutch Working Party on Antibiotic Policy (SWAB) and Dutch Association of Chest Physicians (NVALT). Neth J Med. 2018;76(1):4-13.
23. Wilkes J. ACCP Provides Updated Recommendations on the Management of Somatic Cough Syndrome and Tic Cough. Am Fam Physician. 2016;93(5):416.
24. Benson AB, Venook AP, Al-Hawary MM, Cederquist L, Chen YJ, Ciombor KK, et al. NCCN Guidelines Insights: Colon Cancer, Version 2.2018. Journal of the National Comprehensive Cancer Network. 2018;16(4):359-69. doi:10.6004/jnccn.2018.0021.
25. Benton S, Steele R, Logan R, Djedovic N, Smith S, Addison C. NICE referral guidelines for suspected cancer: colorectal cancer and faecal occult blood testing. Ann Clin Biochem. 2016;53(Pt 1):7-9. doi:10.1177/0004563215612507.
26. Bisschops R, East JE, Hassan C, Hazewinkel Y, Kaminski MF, Neumann H, et al. Advanced imaging for detection and differentiation of colorectal neoplasia: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2019. Endoscopy. 2019;51(12):1155-79. doi:10.1055/a-1031-7657.
27. Cubiella J, Marzo-Castillejo M, Mascort-Roca JJ, Amador-Romero FJ, Bellas-Beceiro B, Clofent-Vilaplana J, et al. Clinical practice guideline. Diagnosis and prevention of colorectal cancer. 2018 Update. Gastroenterologia y hepatologia. 2018;41(9):585-96. doi:10.1016/j.gastrohep.2018.07.012.
28. Ramage JK, De Herder WW, Delle Fave G, Ferolla P, Ferone D, Ito T, et al. ENETS Consensus Guidelines Update for Colorectal Neuroendocrine Neoplasms. Neuroendocrinology. 2016;103(2):139-43. doi:10.1159/000443166.
29. Read AJ, Weissman A, Schoenfeld PS, Saini S, Menees SB, Saini SD. The Effect of Endoscopist Recommendations on PCPs Choosing Wisely about Colonoscopy. American Journal of Gastroenterology. 2016;111(5):749-51. doi:10.1038/ajg.2016.77.
30. Zehnbauer B, Temple-Smolkin R, Monzon FA. Guidelines for Colorectal Cancer Testing: Evidence-Based Practice Recommendations. J Mol Diagn. 2017;19(2):183-6. doi:10.1016/j.jmoldx.2017.01.002.
31. Ministry of Public Health Qatar. National Clinical Guideline: The diagnosis and management of asthma in adults. 2019. Available from: https://www.moph.gov.qa/_layouts/download.aspx?SourceUrl=/Admin/Lists/ClinicalGuidelinesAttachments/Attachments/18/The%20Diagnosis%20and%20Management%20of%20Asthma%20in%20Adults.pdf.
32. Ministry of Public Health Qatar. National Clinical Guideline: The diagnosis and management of asthma in children. 2019. Available from: https://www.moph.gov.qa/_layouts/download.aspx?SourceUrl=/Admin/Lists/ClinicalGuidelinesAttachments/Attachments/19/The%20Diagnosis%20and%20Management%20of%20Asthma%20in%20Children.pdf.

33. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. Nationale VersorgungsLeitlinie Asthma - Langfassung. 4. Auflage. 2020. Available from: https://register.awmf.org/assets/guidelines/nvl-002l_S3_Asthma_2020-09.pdf. 10.6101/AZQ/000469.
34. National Institute for Health and Care Excellence. Asthma: diagnosis and monitoring of asthma in adults, children and young people. 2017. Available from: https://www.nice.org.uk/guidance/ng80.
35. Bottema JW, Bouma M, Broekhuizen L, Chavannes NH, Frankemölle LAM, Hallensleben C, et al. NHG-Standaard Astma bij volwassenen (M27). Utrecht 2020. Available from: https://richtlijnen.nhg.org/standaarden/astma-bij-volwassenen.
36. Scottish Intercollegiate Guidelines Network. SIGN 158. British guideline on the management of asthma. 2019. Available from: https://www.sign.ac.uk/media/1773/sign158-updated.pdf.
37. Arakawa H, Hamasaki Y, Kohno Y, Ebisawa M, Kondo N, Nishima S, et al. Japanese guidelines for childhood asthma 2017. Allergol. 2017;66(2):190-204. doi:10.1016/j.alit.2016.11.003.
38. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention. 2020. Available from: www.ginasthma.org.
39. Ichinose M, Sugiura H, Nagase H, Yamaguchi M, Inoue H, Sagara H, et al. Japanese guidelines for adult asthma 2017. Allergol. 2017;66(2):163-89. doi:10.1016/j.alit.2016.12.005.
40. Cloutier MM, Dixon AE, Krishnan JA, Lemanske RF, Jr., Pace W, Schatz M. Managing Asthma in Adolescents and Adults: 2020 Asthma Guideline Update From the National Asthma Education and Prevention Program. Jama. 2020;324(22):2301-17. doi:10.1001/jama.2020.21974.

Appendix 2. Data extraction form 'Evidence in diagnostic recommendations in clinical practice guidelines'

Recommendation id:	Unique id
Recommendation:	Text of the recommendation
Data extraction by:	First author
Data extraction checked by:	Name of the author that checked the data
Question	Answer
General information recommendation	
Is the recommendation graded? (e.g. strong/weak (conditional), level A, level 1) *This should be stated in the text of the recommendation*	☐ Yes ☐ No ☐ Unclear
Direction of the recommendation *This should be stated in the text of the recommendation*	☐ For use of the diagnostic test ☐ Against use of the diagnostic test ☐ Unclear
What is the target condition that the recommendation is aimed at? *i.e. diagnosis, e.g. pneumonia, colorectal cancer*	Narrative
Does the recommendation concern a single test or does it concern multiple tests?	☐ Single test ☐ Multiple tests ☐ Unclear
General information clinical practice guideline *The information needed for these questions is typically found in general sections of the guideline (e.g. title page, introduction, methods section)*	
Scope of the CPG	☐ Diagnosis only ☐ Broader than diagnosis ☐ Unclear
Composition of the CPG panel	☐ Monodisciplinary: one professional stakeholder group is represented ☐ Multidisciplinary: several stakeholder groups are represented ☐ Unclear
Target audience of the CPG	☐ Monodisciplinary ☐ Multidisciplinary ☐ Unclear
Quality of the clinical practice guideline	
Did the authors of the CPG report use of the GRADE approach?	☐ Yes, this is literally stated ☐ No, not reported ☐ Unclear
Do you recognize elements of the GRADE approach in the CPG? *E.g. GRADE Evidence Profiles, Evidence-to-decision frameworks, rate the importance of outcome measures*	☐ Yes ☐ No ☐ Unclear

AGREE Domain methodology (items 7-12) *See https://www.agreetrust.org/wp-content/uploads/2017/12/AGREE-II-Users-Manual-and-23-item-Instrument-2009-Update-2017.pdf pages 20-25 for guidance on the following questions*	
Systematic methods were used to search for evidence	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree
The criteria for selecting the evidence are clearly described	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree
The strengths and limitations of the body of evidence are clearly described	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree
The methods for formulating the recommendations are clearly described	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree
The health benefits, side effects, and risks have been considered in formulating the recommendations	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree
There is an explicit link between the recommendations and the supporting evidence	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree

Other aspects of the quality of the guideline	
Patient involvement: The views and preferences of the target population (patients, public, etc.) have been sought *See https://www.agreetrust.org/wp-content/uploads/2017/12/AGREE-II-Users-Manual-and-23-item-Instrument-2009-Update-2017.pdf page 17 for guidance on this question*	☐ 1 – Strongly disagree ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 – Strongly agree
Methodologist involvement *This information can be found in the information about the CPG panel, in the introduction or methods section of the guideline, or in general information of the developer*	☐ Yes, in CPG panel ☐ No, not in CPG panel, but at bureau level or in technical team, etc ☐ No involvement ☐ Unclear
The evidence-base of the recommendation *This information can be found in the text or conclusions that are the basis of the recommendation*	
Diagnostic accuracy Defined as information about sensitivity, specificity, predictive values, false positives, false negatives, etc.	
Did the authors report diagnostic accuracy?	☐ Yes ☐ No → skip next two questions ☐ Unclear
If the answer on the previous question is YES: Is the information about diagnostic accuracy based on a systematic review of the literature?	☐ Yes ☐ No → skip next question ☐ Unclear
If the answer on the previous question is YES: Is the quality of the evidence about diagnostic accuracy assessed?	☐ Yes ☐ No ☐ Unclear
Test burden Defined as side effects of the test, complications of the test	
Did the authors report test burden?	☐ Yes ☐ No → skip next two questions ☐ Unclear
If the answer on the previous question is YES: Is the information about test burden based on a systematic review of the literature?	☐ Yes ☐ No → skip next question ☐ Unclear
If the answer on the previous question is YES: Is the quality of the evidence about test burden assessed?	☐ Yes ☐ No ☐ Unclear
Natural course *This is important to judge the value of the test, e.g. in case of mild natural course or a large proportion of false negative test results*	
Did the authors report natural course of the disease?	☐ Yes ☐ No → skip next two questions ☐ Unclear

If the answer on the previous question is YES: Is the information about natural course based on a systematic review of the literature?	☐ Yes ☐ No → skip next question ☐ Unclear
If the answer on the previous question is YES: Is the quality of the evidence about natural course assessed?	☐ Yes ☐ No ☐ Unclear
<u>Effectiveness of disease management</u> *E.g. treatment/therapy*	
Did the authors report effectiveness of disease management in the evaluation of the test?	☐ Yes ☐ No → skip next two questions ☐ Unclear
If the answer on the previous question is YES: Is the information about effectiveness of disease management based on a systematic review of the literature?	☐ Yes ☐ No → skip next question ☐ Unclear
If the answer on the previous question is YES: Is the quality of the evidence about effectiveness of disease management assessed?	☐ Yes ☐ No ☐ Unclear
<u>Linked evidence</u> *Defined as the link between test result and patient outcomes, expressed in studies that link diagnostic accuracy outcome measures (TP, FP, TN, FN) to patient important outcomes (eventually covered in the overall certainty in the evidence)*	
Did the authors report 'linked evidence'?	☐ Yes ☐ No ☐ Unclear
Final remarks	Narrative

Chapter 4.

Required knowledge for guideline panel members to develop healthcare related testing recommendations – a developmental study

Mariska Tuut
Jako Burgers
Hans de Beer
Patrick Bindels
Patrick Bossuyt
Jochen Cals
Mariska Leeflang
Reem Mustafa
Hester Rippen
Corinna Schaefer
Holger Schünemann
Trudy van der Weijden
Miranda Langendam

Journal of Clinical Epidemiology 2024; https://doi.org/10.1016/j.jclinepi.2024.111438

Abstract

Objective: To define the minimum knowledge required for guideline panel members (healthcare professionals and consumers) involved in developing recommendations about healthcare related testing.

Study design and setting: A developmental study with a multi-staged approach. We derived a first set of knowledge components from literature and subsequently performed semi-structured interviews with nine experts. We refined the set of knowledge components and checked it with the interviewees for final approval.

Results: Understanding the test-management pathway, e.g., how test results should be used in context of decisions about interventions, is the key knowledge component. The final list includes 26 items on the following topics: health question, test-management pathway, target population, test, test result, interpretation of test results & subsequent management, and impact on people-important outcomes. For each item, the required level of knowledge is defined.

Conclusion: We developed a list of knowledge components required for guideline panels to formulate recommendations on healthcare related testing. The list could be used to design specific training programs for guideline panel members when developing recommendations about tests and testing strategies in healthcare.

Plain language summary

Healthcare professionals and consumers need to have specific knowledge when they develop guidelines about testing. In this study we defined what guideline panel members need to know. This will help to create training for them. It is important to understand that testing is only useful if it has a positive impact on the people tested. The 26 defined knowledge components all relate to this.

Keywords: guidelines, healthcare related testing, methodology, education in guideline methods, test-management pathway, people-important outcomes

Introduction

In healthcare, tests are used to screen for a disorder or disease (such as mammography in asymptomatic women at risk of breast cancer) and to confirm or exclude a diagnosis (such as a haemoglobin test to diagnose anaemia). Other purposes of testing include risk assessment (e.g. weight, blood pressure and cholesterol measurements to determine the likelihood of getting a cardiovascular event) and monitoring: to follow-up patients with a known disease (such as checklists to monitor rehabilitation) [1]. The benefits and harms of testing will depend on population characteristics (context and setting, related to the pre-test probability of having a particular condition), test characteristics (such as sensitivity and specificity), testing process (e.g. burden), and the impact of management, guided by test results, on people-important outcomes (also called patient important outcomes), such as mortality, morbidity and quality of life [2].

Assessing the net benefit of healthcare related testing in daily practice is complex, and likely more complex than assessing those benefits for other interventions, such as treatment [3, 4]. Therefore, guidelines can be developed to provide support in decision making for healthcare professionals and consumers[5]. Trustworthy guidelines should be based on systematic reviews of the evidence (or systematically and transparently extracted evidence if scientific evidence is missing), should consider important population subgroups and people's values and preferences, should be based on an explicit and transparent process that minimizes biases, should provide a clear explanation of the relationship between policy options and health outcomes, should be explicit about the certainty of evidence and the strength of recommendations, and should be reconsidered in case of new evidence [5-7]. This includes critical appraisal of the evidence on testing beyond their clinical performance, in particular assessment of the impact on people-important outcomes, which appeared to be challenging [8-10]. Additionally, a lack of transparency in processing the evidence and considerations that lead to recommendations about healthcare related testing in guidelines was observed [11].

Facilitation of guideline development on testing could improve this process and might eventually lead to better guidelines and improvement of the quality of care. This not only applies to critical appraisal of the evidence but also to other essential aspects of guideline development, such as the formulation of questions, the definition of the role and purpose of a test, and the evidence-to-decision processes [12]. Currently available competency-based frameworks for guideline developers do not focus on expertise about test evaluation [13, 14]. Clarification about the knowledge needed for guideline

developers responsible for developing recommendations about healthcare related testing seems necessary.

Guidelines are typically developed by panels (also called guideline development groups), consisting of experts from different backgrounds and perspectives: (1) healthcare professionals, such as doctors and nurses, with expertise on the topic of interest, (2) healthcare consumers, such as patients, with experience in the topic of interest, (3) methodologists with specific expertise on guideline methodology and (4) chairs leading the guideline panel. Our study aimed to define the knowledge required for healthcare professionals and consumers in guideline panels (further referred to as guideline panel members) to adequately contribute to the development of guideline recommendations about healthcare related testing. The results of this study can inform specific training programs for guideline panel members involved in developing healthcare related testing recommendations.

Methods

Design

This was a developmental study with a multi-staged approach. We set up a project team (MKT, JSB, TvdW, MWL) to conduct the study. Based on an exploratory literature review, the project team drafted a first set of minimum knowledge components, and then conducted qualitative semi-structured interviews with experts to reflect and to comment on the set. Then, the refined final set of knowledge components was approved by the interviewees. This study was conducted between January 2022 and September 2023.

Literature review

The aim of the literature review was to collect potentially relevant knowledge components from existing literature. The first author of this study (MKT) systematically searched the literature, focusing on knowledge needed for guideline development and evidence-based medicine using Medline on January 7th, 2022, using the following search string:

> (((((("Evidence-Based Medicine"[Majr]) OR ("Evidence-Based Practice"[Majr])) OR (evidence based[Title])) OR ("Guidelines as Topic"[Majr])) OR (protocol*[Title])) OR (recommendation*[Title])) OR (guideline*[Title])) AND ((((((knowledge*[Title]) OR (competenc*[Title])) OR ("Knowledge"[Majr])) OR ("Health Knowledge, Attitudes, Practice"[Majr])) OR ("Professional Competence"[Majr])) OR ("Clinical Competence"[Majr])) with filters for studies published in the last 10 years, and studies published in English, Dutch or German language.

Next step was to select abstracts from potentially relevant studies about knowledge needed to develop guidelines using Rayyan [15]. Then, full-text articles of selected abstracts were retrieved, and analysed and relevant data were extracted. Additionally, relevant items of the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (draft version) and the GRADE (Grading of Recommendations Assessment, Development and Evaluation) Handbook were searched and adopted [16, 17]. Finally, 'snowballing' was applied to retrieve more relevant evidence. All steps of the literature review conducted by the first author (MKT) were discussed and approved by the other members of the project team (JSB, TvdW, MWL).

Creation of a draft list of knowledge components

Based on the results of the literature review, the first author (MKT) generated a draft list with specific knowledge components for the development of healthcare related testing recommendations in guidelines. General competencies for guideline development (such as the performance of a systematic review or group process techniques) and knowledge required to further develop guideline methods were not addressed. The draft list of knowledge components was discussed in detail within the project team for identifying overlapping items, deleting irrelevant items, and structuring the items into several groups.

Semi-structured interviews

The draft list of knowledge components was discussed in one-hour semi-structured online interviews. Relying on the Dutch and international network of the project team, we created a purposeful sample of nine internationally respected experts in the field of guideline development with specific knowledge and experience in:

- Using tests in clinical practice
- Test evaluation
- Guideline development and GRADE for tests
- Involvement of the public in guideline development, including testing recommendations
- Training about guideline development concerning tests

The interviewees received an information leaflet (*Appendix 1*) and the draft list (*Appendix 2*) beforehand. Two authors (MKT, MWL) undertook the interviews via Zoom, using an interview guide (*Appendix 3*) and tailoring the questions to the specific expertise of the interviewee. The interviews were videorecorded to facilitate data extraction and analysis.

To differentiate between levels of knowledge, we used a modified version of the cognitive domain of Bloom's taxonomy with the following levels [18]:

- Not necessary to know
- Remember (recall or recognize information)
- Understand (understand meaning, re-state data in one's own words)
- Apply (use or apply knowledge, put theory into practice)

We asked the interviewees to indicate a required level of knowledge for each discussed component.

Data analysis and creation of the final list of knowledge components

The first author of this study (MKT) used the video recordings of the interviews to select comments from the interviewees based on perceived relevance. This data extraction was checked by the last author of this study (MWL). Second, a member check was performed by sending the interview report to the interviewee asking for approval. Feedback from the first interviewee was incorporated in a subsequent version of the list of knowledge components and sent to the second interviewee. The third interviewee received the list including feedback from the first and second interviewee (and so on) to enable reflection on earlier comments.

Based on the feedback from the interviewees, potential changes were defined for adjustment (such as rephrasing, deleting, reordering, or combining components). All project group members commented on the interviewees' feedback and the proposals for adjustment and resolved conflicting feedback by reaching consensus through discussion. The final list of knowledge components was approved by the experts interviewed.

Results

Literature review

The literature search retrieved 3,299 potentially relevant studies. Of these, 3,290 were excluded based on the abstract, since the studies did not describe knowledge needed to develop guidelines. Nine articles were selected for further analysis [12-14, 19-24]. Furthermore, the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (draft version) and the GRADE Handbook provided additional information [16, 17]. Snowballing (of included articles and handbooks) resulted in the addition of eight articles including standards for reporting [25-32]. The literature selection process is presented in *figure 1*. A detailed description of the retrieved information is provided in *Appendix 4*.

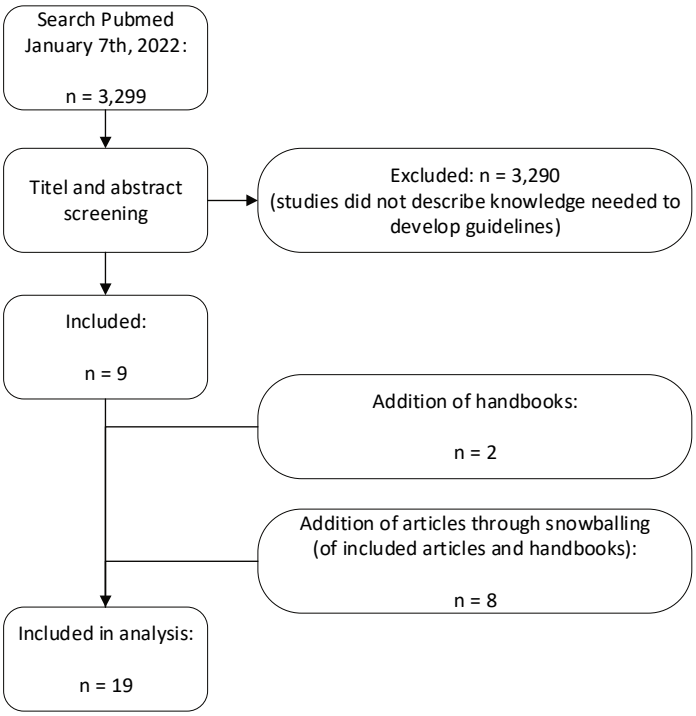


Figure 1. Literature flow

Draft list of components

Based on the literature review, we drafted a list of forty-one potentially relevant knowledge components as a starting point for the interviews. The items were categorized in three domains: diagnostic process in clinical practice, medical test evaluation, and clinical practice guideline development.

See *Appendix 2* for the draft list.

Interviews

All nine experts approached, agreed to be interviewed. Baseline characteristics of the interviewees are described in *table 1*.

The online interviews were conducted during the Summer and Autumn of 2022. All interviewees acknowledged the importance of determination of required knowledge components for developing healthcare related testing recommendations. General comments on the draft list concerned the phrasing and wording of the components, the need for a glossary, the structure of the list, and differentiation between core and detailed components.

Table 1. Characteristics of the interviewees

Inter-viewee	Background	Expertise	Country
1	Researcher and teacher	Test evaluation	The Netherlands
2	Researcher	Involvement of the public in guideline development, including testing recommendations	Germany
3	Guideline methodologist and teacher	▪ Guideline development and GRADE for tests ▪ Training about guideline development concerning tests	The Netherlands
4	Patient representative	Involvement of the public in guideline development, including testing recommendations	The Netherlands
5	Researcher, clinician and teacher	▪ Guideline development and GRADE for tests ▪ Training about guideline development concerning tests	USA
6	Researcher	Using tests in clinical practice	The Netherlands
7	Researcher, clinician and teacher	Using tests in clinical practice Test evaluation	The Netherlands
8	Researcher and teacher	Test evaluation	The Netherlands
9	Researcher and teacher	▪ Test evaluation ▪ Guideline development and GRADE for tests ▪ Training about guideline development concerning tests	Canada

There was no disagreement between the interviewees in their feedback on most of the components in the draft list. However, some components led to substantial feedback, particularly the components about Bayes' theorem, test burden, the performance of a test in specific circumstances, diagnostic accuracy as a surrogate outcome for people-important outcomes and the balance between desirable and undesirable consequences of a test. Feedback also concerned the importance of draft components, the formulation, combination or splitting of components and the required level of knowledge for guideline panel members. A summary of the interviewees' feedback is provided in *Appendix 5*.

Data analysis and creation of the final list

Based on the results of the interview, we reformulated frequently used terms to be more inclusive. For instance, by formulating: (a) 'guidelines' instead of 'clinical practice guidelines' and 'people-important outcomes' instead of 'patient important outcomes' to include the public health domain; (b) 'tests/testing' instead of 'diagnostic test' or 'medical test' to include self-testing, screening and settings other than medical; and (c) 'test-management pathway' instead of 'test-treatment pathway or strategy' to include no treatment, reassurance, or additional tests as subsequent management. We also added a glossary of terms and definitions.

After the interviews, we changed the structure of the knowledge components. The interviewees emphasized that tests do not stand on their own but are part of a test-management pathway and aim to influence the health of the people tested. Since understanding the concept of the test-management pathway as the key element, we restructured the knowledge components according to the test-management pathway. Additionally, we combined knowledge components to reduce the number of items and deleted knowledge components considered less relevant. We restricted the target group of the defined knowledge components to guideline panel members (defined as healthcare professionals and consumers involved), and stated that methodologists and chairs should have more in-depth knowledge.

In the final list of knowledge components, we defined 'recall', 'understand' and 'able to apply/able to interpret' as levels of required knowledge. The final list includes 26 knowledge components grouped into 7 domains (*box 1)*.

Discussion

Main findings

In this study, we defined knowledge components and the minimum level of knowledge required for the development of healthcare related testing recommendations in guidelines. Understanding the concept of the test-management pathway illustrated in *figure 2*, is the key component, connecting all knowledge components.

Strengths and limitations

This is the first study defining required knowledge for developing guideline recommendations about healthcare related testing. We believe these results fill a gap, since the development of healthcare related testing recommendations is complex and these recommendations often lack a focus on people-important outcomes [8, 10, 11]. Our study completes available frameworks for guideline developers and training programs (such as Dutch guideline courses and INGUIDE (International Guideline Training and Certification Program)) that do not yet focus on testing [13, 14].

This study has some limitations. First, the literature search was restricted to Medline and conducted in January 2022. Searching in additional databases could have resulted in more relevant articles. We tried to mitigate this by incorporating handbooks and snowballing. An update of the literature search, conducted in September 2023, revealed no new relevant publications. Inquiries with the interviewees indicated that they were not aware of any additional literature.

Box 1. Required knowledge for guideline panel members who develop recommendations about healthcare related testing

The development of guideline recommendations about healthcare related testing requires specific knowledge, in addition to the knowledge required to develop guidelines in general. *Figure 2* illustrates the key elements of the development of healthcare related testing recommendations in guidelines. This framework is used to structure the specific required knowledge components for guideline panel members[a] [12, 33]. The listed knowledge components have different levels of cognitive learning (recall, understand, able to interpret/formulate), according to Bloom's revised taxonomy [18].

Definitions:

- Test/testing: this includes all healthcare related tests and test strategies [34]. A medical test is a medical procedure performed to detect, diagnose, or monitor diseases, risks and treatment [35].
- Test-management pathway (also called test-treatment pathway, management pathway, care pathway, clinical pathway, test-management strategy, test-treatment strategy): a schematic pathway that includes all aspects in time related to the application of a healthcare related test and consequences for management that may follow such as (re)treatment, monitoring, side effects and complications as a result of testing and/or treatment [36].
- Target population: population eligible for the test, including the context in which the test is performed (such as earlier tests received) and setting (such as public health, primary care, secondary care).
- Burden: undesirable aspects of the test or treatment for healthcare consumers, patients, or caregivers (e.g., family) with psychosocial, physical, or practical impact, such as need to take medication, the inconvenience of visiting the doctor's office, financial costs, pain or anxiety.
- People-important outcome (also called patient important outcome, patient relevant outcome, patient relevant outcome measure, patient centred outcome): a component of a participant's clinical or functional status after an intervention has been applied that is used to assess the effectiveness of an intervention [16]. Depending on the condition of interest people-important outcomes may include consequences of having a certain test result (such as reassurance or labelling), consequences of the test or management (such as side effects (including adverse events and complications) and diagnostic and therapeutic yield), and society relevant outcomes (such as public health outcomes and costs).
- Modelling: decision analytic modelling, often undertaken when evidence is limited, involving prediction based on probabilities of possible outcomes (e.g., modelling the relation between pre-test probabilities, clinical performance/test accuracy, treatment strategies and people-important outcomes). This includes formal (complex, statistical) and informal ('back of the envelope') modelling [37].

Knowledge components:

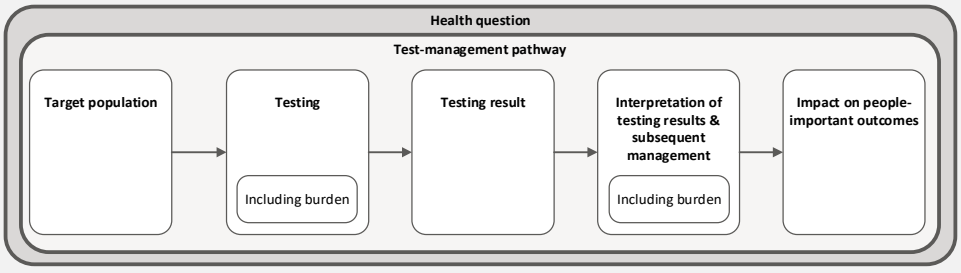


Figure 2. Analytical framework for the development of a healthcare related testing recommendation

Health question:

- Defining health question: The development of a testing recommendation starts with defining a health question. A guideline panel member is able to formulate a health question which includes definition of the target population, the test, the test-management pathway, and people-important outcomes.

Test-management pathway:

- Tests are part of test-management pathway: A guideline panel member understands that tests are part of a test-management pathway.[b]
- Test evaluation concepts: (In case the guideline development situation demands:) a guideline panel member can recall that tests are evaluated using the following concepts: analytical performance, clinical performance, clinical effectiveness, cost-effectiveness, and broader impact[c] [38].
- Clinical effectiveness of testing: A guideline panel member understands that the clinical effectiveness (desirable and undesirable health effects) of testing is determined by evaluating the test-management pathway.
- Direct and indirect evaluation: A guideline panel member can recall that a test-management pathway can be evaluated directly (in a diagnostic randomised controlled trial with sufficient follow-up to reach a change in people-important outcomes) and indirectly if direct evidence is lacking. Indirect evaluation of a test-management pathway includes assessment of all components of the test-management pathway.
- Certainty assessment: A guideline panel member can recall that evaluation of a test-management pathway includes assessment of the certainty of the evidence of the (components of the) pathway.
- Balance of desirable and undesirable consequences: A guideline panel member understands that evaluation of a test-management pathway includes consideration of the balance between desirable and undesirable consequences related to the test.
- Modelling: A guideline panel member can recall that desirable and undesirable consequences related to the test should be balanced by modelling.

Target population:

- Pre-test probability: A guideline panel member understands that each target population has a (specific) pre-test probability (related to the context and setting in which the test is performed)[d] of having a certain condition (e.g. a disease) [39].

Test:

- Purpose: Tests can be used for different purposes: screening or surveillance, risk assessment, classification and staging, diagnosis, treatment decisions, treatment monitoring, and estimating prognosis [34].
  A guideline panel member is able to formulate the purpose of the test of interest.
- Role: New tests can have four main roles: replacement of a test, triage, add-on or parallel/combined [34].
  A guideline panel member is able to formulate the role of the test of interest.
- Test burden: A guideline panel member is able to formulate burden, side effects and societal costs related to the test of interest.

Test result:

- Test accuracy as informative step: A guideline panel member understands that, in the absence of direct evidence, evaluation of the clinical performance/test accuracy is an informative and essential step to be able to determine the impact of a test on people-important outcomes.
- Clinical performance in target population: A guideline panel member understands that clinical performance/test accuracy depends on the target population (including context and setting) in which the test is used and evaluated.[e]
- Post-test probability: A guideline panel member can recall that the post-test probability of having a certain condition depends on the pre-test probability and the test result.
- Threshold for test-positivity: A guideline panel member can recall that a cut-off point of a certain test result determines the threshold for test positivity (i.e., the test result deviates from normal), and that changing the threshold for test positivity affects sensitivity and specificity (in opposite direction) as well as management following the test results. This may relate to over- and underdiagnosis [40].
- Test results: A guideline panel member understands that test results can be true positive, true negative, false positive, false negative and inconclusive (neither positive, nor negative). In guideline development, it is important to focus on these outcomes rather than sensitivity and specificity.
- Incorrect classification: A guideline panel member understands that false positive and false negative test results are related to incorrect classification of (not) having the condition of interest.

- Interpreting false positives/negatives: A guideline panel member is able to interpret false positive and false negative test results in terms of people-important outcomes.
- Testing reflects moment: A guideline panel member understands that the test result relates to the moment in time when the test was taken, meaning that test results can vary over time within a person.

Interpretation of test results & subsequent management:

- Management following testing: A guideline panel member understands that a test can lead to additional tests and/or treatment and/or other management, depending on the test result.
- Link between test result and management: A guideline panel member understands that not all patients with a specific test result get the recommended management.
- Management burden: A guideline panel member is able to formulate the burden, side effects and societal costs of the management of interest.

Impact on people-important outcomes:

- Testing aim: A guideline panel member understands that the principle aim of a test is to improve people-important outcomes and/or to reduce deterioration of people-important outcomes.
- Direct and indirect impact of management: A guideline panel member understands that management following a test result may directly or indirectly affect people-important outcomes.
- Management effectiveness: A guideline panel member understands that management following a test result may or may not be effective in the improvement or prevention of deterioration of people-important outcomes.

Footnotes:

a. Several experts are involved in guideline development: healthcare professionals who have expertise on the guideline topic of interest, healthcare consumers who have experience in the guideline topic of interest, methodologists who have expertise on guideline methodology and chairs leading the guideline panel. In this document, a guideline panel member refers to healthcare professionals and healthcare consumers in a guideline panel.

b. This relates to the assumption that a test is performed in a certain context, and is usually followed by actions, such as clinical management, self-management, or watchful waiting.

c. The ability to correctly detect or monitor a measurand is called the analytical performance of a test; This is evaluated by parameters such as trueness, validity, imprecision, limits of detection and cross-reactivity.
The ability to correctly classify those with and without the target condition is called the clinical performance of a test (also called test accuracy or diagnostic accuracy); this can be determined by comparing the index test (test of interest) with a reference test (also called reference standard) and is evaluated in a 2x2 table in which people with and without the target condition are classified according to their test result and parameters such as sensitivity and specificity, true positives, true negatives, false positives, false negatives, and inconclusive results.
The ability of a test to improve people-important outcomes is called the clinical effectiveness of a test (also called clinical utility). The clinical effectiveness is determined by evaluating the test-management pathway.
The evaluation of the balance between a change in people-important outcomes and costs due to the introduction of a test is called the cost-effectiveness of a test.
The evaluation of consequences of introducing or using a test beyond clinical effectiveness and cost-effectiveness is called the broader impact of a test. This is evaluated by parameters such as acceptability (including robustness of a test in practice), feasibility and implementability.

d. The pre-test probability is the proportion of people in the population at risk who have the condition of interest at a specific time or time interval, i.e., the point prevalence or the period prevalence of the condition of interest. Pre-test probabilities may be estimated from routine data, practice data or clinical judgement.

e. Test characteristics (such as sensitivity and specificity) vary between populations.

Second, the literature review, including data extraction and generation of a draft list of knowledge components, was conducted by a single author. This may have introduced random error and risk of bias. However, the project team closely monitored this process and discussed the draft list of knowledge components in detail. Furthermore, the interviewees were invited to supplement the draft knowledge components with their own knowledge, derived from published evidence as well as from their own expertise and experience. Consequently, it is unlikely that the review would contain significant gaps in the published knowledge base.

Third, we included a purposeful sample of nine experts in the field. However, these experts are internationally respected opinion leaders. They enlightened the topic of the study from different perspectives (clinical, consumer, researcher, methodologist, and teaching). Besides that, the experts complemented each other, and data saturation was reached. This resulted in a sound set of knowledge components required to develop healthcare related testing recommendations in guidelines.

Fourth, we used a modified version of the cognitive domain of Bloom's taxonomy. This domain consists of six levels (remember, understand, apply, analyse, evaluate, create) of which we used the first three and added a level zero 'not necessary to know'. We deemed this sufficient since the scope of this study is restricted to guideline development and does not include the guideline methodology development, which would have required the latter levels of knowledge. We considered using other methods of distinguishing knowledge, such as 'entrustable professional activities (EPA)' and CanMeds [41, 42]. However, these methods highlight knowledge and competencies from the healthcare professionals' perspective, which might put healthcare consumers in second place.

Implications for practice

Guideline panel members are often trained before or during their participation in a guideline panel. Examples are the INGUIDE course for guideline panel members and Dutch training programs about evidence-based guideline development [43]. We suggest extending available training programs with a module about the development of healthcare related testing recommendations. It can be confirmed that this work will be used to develop dedicated modules for INGUIDE (inguide.org), a Guidelines International Network led guideline credentialing and certification program. Such modules can be developed based on our findings and the scientific principles of developing educational courses.

Implications for research

This study identified required knowledge components for guideline panel members to develop healthcare related testing recommendations in guidelines. Next steps are (a) to assess the variation in knowledge of guideline panel members in order to offer suitable training and (b) to define required knowledge for guideline methodologists and guideline panel chairs involved in developing healthcare related testing recommendations.

Conclusion

This study defined the minimum knowledge required for guideline panel members involved in the development of guideline recommendations about healthcare related testing. The key component is the test-management pathway concept, which helps focussing on people-important outcomes. Other required components, such as the ability to formulate a health question concerning the benefit of a test, fit in this concept. The results of this study provide input to design specific training programs for guideline panel members when developing healthcare related testing recommendations.

References

1. Bossuyt PM, Deeks JJ, Leeflang MM, Takwoingi Y, Fleming E. Preface. 2023 In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 20 (update July 2023). Cochrane. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
2. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jama. 1994;271(5):389-91. doi:10.1001/jama.271.5.389.
3. Noguchi Y, Matsui K, Imura H, Kiyota M, Fukui T. Quantitative evaluation of the diagnostic thinking process in medical students. J Gen Intern Med. 2002;17(11):839-44. doi:10.1046/j.1525-1497.2002.20139.x.
4. Elstein AS. Thinking about diagnostic thinking: a 30-year perspective. Adv Health Sci Educ Theory Pract. 2009;14 Suppl 1:7-18. doi:10.1007/s10459-009-9184-0.
5. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Clinical Practice Guidelines We Can Trust. In: Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US). Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011 isbn: doi:10.17226/13058.
6. Schunemann HJ, Lerda D, Dimitrova N, Alonso-Coello P, Grawingholt A, Quinn C, et al. Methods for Development of the European Commission Initiative on Breast Cancer Guidelines: Recommendations in the Era of Guideline Transparency. Annals of internal medicine. 2019;171(4):273-80. doi:10.7326/M18-3445.
7. Regional Office for Europe World Health Organisation. (eds.). Strengthening countries' capacities to adopt and adapt evidence-based guidelines: a handbook for guideline contextualization. Copenhagen. World Health Organisation; 2023. Available from: https://iris.who.int/bitstream/handle/10665/372275/9789289060028-eng.pdf?sequence=1&isAllowed=y.
8. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014;67(7):760-8. doi:10.1016/j.jclinepi.2014.01.006.
9. Gopalakrishna G, Leeflang MM, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6(9):e010549. doi:10.1136/bmjopen-2015-010549.
10. Tuut MK, de Beer JJA, Burgers JS, van de Griendt EJ, van der Weijden T, Langendam MW. Applying GRADE for diagnosis revealed methodological challenges: an illustrative example for guideline developers. J Clin Epidemiol. 2020;131:123-32. doi:10.1016/j.jclinepi.2020.11.021.
11. Tuut MK, Burgers JS, van der Weijden T, Langendam MW. Do clinical practice guidelines consider evidence about diagnostic test consequences on patient-relevant outcomes? A critical document analysis. Journal of evaluation in clinical practice. 2021 doi:10.1111/jep.13619.
12. Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
13. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schunemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
14. Norris SL, Meerpohl JJ, Akl EA, Schünemann HJ, Gartlehner G, Chen Y, et al. The skills and experience of GRADE methodologists can be assessed with a simple tool. J Clin Epidemiol. 2016;79:150-8.e1. doi:10.1016/j.jclinepi.2016.07.001.
15. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Systematic reviews. 2016;5(1):210. doi:10.1186/s13643-016-0384-4.
16. GRADE Working Group. GRADE Handbook. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013. Available from: https://gdt.gradepro.org/app/handbook/handbook.html.
17. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (eds.). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (update July 2023). Cochrane; 2023. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.

18. Anderson LW, Krathwohl DR, Airasian PW, Cruikshank KA, Mayer RJ, Pintrich PR, et al. (eds.). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman; 2001. isbn:080131903X.
19. Albarqouni L, Hoffmann T, Straus S, Olsen NR, Young T, Ilic D, et al. Core Competencies in Evidence-Based Practice for Health Professionals: Consensus Statement Based on a Systematic Review and Delphi Survey. JAMA Netw Open. 2018;1(2):e180281. doi:10.1001/jamanetworkopen.2018.0281.
20. Hinneburg J, Hecht L, Berger-Höger B, Buhse S, Lühnen J, Steckelberg A. Development and piloting of a blended learning training programme for physicians and medical students to enhance their competences in evidence-based decision-making. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2020;150-152:104-11. doi:10.1016/j.zefq.2020.02.004.
21. Wieringa S, Dreesens D, Forland F, Hulshof C, Lukersmith S, Macbeth F, et al. Different knowledge, different styles of reasoning: a challenge for guideline development. BMJ Evid Based Med. 2018;23(3):87-91. doi:10.1136/bmjebm-2017-110844.
22. Berger B, Gerlach A, Groth S, Sladek U, Ebner K, Mühlhauser I, et al. Competence training in evidence-based medicine for patients, patient counsellors, consumer representatives and health care professionals in Austria: a feasibility study. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2013;107(1):44-52. doi:10.1016/j.zefq.2012.11.013.
23. Messerli FH, Hofstetter L, Agabiti-Rosei E, Burnier M, Elliott WJ, Franklin SS, et al. Expertise: no longer a sine qua non for guideline authors? J Hypertens. 2017;35(8):1564-6. doi:10.1097/hjh.0000000000001435.
24. Zuiderent-Jerak T, Forland F, Macbeth F. Guidelines should reflect all knowledge, not just clinical trials. Bmj. 2012;345:e6702. doi:10.1136/bmj.e6702.
25. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92. doi:10.1136/bmj.332.7549.1089.
26. Dickersin K. The existence of publication bias and risk factors for its occurrence. Jama. 1990;263(10):1385-9.
27. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. Jama. 1998;279(4):281-6. doi:10.1001/jama.279.4.281.
28. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, and the P-DTAG, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. Jama. 2018;319(4):388-96. doi:10.1001/jama.2017.19163.
29. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25. doi:10.1186/1471-2288-3-25.
30. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36. doi:10.7326/0003-4819-155-8-201110180-00009.
31. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009;64(8):1109-16. doi:10.1111/j.1398-9995.2009.02083.x.
32. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.
33. Schunemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2014;186(3):E123-42. doi:10.1503/cmaj.131237.
34. Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, et al. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PLoS One. 2015;10(10):e0134553. doi:10.1371/journal.pone.0134553.
35. Medical test. Wikipedia. Available from: https://en.wikipedia.org/wiki/Medical_test.
36. Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Guidelines for guideline developers: a systematic review of grading systems for medical tests. Implement Sci. 2013;8:78. doi:10.1186/1748-5908-8-78.

37. Schunemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82. doi:10.1016/j.jclinepi.2019.02.003.
38. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta. 2014;427:49-57. doi:10.1016/j.cca.2013.09.018.
39. Ruf M, Morgan O, Mackenzie K. Diagnosis and screening - Pre- and post-test probability. London UK Faculty of Public Health; 2017updated 2017. Available from: https://www.healthknowledge.org.uk/content/pre-and-post-test-probability.
40. Doust J, Vandvik PO, Qaseem A, Mustafa RA, Horvath AR, Frances A, et al. Guidance for Modifying the Definition of Diseases: A Checklist. JAMA internal medicine. 2017;177(7):1020-5. doi:10.1001/jamainternmed.2017.1302.
41. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. Med Teach. 2015;37(11):983-1002. doi:10.3109/0142159X.2015.1060308.
42. Royal College of Physicians and Surgeons of Canada. CanMEDS 2015 Physician Competency Framework. Ottawa Royal College of Physicians and Surgeons of Canada; 2015. Available from: https://canmeds.royalcollege.ca/uploads/en/framework/CanMEDS%202015%20Framework_EN_Reduced.pdf.
43. INGUIDE. International Guideline Training and Certification Program. Available from: www.inguide.org.

Appendix 1. Information leaflet for interviewees

Required knowledge to develop diagnostic test recommendations in clinical practice guidelines

This study aims to define the minimum knowledge components needed for guideline developers when making recommendations about diagnostic tests and to measure current knowledge among guideline developers. Currently available competencies and competency-based frameworks for guideline developers do not include diagnostic test evaluation [1, 2]. The results of this study provide input for designing specific training programs for guideline panel members when developing diagnostic test recommendations.

Definitions and concepts

We use the following definitions and concepts:

- Clinical practice guidelines: Clinical practice guidelines are statements that include recommendations intended to optimize patient care. To be trustworthy according to the Institute of Medicine, guidelines should:
  - be based on a systematic review of the existing evidence;
  - be developed by a knowledgeable, multidisciplinary panel of experts and representatives from key affected groups;
  - consider important patient subgroups and patient preferences, as appropriate;
  - be based on an explicit transparent process that minimizes distortions, biases, and conflicts of interest;
  - provide a clear explanation of the logical relationships between alternative care options and health outcomes and provide ratings of both the quality of evidence and the strength of recommendations;
  - be reconsidered and revised as appropriate when important new evidence warrants modifications of recommendations [3].
- Diagnostic test: Diagnostic tests are medical tests undertaken in patients who present to health services with signs or symptoms. A medical test refers to any procedure performed on a person's fluids, cells, tissue, or on the person themself, to detect, diagnose or monitor a condition or the course of a condition. Medical tests come in many different forms, from patient history and physical and visual examination to lab tests and imaging, as well as risk scores that combine multiple pieces of information from different sources [4].
- Diagnostic process: The diagnostic process is an empirical iterative process [5]. It has inductive and deductive elements, based on Bayes' theorem [6].

- Test-treatment pathway:

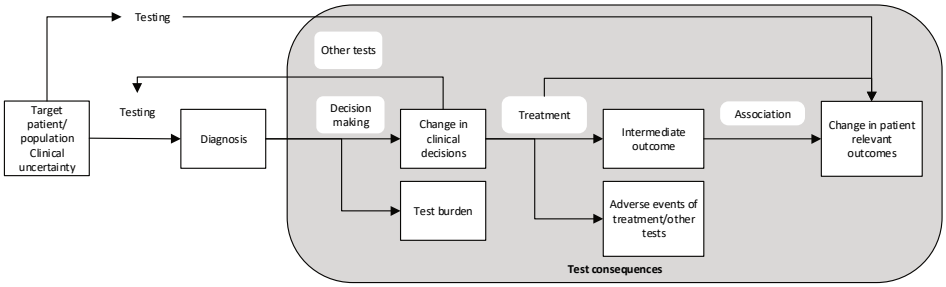


Figure 1. Schematic test-treatment pathway (adapted from Harris et al.) [7]

Brief description of this study

We created a draft list of knowledge components required to adequately develop diagnostic test recommendations. We reviewed current evidence, such as existing guideline competency frameworks, methodological literature about the evaluation of medical tests and clinical practice guidelines, and handbooks about test evaluation and clinical guideline development.

The next steps are review of the draft list by experts in the field of diagnostic test evaluation and guideline development and interviews of these experts. The results of the interviews will be incorporated in a new version of the list with knowledge components. This list will be sent out in an internet survey to a broader group, in which participants will be asked to score each knowledge component per role (health care provider, health care consumer representative, methodologist, guideline panel chair) in the clinical guideline panel.

Interviews

We plan to interview experts in the field who are specialized in specific domains:

- Diagnostic process in clinical practice
- Diagnostic test evaluation
- Clinical practice guideline development and GRADE for medical tests
- Patient involvement in clinical practice guideline development
- Training about clinical guideline development concerning medical tests

Interview items

Please note:

- The focus of this study is specifically on the development of diagnostic test recommendations in clinical guideline development. General competencies

required for clinical guideline development, such as the performance of a systematic review or group process techniques, are outside the scope of this study.

- This study focuses on the development of a clinical practice guideline in a guideline panel. Knowledge required to develop or improve guideline methods, such as improvement of GRADE methodology, is outside the scope of this study.
- This study focuses on knowledge components, which might be required in different levels to develop diagnostic test recommendations. We used a modified version of Bloom's taxonomy to distinguish between the levels of knowledge which might be required. A simple illustration of Bloom's taxonomy is given below:

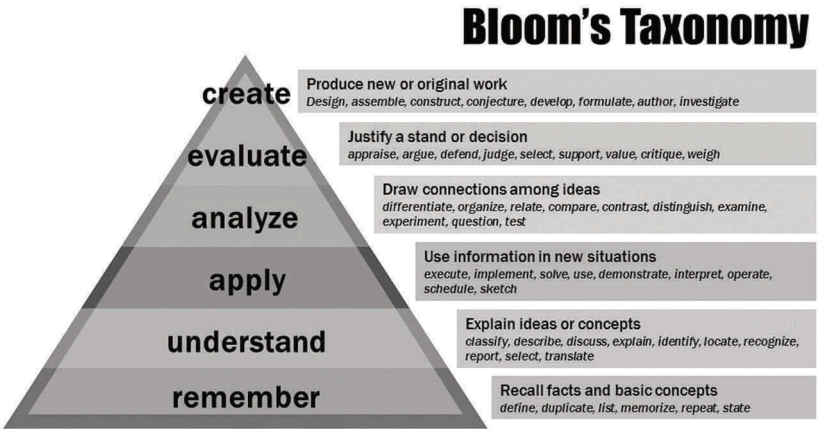


Figure 2. Bloom's taxonomy, created by the Vanderbilt University Center for Teaching. Compared to Bloom's taxonomy we added one level prior to the first level (remember): 'not necessary to know'.

References

1. Norris SL, Meerpohl JJ, Akl EA, Schunemann HJ, Gartlehner G, Chen Y, et al. The skills and experience of GRADE methodologists can be assessed with a simple tool. J Clin Epidemiol. 2016;79:150-8.e1. doi:10.1016/j.jclinepi.2016.07.001.
2. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schunemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
3. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Clinical Practice Guidelines We Can Trust. In: Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US). Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011 isbn: doi:10.17226/13058.
4. Deeks JJ, Bossuyt PM. Chapter 2: Evaluating medical tests. 2023 In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 20 (updated July 2023). Cochrane. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
5. Norman G, Barraclough K, Dolovich L, Price D. Iterative diagnosis. BMJ. 2009;339:b3490. doi:10.1136/bmj.b3490.
6. Wulff HR. (eds.). Principes van klinisch denken en handelen; Nederlandse bewerking. Utrecht: Bohn, Scheltema & Holkema; 1980. isbn:90 313 0399 2.
7. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21-35. doi:10.1016/s0749-3797(01)00261-6.

Appendix 2. Draft list 'Required knowledge to develop medical test recommendations in clinical practice guidelines'

Possible required knowledge components ↓	Role in guideline panel			
	Health care provider	Health care consumer	Methodologist	Guideline panel chair
Diagnostic process in clinical practice				
The diagnostic process has inductive elements (to generate a general diagnosis, hypothesis generation) and deductive elements (to confirm or rule out a specific diagnosis, hypothesis testing)				
The diagnostic process is a combination of sense (e.g., common sense) and science (e.g., scientific evidence)				
The diagnostic process is based on Bayes' theorem; Bayes' theorem states that the probability of a particular chance depends on a priori chances (here: pre-test probability) and the occurrence of events (here: test result)				
Clinical experience, including gut feelings, is essential in patient care				
Medical tests can have different purposes (to confirm or exclude a clinical diagnosis, to test the likelihood of a clinical diagnosis, for follow-up of patients)				
Medical tests are part of a test-treatment pathway				
The aim of a test is to improve patient relevant outcomes				
In general, several steps are essential to move from medical test to patient relevant outcome (test-treatment pathway)				
A test rarely is 100% accurate				
Test results can be true positive, true negative, false positive, false negative, and inconclusive				
A test might have side effects, adverse events and complications and can lead to stress (test burden)				
A test has costs, which may be direct and/or indirect and medical and/or non-medical				
A test may have acceptability issues, such as preparation by the patient and travel and waiting time				
A test result can lead to additional tests and/or treatment				
Not all patients with a specific (positive or negative) test result get the recommended follow-up test or treatment				
Treatment following a test result may directly influence patient relevant outcomes				

Possible required knowledge components ↓	Role in guideline panel			
	Health care provider	Health care consumer	Methodologist	Guideline panel chair
Treatment following a test result may influence surrogate/intermediate outcomes, related to patient relevant outcomes				
Treatment following a test result may or may not be effective in improving patient relevant outcomes				
Treatment following a test result may have side effects, adverse events, and complications				
Treatment following a test result has costs, which may be direct and/or indirect and medical and/or non-medical				
Medical test evaluation				
Analytical performance of a test is the ability to correctly detect or measure a measurand (trueness/validity, imprecision, limits of detection, cross-reactivity)				
Clinical performance of a test is the ability of a test to correctly classify those with and without the target condition (also called diagnostic accuracy)				
Diagnostic accuracy measures can be determined based on comparisons between the index test (test of interest) and reference test (also called reference standard, this may be the gold standard)				
Diagnostic accuracy measures can be derived from a 2x2 table in which patients with and without the disease of interest are classified according to their test result				
Diagnostic accuracy measures include true positives, true negatives, false positives, false negatives, inconclusive test results, sensitivity and specificity, positive and negative predictive value, likelihood ratio for positive and negative tests, diagnostic odds ratio, ROC-curve, and area under the curve				
Test accuracy describes the performance of a test in specific circumstances (and may therefore vary), e.g., in specific pre-test probabilities (prevalence of the disease)				
Positive and negative test results are defined based on a threshold for test positivity. Changing the threshold may change the test results				
Clinical effectiveness (also called clinical utility) of a test is the ability of a test to improve patient relevant outcomes				
Clinical effectiveness of a test is evaluated by evaluating the test-treatment pathway				
Clinical effectiveness of a test can be evaluated directly by performing a diagnostic RCT				
Clinical effectiveness of a test is mostly evaluated indirectly by assessing the different steps of the test-treatment pathway				

Possible required knowledge components ↓	Role in guideline panel			
	Health care provider	Health care consumer	Methodologist	Guideline panel chair
Cost-effectiveness analysis of a test is the evaluation of the balance between patient relevant outcomes and costs due to the introduction of a test				
The broader impact of a test refers to consequences of introducing or using a test beyond clinical effectiveness and cost-effectiveness (e.g., acceptability, implementability)				
Clinical practice guideline development				
Clinical practice guideline development about the value of a medical test starts with the definition of a clinical or health question. In the formulation of the question the proposed role (triage, add-on, replacement) of the medical test is described, the test-treatment pathway is outlined, and the patient relevant outcomes are determined				
Content expertise about the test and disease of interest is necessary				
Diagnostic accuracy can be considered a surrogate outcome for patient relevant outcomes				
All steps of the test-treatment pathway should be systematically and critically assessed, starting with the evaluation of diagnostic accuracy				
Specific methods (e.g., search filters, risk of bias tools, meta-analysis techniques, reporting standards) should be used for conducting systematic reviews of diagnostic test accuracy				
False negative and false positive results should be interpreted in terms of patient relevant outcomes				
The certainty of the evidence should be assessed. This includes test accuracy, effects of the test (direct benefit, adverse effects, burden), natural course of the disease of interest, effect of management guided by test results, the link between test results and management decisions, and the overall certainty of the evidence				
Desirable and undesirable consequences of a medical test should be balanced (by formal or informal modelling)				

Appendix 3. Interview guide

Item	Time
Introduction and questions beforehand START RECORDING	5 minutes
SHARE SCREEN WITH LIST In the list we classified several possible required knowledge components in three domains: diagnostic process in clinical practice, medical test evaluation and clinical practice guideline development. *The interviewees can have different points of attention that they spend more time on. Then keep questioning.* 1. Could you reflect upon the items mentioned under 'diagnostic process in clinical practice'? a. Are the items relevant? b. Is the formulation adequate? c. Do you miss important items?	10 minutes
2. Could you reflect upon the items mentioned under 'medical test evaluation'? a. Are the items relevant? b. Is the formulation adequate? c. Do you miss important items?	10 minutes
3. Could you reflect upon the items mentioned under 'clinical practice guideline development'? a. Are the items relevant? b. Is the formulation adequate? c. Do you miss important items?	10 minutes
4. In the list we distinguished four roles in the clinical guideline panel: health care provider, health care consumer representative, methodologist, and guideline panel chair. Do you think this is a logical classification?	5 minutes
We will probably not ask the question about the Bloom's taxonomy explicitly, but this will be discussed within the other questions - there is a separate slide for this - share screen if necessary. 5. We used a modified version of Bloom's taxonomy to distinguish between the different levels of required knowledge. To your opinion, which of the following levels are useful to incorporate in our survey? a. Not necessary to know b. Remember c. Understand d. Apply e. Analyse f. Evaluate g. Create	5 minutes
STOP SCREEN SHARING 6. Do you have any other additional comments or questions?	10 minutes
Closing remarks STOP RECORDING	5 minutes
TOTAL	60 minutes

Appendix 4. Description of the evidence retrieved through the literature review

The main characteristics of the evidence retrieved are described in *table 1*.

Table 1. Characteristics of the retrieved literature

First author (publication year)	Short description of the study (design/type)	Main results concerning knowledge to develop guideline recommendations about healthcare related testing
Albarquoni (2018) [1]	Systematic review and Delphi study to determine core competences in evidence-based practice for healthcare professionals	No information about knowledge required to develop recommendations about testing
Berger (2013) [2]	Description of a program for competence training in evidence-based medicine for healthcare consumers	No information about knowledge required to develop recommendations about testing
Hinneburg (2020) [3]	Description of a learning program for physicians/ medical students to enhance competencies in evidence-based decision-making	No information about knowledge required to develop recommendations about testing
Messerli (2017) [4]	Editorial about needed experience for guideline development	No information about knowledge required to develop recommendations about testing
Norris (2016) [5]	Development set minimum skills and experience for GRADE methodologists	▪ Perform one or more systematic reviews or develop guidelines that involved a question of diagnostic test accuracy or value ▪ The development of key (PICO) questions and the synthesis and assessment of evidence related to diagnostic tests have many unique considerations and the GRADE methodologist should have relevant experience
Schünemann (2016) [6]	Presentation of evidence-to-decision frameworks for tests in clinical practice and public health	▪ A test question should include test and subsequent management strategies as pathways to important outcomes ▪ Judgements about test accuracy should be based on a systematic review ▪ Judgements about benefits and harms require summaries of findings for desirable and undesirable effect on health outcomes (direct effects, downstream consequences, and linked evidence) ▪ Rating the certainty of the evidence requires consideration of each element of the linked evidence

First author (publication year)	Short description of the study (design/type)	Main results concerning knowledge to develop guideline recommendations about healthcare related testing
Schünemann (2008) [7]	Presentation of grading quality of the evidence and strength of recommendation for tests and test strategies	▪ Evaluation of a test in a guideline perspective should be seen in relation to people-important outcomes ▪ Usually indirect evidence is used
Brozek (2009) [8]	Presentation of grading quality of evidence about tests and test strategies	▪ Interpretation false negative and false positive test results is important to evaluate the value of a test ▪ Understanding of the proposed place of a new test in the pathway and suggested benefits is required
Sultan (2020) [9]	Development of competency framework for guideline developers	No information about knowledge required to develop recommendations about testing
Wieringa (2017) [10]	Description of using different kinds of knowledge for guideline development	No information about knowledge required to develop recommendations about testing
Zuiderent-Jerak [11]	Plead for guidelines to reflect all knowledge, not only RCT's	No information about knowledge required to develop recommendations about testing
Cochrane Handbook for systematic reviews of diagnostic test accuracy (draft version) (2022) [12]	Handbook	▪ Follow PRISMA-DTA reporting standards ▪ Include clinical and methodological expertise ▪ Be aware of direct harms of a test and impact thereof, and harms associated with false positive and false negative test results ▪ Assess failure and non-diagnostic findings, inconclusive results ▪ Evaluate acceptability of a test ▪ Utility depends on sensitivity and specificity and is influenced by the proportion with the target condition among those tested ▪ Report accuracy measures including confidence intervals ▪ Test accuracy depends on the threshold for test positivity ▪ Test accuracy describes test performance in specific circumstances ▪ Test can have different roles ▪ The clinical pathway contains setting and patient groups, index test and comparator tests, subsequent steps after testing driven by test result ▪ Specify the purpose of testing ▪ Report number of true positives, false negatives, true negatives, false negatives ▪ Use QUADAS-2 to evaluate risk of bias

First author (publication year)	Short description of the study (design/type)	Main results concerning knowledge to develop guideline recommendations about healthcare related testing
GRADE Handbook (2013) [13]	Handbook	▪ Be explicit about the purpose of the test ▪ Establish the role of the test ▪ Determine the standard diagnostic pathway ▪ Recommendations regarding use of medical tests require inference about consequences of false positive and false negative test results ▪ Diagnostic intervention studies can be evaluated using GRADE for interventions ▪ Otherwise, focus on test accuracy studies and make inferences about the likely impact on people-important outcomes ▪ Recommendations about tests require evaluation of the balance between desirable and undesirable consequences of that test, based on systematic reviews. ▪ Test accuracy studies are vulnerable to limitations, mostly due to indirect evidence

Albarqouni et al. determined core competencies in evidence-based practice for healthcare professionals in a systematic review and Delphi study [1]. Critical appraisal and interpretation of diagnostic accuracy studies as well as distinguishing evidence-based from opinion-based clinical practice guidelines were mentioned in the set of core competencies. However, the knowledge needed to develop guidelines about tests was not appointed. This also goes for the study of Berger et al. in which they described a program for competence training in evidence-based medicine for patients, patients counsellors, consumer representatives and health care professionals [2], as well as for the study of Hinneburg et al. who described a learning program for physicians and medical students to enhance competencies in evidence-based decision-making [3]. Messerli et al. advocate in an editorial that clinical expertise about the topic of interest is crucial to develop clinical practice guidelines for acceptation by health care professionals, by evaluating hypertension guidelines [4]. The authors do not specify knowledge needed to develop guidelines concerning tests.

Norris et al. described a set of minimum skills and experience required for GRADE methodologists working on the development of guidelines [5]. One of the recommended components of required experience is about tests: ‘Perform one or more systematic reviews or develop guidelines that involved a question of diagnostic test accuracy or value – The development of key (PICO) questions and the synthesis and assessment of evidence related to diagnostic tests have many unique considerations and the GRADE methodologist should have relevant experience’.

In the GRADE Guidelines Series, Schünemann et al. presented evidence-to-decision frameworks for tests in clinical practice and public health [6]. In this guidance, attention is given to aspects that typically belong to the evaluation of medical tests:

- 'Formulating a question about a test in the PICO format should include the test and subsequent management strategies as pathways to important outcomes as well as identifying subgroups that might require different recommendations or options.' This is also known as the test-treatment pathway.
- Determination whether the problem is a priority: this depends on the perspective (e.g., individual patient or population). For screening topics, this could also be a public health perspective.
- 'Judgments about test accuracy should be based on a summary of findings from a systematic review of test accuracy studies'.
- Benefits and harms: 'Judgments about the benefits and harms of using a test require preparation of a summary of findings for the modelled desirable and undesirable effects on health outcomes'. And: 'This includes information about direct benefits and harms of the test and the downstream consequences of interventions. In particular, judgments about the effects of the interventions that follow based on the test results (linked evidence) should be informed by a summary of findings table'.
- 'Rating the certainty of the evidence for the effects of tests requires consideration of each element of the linked evidence used to inform judgments about their benefits and harms'. This includes certainty of the evidence of:
  - Test accuracy.
  - Test related direct benefit, adverse effect, or burden of the test.
  - Natural course of the condition and the effect of management guided by test results.
  - Link between test result and management decisions.
  - Overall quality of the evidence.
- Valuing main outcomes: 'For tests, this includes adverse effects and any burden associated with the test, as well as downstream outcomes of linked interventions'.
- Balance between desirable and undesirable effects: 'For tests, this judgment is informed by the results of either formal or informal modelling of the anticipated desirable and undesirable effects of linked interventions'.
- Resource use: 'This includes judgments about how large the resource requirement was, the certainty of the evidence of resource requirement and the cost-effectiveness of interventions. This includes consideration of downstream costs....'.

- Equity, acceptability, and feasibility: 'For tests, assessment of equity, acceptability, and feasibility include consideration of both the test and linked interventions'.

In the forementioned article, another paper from Schünemann et al. is referenced. This is an article about grading the quality of the evidence and strength of recommendations for diagnostic tests and strategies [7]. In this article, it is emphasized that the evaluation of the quality of evidence of a test in a guideline perspective should be seen in relation to people-important outcomes. Since direct evidence (diagnostic RCT's) is scarce, usually indirect evidence is used to make inferences about impact on people-important outcomes. In the GRADE Guidelines Series paper, also a paper from Brozek et al. is referenced, which is about grading quality of evidence about diagnostic tests and test strategies [8]. In their paper, Brozek et al. mention the importance of the interpretation of false negative and false positive test results to be able to evaluate the value of test. They also state that 'this approach requires a clear understanding of the proposed place of a new test in a diagnostic pathway and its suggested benefits, as well as careful consideration of whether the patients detected by the new test are representative of the patients included in management trials'.

Sultan et al. developed a competency framework for guideline developers [9]. They describe different competencies, sub competencies and milestones, such as 'facilitate the development of guideline structure and setup'. No specific attention is given to competencies needed to develop testing recommendations.

Wieringa et al. described the use of different kinds of knowledge as a challenge for guideline development [10]. They criticize the focus on frequency-based reasoning and emphasize the importance of taking other knowledge into account. This study does not pay specific attention to the development of medical testing recommendations.

Zuiderent-Jerak et al. published their view on guideline development, in which they criticized the focus on RCT's in guideline development [11]. They state that guidelines should reflect all knowledge. No specific attention is given to the development of testing recommendations.

The Cochrane Collaboration has published a draft version of the second edition of the Cochrane Handbook for Systematic Reviews of DTA (*available for Cochrane members only*) [12]. Aspects of this handbook that might be relevant for the knowledge needed to develop testing recommendations in guidelines are stated below:

- 'The PRISMA-DTA reporting standards should be followed [14].
- Review teams should include clinical and methodological expertise in the topic area being reviewed, as well as the perspectives of stakeholders. For systematic reviews of test accuracy, it is often helpful to include both health professionals who use the index test in daily practice for the purpose specified in the review, and

experts who are familiar with the relevant technical details related to their implementation.

- For systematic reviews of test accuracy, author teams should include members with expertise in: literature searching, completing systematic reviews, test research methods, and statistics. The information specialists and statistical experts must be aware of the particular methodology for searching and data analysis for systematic reviews of test accuracy.
- Tests may directly harm patients if they are invasive. It is important to be aware of what these harms may be and how frequently they are encountered.
- It is also important to be wary of the harms associated with false positive and false negative diagnoses.
- It is important to assess failure rates and non-diagnostic findings.
- Evaluation of the acceptability of a test is important to assess whether patients are willing to undergo a procedure.
- Before starting a review, it is essential to understand what kind of primary study would ideally fit the review question.
- Before undertaking a test accuracy meta-analysis, it is necessary to understand the distinct types of data, as well as the presentation and meaning of statistical summaries of test accuracy reported in the primary studies.
- To inform decision-making, researchers should report the results or outcome for all participants undergoing testing. For many tests, this means that researchers should also report the number of persons tested for whom a conclusive result – a clear positive or a clear negative – could not be obtained, and the reasons why.
- The clinical utility of a test will always depend on both sensitivity and specificity and will also be influenced by the proportion with the target condition among those tested. It is therefore crucially important to always report sensitivity and specificity in pairs.
- Accuracy measures:
  - Sensitivity
  - Specificity
  - True positives, false positives, true negatives, false negatives
  - Positive predictive value
  - Negative predictive value
  - Positive likelihood ratio
  - Negative likelihood ratio
  - Youden's index
  - Overall accuracy

  - Diagnostic odds ratio
  - Receiver operation characteristic curve (and area under the curve)
- Results from individual studies should be reported with confidence intervals for each measure.
- In studies of the accuracy of tests with ordinal and continuous results, positive and negative test results are defined based on a threshold for test positivity and change if the threshold is altered. This dependence on threshold is a fundamental aspect of test accuracy evaluation.
- Test accuracy is not a fixed property of a test: accuracy describes the performance of a test in specific circumstances. The accuracy of a test may therefore vary with the intended use (e.g., screening versus diagnosis), population (e.g., children versus adults), setting (rural health centre in a low-income country versus urban hospital), prior tests (e.g., only signs and symptoms, or also an X-ray before CT-scanning), level of training (novice versus expert readers), and many more elements.
- In general, three roles can be defined for a new test relative to an existing test: (1) to select patients for whom follow-up testing may be useful (triaging); (2) to increase the accuracy of a testing strategy, by adding an extra test to the existing strategy (add-on); and (3) to replace one or more tests in the existing strategy with the (new) index test (replacement) [15]
- A description of the clinical pathway should contain the following elements: (1) the setting and patient groups to be tested, including relevant prior testing; (2) the index test and any comparator index tests; (3) subsequent steps after testing, driven by the test result, such as further testing or treatment.
- The purpose of testing should be specified explicitly, as well as the intended use population (asymptomatics versus symptomatics).
- Details of index tests should be collected.
- It is important to identify where a test is being used in a clinical pathway in each study.
- The definition of the target condition and the reference standard used to identify the presence or absence of the target condition must be collected.
- Collection of information about the harmful effects of testing may be desirable depending on the nature of the test.
- Forest plots for diagnostic test accuracy report the number of true positives and false negatives in participants with the target condition (diseased), and true negatives and false positives in participants who do not have the target condition (non-diseased) in each study, and the estimated sensitivity and specificity, together

with confidence intervals. The plots are known as coupled forest plots as they contain two graphical sections: one depicting sensitivity, and one specificity.
- A SROC plot is a scatterplot of the results of individual studies in ROC space where each study is plotted as a single (specificity, sensitivity) point.
- It is clear that the determinants of publication bias for reviews of randomized trials are unlikely to be generalizable to reviews of diagnostic accuracy studies [16, 17]'.

QUADAS-2 is a tool that is recommended by Cochrane for use in systematic reviews to evaluate the risk of bias and applicability and mentioned by the GRADE Working Group as a suitable risk of bias tool of primary diagnostic studies [18].
In the GRADE Handbook a chapter is published about the GRADE approach for diagnostic tests and strategies [13]. Some aspects of this chapter that are potentially relevant for the knowledge needed to develop testing recommendations are stated below:
- 'Guideline panels should be explicit about the purpose of the test in question.
- Guideline panels and authors of systematic reviews should also clearly establish the role of a diagnostic test or strategy. This process should begin with determining the standard diagnostic pathway – or pathways – for the target patient presentation and identify the associated limitations.
- It follows that recommendations regarding the use of medical tests require inferences about the consequences of falsely identifying patients as having or not having the disease.
- When diagnostic intervention studies (RCTs or observational studies) comparing alternative diagnostic strategies with assessment of direct patient-important outcomes are available, guideline panels can use the GRADE approach for other interventions.
- If studies measuring the impact of testing on patient-important or population-important outcomes are not available, guideline panels must focus on other studies, such as diagnostic test accuracy studies, and make inferences about the likely impact of using alternative tests on patient-important outcomes. In the latter situation, diagnostic accuracy can be considered a surrogate outcome for patient-important benefits and harms.
- A recommendation associated with a diagnostic question follows from an evaluation of the balance between the desirable and undesirable consequences of the diagnostic test or strategy. It should be based on a systematic review addressing the clinical question as well as information about management after applying the diagnostic test.

- In a typical test accuracy study, a consecutive series of patients suspected of a particular condition are subjected to the index test (the test being evaluated) and then all patients receive a reference or gold standard (the best available method to establish the presence of the target condition). While in the GRADE approach appropriate accuracy studies start as high quality evidence about diagnostic accuracy, these studies are vulnerable to limitations and often lead to low quality evidence to support guideline recommendations, mostly owing to indirectness of evidence associated with diagnostic accuracy being only a surrogate for patient outcomes.
- Several instruments for the evaluation of risk of bias in DTA studies are available. Cochrane Collaboration suggests a selection of the items from the QUADAS and QUADAS-2 instruments [18, 19]. Authors of systematic reviews and guideline panels can use the criteria from the QUADAS list to assess the risk of bias within and across studies.
- If only diagnostic accuracy information is available, the assessment of indirectness requires additional judgments about how the correct and incorrect classification of subjects as having or not having a target condition relates to people-important outcomes.

References

1. Albarqouni L, Hoffmann T, Straus S, Olsen NR, Young T, Ilic D, et al. Core Competencies in Evidence-Based Practice for Health Professionals: Consensus Statement Based on a Systematic Review and Delphi Survey. JAMA Netw Open. 2018;1(2):e180281. doi:10.1001/jamanetworkopen.2018.0281.
2. Berger B, Gerlach A, Groth S, Sladek U, Ebner K, Mühlhauser I, et al. Competence training in evidence-based medicine for patients, patient counsellors, consumer representatives and health care professionals in Austria: a feasibility study. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2013;107(1):44-52. doi:10.1016/j.zefq.2012.11.013.
3. Hinneburg J, Hecht L, Berger-Höger B, Buhse S, Lühnen J, Steckelberg A. Development and piloting of a blended learning training programme for physicians and medical students to enhance their competences in evidence-based decision-making. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2020;150-152:104-11. doi:10.1016/j.zefq.2020.02.004.
4. Messerli FH, Hofstetter L, Agabiti-Rosei E, Burnier M, Elliott WJ, Franklin SS, et al. Expertise: no longer a sine qua non for guideline authors? J Hypertens. 2017;35(8):1564-6. doi:10.1097/hjh.0000000000001435.
5. Norris SL, Meerpohl JJ, Akl EA, Schünemann HJ, Gartlehner G, Chen Y, et al. The skills and experience of GRADE methodologists can be assessed with a simple tool. J Clin Epidemiol. 2016;79:150-8.e1. doi:10.1016/j.jclinepi.2016.07.001.
6. Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
7. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.
8. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009;64(8):1109-16. doi:10.1111/j.1398-9995.2009.02083.x.
9. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schunemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
10. Wieringa S, Dreesens D, Forland F, Hulshof C, Lukersmith S, Macbeth F, et al. Different knowledge, different styles of reasoning: a challenge for guideline development. BMJ Evid Based Med. 2018;23(3):87-91. doi:10.1136/bmjebm-2017-110844.
11. Zuiderent-Jerak T, Forland F, Macbeth F. Guidelines should reflect all knowledge, not just clinical trials. Bmj. 2012;345:e6702. doi:10.1136/bmj.e6702.
12. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (eds.). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (update July 2023). Cochrane; 2023. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
13. GRADE Working Group. GRADE Handbook. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013. Available from: https://gdt.gradepro.org/app/handbook/handbook.html.
14. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, and the P-DTAG, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. Jama. 2018;319(4):388-96. doi:10.1001/jama.2017.19163.
15. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92. doi:10.1136/bmj.332.7549.1089.
16. Dickersin K. The existence of publication bias and risk factors for its occurrence. Jama. 1990;263(10):1385-9.
17. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. Jama. 1998;279(4):281-6. doi:10.1001/jama.279.4.281.
18. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36. doi:10.7326/0003-4819-155-8-201110180-00009.

19. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25. doi:10.1186/1471-2288-3-25.

Chapter 5.

Developing guideline recommendations about tests: educational examples of test-management pathways

Mariska Tuut
Jochen Cals
Jesse Jansen
Jako Burgers

BMJ Evid Based Med [submitted]

Keywords: Diagnosis, Evidence-Based Practice, Overdiagnosis, Policy, Quality of Health Care

Introduction

Recommendations about healthcare related testing in guidelines are common. Tests can be used for several purposes: screening, surveillance, risk classification, diagnosis, staging, treatment triage, determination of prognosis and monitoring/follow-up [1]. The development of testing recommendations in guidelines is challenging, especially because the benefit of a test not only depends on test characteristics, such as sensitivity and specificity, but also on population characteristics and test consequences, such as management [2-4]. Furthermore, the role of a new test in comparison to the existing testing scenario should be defined, since this influences the interpretation of the new test's value. The following roles of new tests have been identified in the literature: triage, replacement, add-on, and parallel/combined [5].

As with treatment, testing can have negative consequences, including physical impairment, psychological distress, disease labelling, and costs [6]. There is limited evidence on harms of testing, and healthcare professionals often overestimate its benefits while underestimating its harms [7]. This is also true for patients' expectations of testing [8]. Additionally, testing occasionally yields unexpected and coincidental findings, which may result in additional testing and treatment.

There is a lack of transparency in processing the evidence and considerations that support testing recommendations in guidelines[9]. To facilitate the development of test recommendations, we determined the minimum required knowledge for guideline panel members involved, supplementing the competency-based framework available for guideline development [10, 11]. The concept of the test-management pathway (*figure 1*) appeared key to understand.

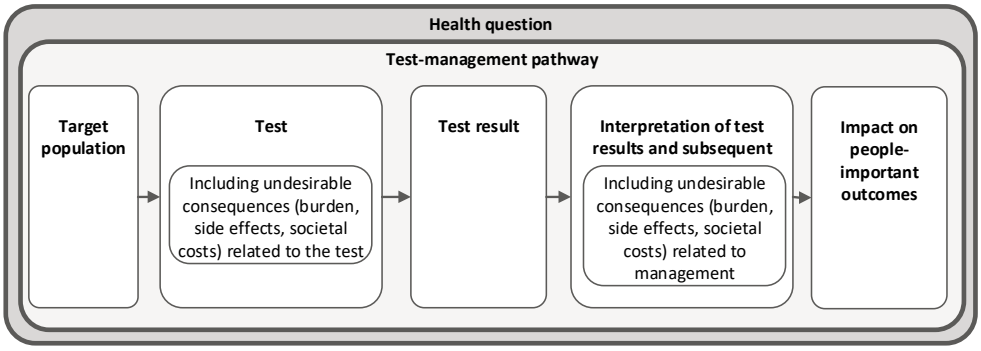


Figure 1. Test-management pathway concept

During our developmental study, the need for practical examples of test-management pathways became apparent [10]. In our subsequent teach-the-teacher workshop at the

2023 GIN conference [12], participants requested additional elaboration of pathways for different test outcomes (such as false positives and false negatives) being helpful for explaining the test-management pathway concept to guideline panel members.

The aim of this paper is to facilitate the understanding and uptake of the test-management pathway concept by offering four test scenarios in different settings and with different purposes and roles (*table 1*). The first scenario is a hypothetical example; the other scenarios are based on existing guidance.

Table 1. Test scenarios including examples

Scenario	Setting	Test	Condition of interest	Role of the test	References
Self-testing	Home	Smart watch with single-lead ECG-app	Atrial fibrillation	Triage	Hypothetical example
Screening	Secondary care	Annual MRI	Breast cancer	Replacement	NICE Guidance familial breast cancer [13]
Diagnostic testing	Primary care	CRP point of care	Severe lower respiratory tract infection	Add-on	Dutch GP guideline on acute cough [14]
Follow-up testing	Primary care	Annual spirometry	COPD	Add-on	Dutch GP guideline on COPD [15]

Self-testing

Self-testing means that individuals take their own samples, conduct a simple test, and interpret the results without assistance. Examples comprise assessment of blood glucose levels via finger prick tests, and detection of pregnancy through a urinary test.

Validated self-tests are easy to perform, are straightforward to interpret, and have a safety net (e.g. healthcare access) in case of unexpected test results. Limitations of self-tests include possible incorrect test execution and interpretation, and low quality tests with limited accuracy (such as commercial home-use HbA1c tests)[16]. Integrating and recommending self-tests in guidelines may be useful.

To illustrate this, we worked out a hypothetical test-management pathway example for the detection of atrial fibrillation (*figure 2a*). A guideline panel might recommend use of an ECG-app on a consumer watch as a triage test for people with symptoms or atrial fibrillation who are at risk of cardiovascular disease. This may lead to early detection of atrial fibrillation and subsequent treatment, possibly lowering stroke risk. A prerequisite would then be sufficient test accuracy, which means an acceptable rate of false positive, false negative, and inconclusive test results. In this scenario, feasibility concerns could lead to obtaining unreliable test, even if the test accuracy

(clinical performance) is suitable in principle. This may lead to unjustified healthcare consultation.

Screening

Screening tests are conducted in asymptomatic individuals to identify a subset of the population for further testing. Their objective is to detect conditions at an earlier stage to enable prompt management, including medical interventions and/or lifestyle adjustments to reduce risks of future events or to maximize treatment effectiveness. Examples include heel prick procedures to detect treatable congenital diseases in neonates, and faecal blood testing for early detection of colon cancer in people aged 55-75 years.

Screening tests could be beneficial if early detection of a condition leads to better people-important outcomes and these outcomes outweigh potential (physical and mental) harms of screening. Individuals should be informed about the benefits and harms of screening tests before decision making.

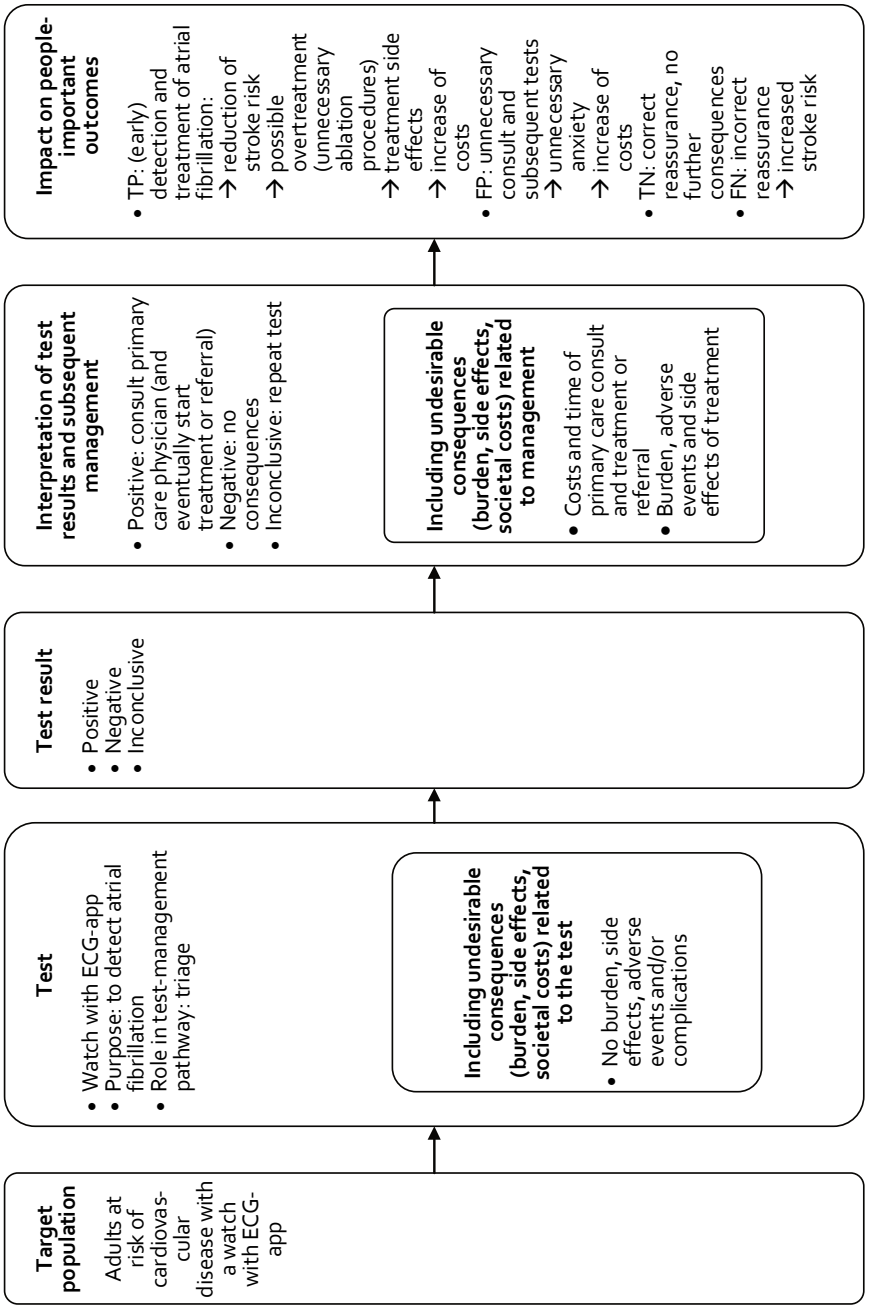
To illustrate, a guideline panel may suggest MRI screening as substitute for mammography in women at high risk due to its greater sensitivity [13]. When following the test-management pathway, it becomes evident that drawbacks of such screening should be considered, such as higher costs and potentially larger groups with false positive test results (*figure 2b*). Moreover, the impact of MRI screening on people-important outcomes, such as breast cancer-related mortality, needs to be evaluated with adequate follow-up time.

Diagnostic testing

A diagnostic test aims to confirm or exclude a particular disease. Examples include a urine dipstick to detect urinary tract infections, and X-rays to identify bone fractures.

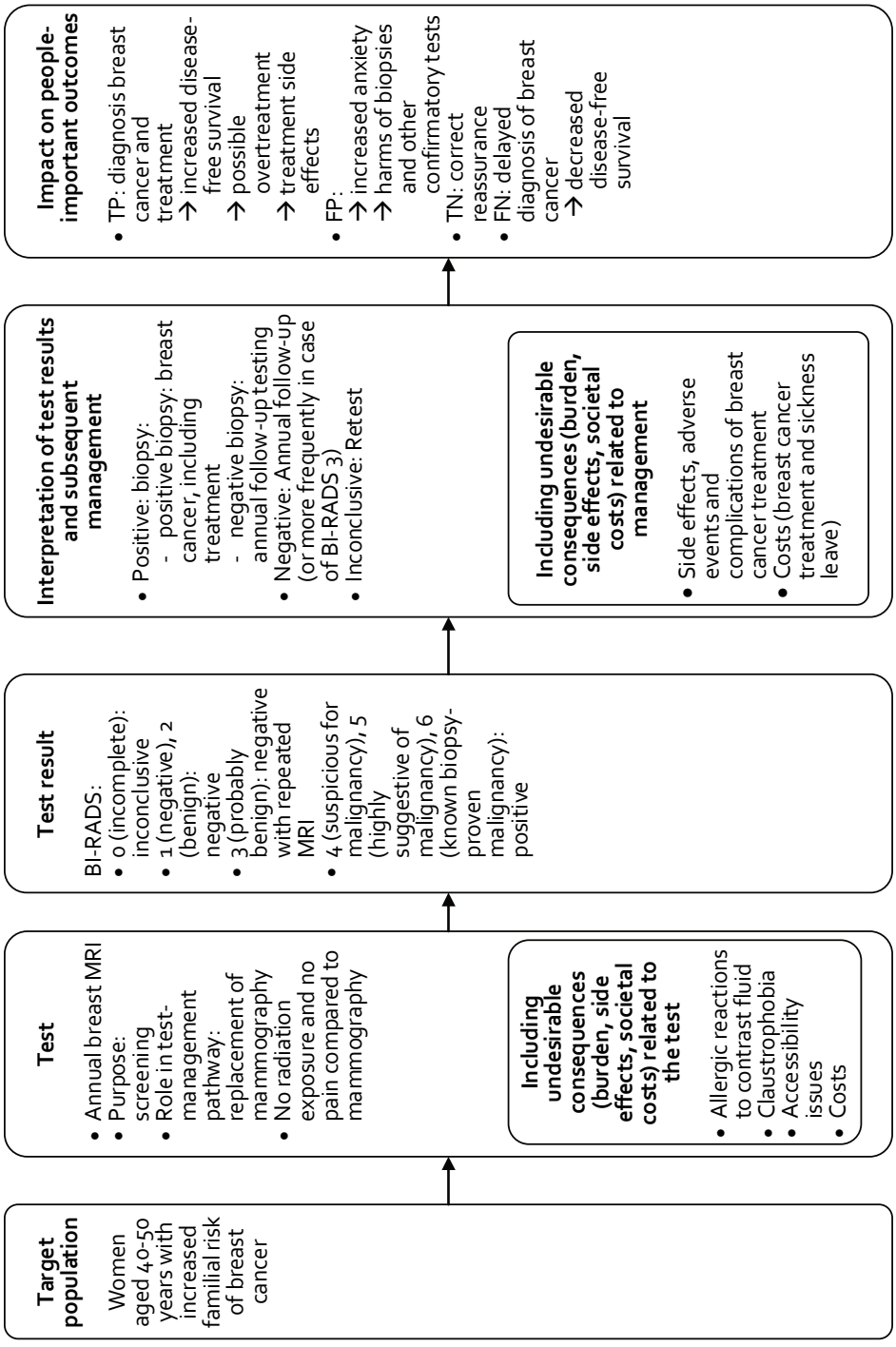
The final aim of diagnostic testing is to improve people-important outcomes. The required clinical performance of a test relies on its intended purpose. For diagnosing a condition, the test should have sufficient specificity (i.e. low false positives). Conversely, if the aim is to exclude a diagnosis, the test should have sufficient sensitivity (i.e. low false negatives).

We illustrated a test-management pathway for CRP testing in primary care patients with acute cough (*figure 2c*) [14]. In this scenario, a false-negative test result can lead to an increase of symptoms and complications, rather than a decrease. Additionally, clinicians might still feel uncertain in case of inconclusive test results. Finally, false-positive testing can lead to undesirable consequences of antibiotic management, such as side effects and antibiotic resistance.

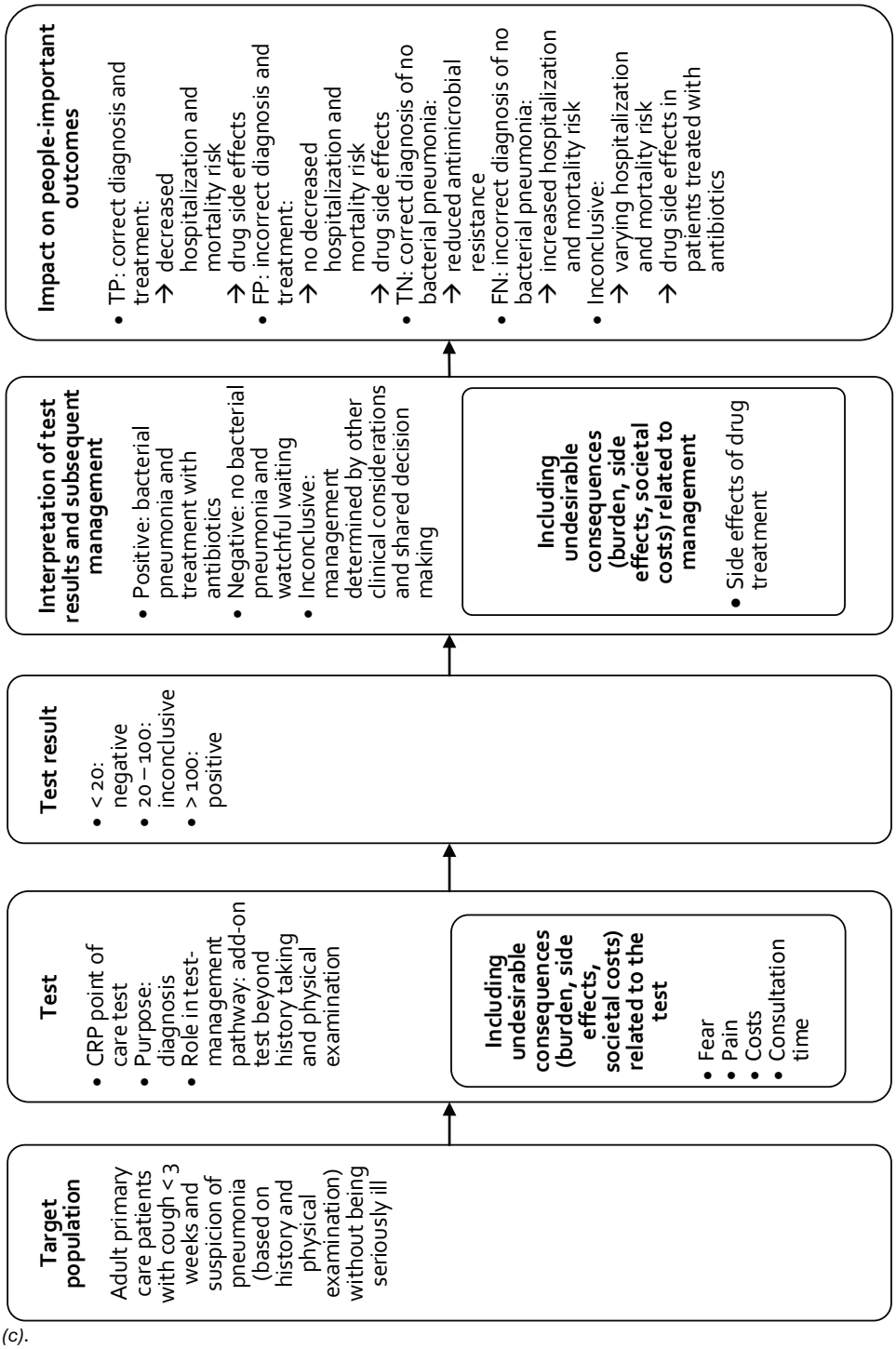


(a).

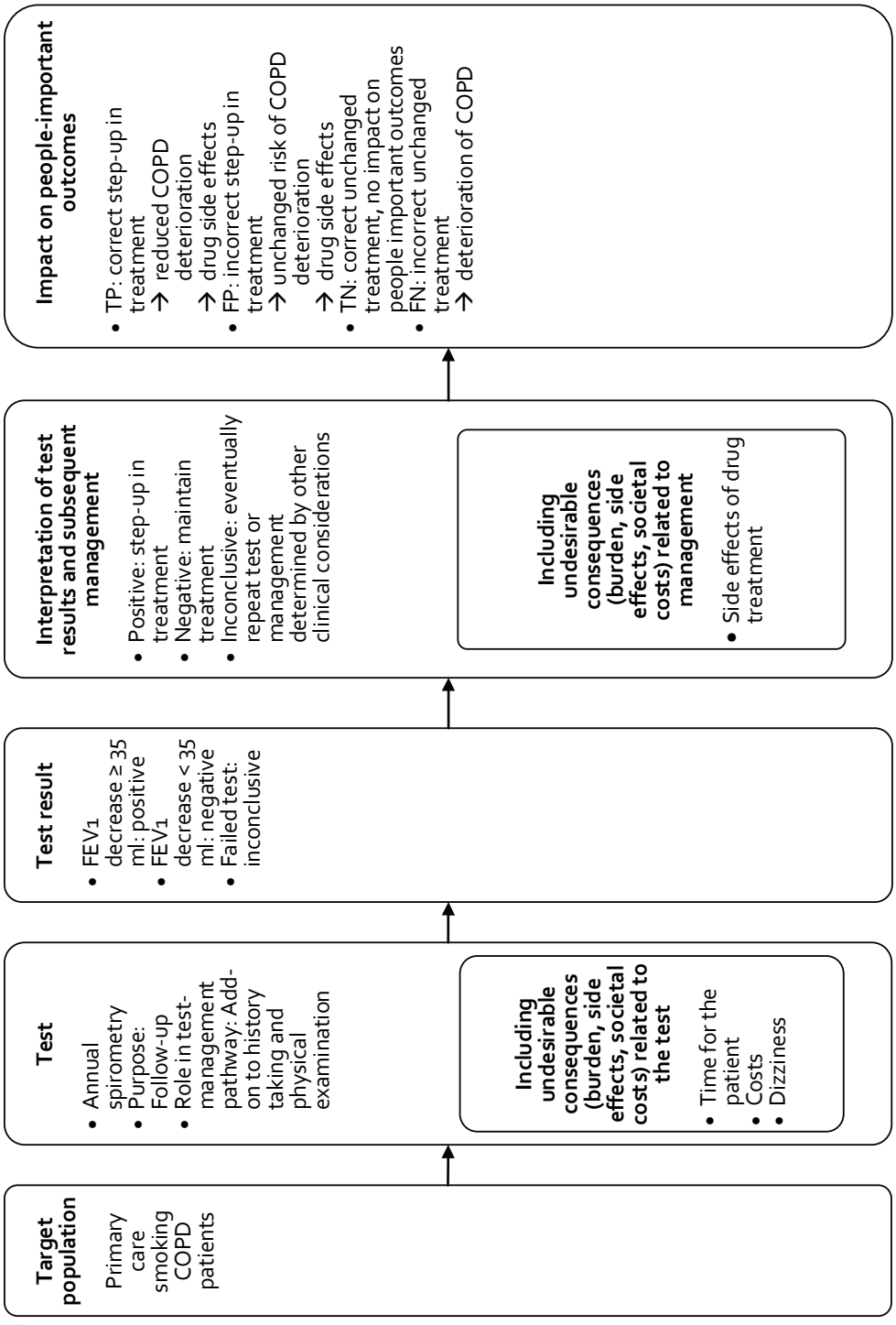
Figure 2. Test-management pathway examples on (a) self-testing, (b) screening, (c) diagnostic testing, and (d) follow-up testing. TP: true positives; FP: false positives; TN: true negatives; FN: false negatives. Note: The examples provided are intended solely as such and may not necessarily reflect the cut-off values or recommended management in currently applicable guidelines.


Target population
Women aged 40-50 years with increased familial risk of breast cancer
Test
• Annual breast MRI
• Purpose: screening
• Role in test-management pathway: replacement of mammography
• No radiation exposure and no pain compared to mammography
Including undesirable consequences (burden, side effects, societal costs) related to the test
• Allergic reactions to contrast fluid
• Claustrophobia
• Accessibility issues
• Costs
Test result
BI-RADS:
• 0 (incomplete): inconclusive
• 1 (negative), 2 (benign): negative
• 3 (probably benign): negative with repeated MRI
• 4 (suspicious for malignancy), 5 (highly suggestive of malignancy), 6 (known biopsy-proven malignancy): positive
Interpretation of test results and subsequent management
• Positive: biopsy:
- positive biopsy: breast cancer, including treatment
- negative biopsy: annual follow-up testing
• Negative: Annual follow-up (or more frequently in case of BI-RADS 3)
• Inconclusive: Retest
Including undesirable consequences (burden, side effects, societal costs) related to management
• Side effects, adverse events and complications of breast cancer treatment
• Costs (breast cancer treatment and sickness leave)
Impact on people-important outcomes
• TP: diagnosis breast cancer and treatment
→ increased disease-free survival
→ possible overtreatment
→ treatment side effects
• FP:
→ increased anxiety
→ harms of biopsies and other confirmatory tests
• TN: correct reassurance
• FN: delayed diagnosis of breast cancer
→ decreased disease-free survival

(b).


Target population
Adult primary care patients with cough < 3 weeks and suspicion of pneumonia (based on history and physical examination) without being seriously ill
Test
• CRP point of care test
• Purpose: diagnosis
• Role in test-management pathway: add-on test beyond history taking and physical examination
Including undesirable consequences (burden, side effects, societal costs) related to the test
• Fear
• Pain
• Costs
• Consultation time
Test result
• < 20: negative
• 20 – 100: inconclusive
• > 100: positive
Interpretation of test results and subsequent management
• Positive: bacterial pneumonia and treatment with antibiotics
• Negative: no bacterial pneumonia and watchful waiting
• Inconclusive: management determined by other clinical considerations and shared decision making
Including undesirable consequences (burden, side effects, societal costs) related to management
• Side effects of drug treatment
Impact on people-important outcomes
• TP: correct diagnosis and treatment:
→ decreased hospitalization and mortality risk
→ drug side effects
• FP: incorrect diagnosis and treatment:
→ no decreased hospitalization and mortality risk
→ drug side effects
• TN: correct diagnosis of no bacterial pneumonia:
→ reduced antimicrobial resistance
• FN: incorrect diagnosis of no bacterial pneumonia:
→ increased hospitalization and mortality risk
• Inconclusive:
→ varying hospitalization and mortality risk
→ drug side effects in patients treated with antibiotics

(c).


Target population
Primary care smoking COPD patients
Test
• Annual spirometry
• Purpose: Follow-up
• Role in test-management pathway: Add-on to history taking and physical examination
Including undesirable consequences (burden, side effects, societal costs) related to the test
• Time for the patient
• Costs
• Dizziness
Test result
• FEV1 decrease ≥ 35 ml: positive
• FEV1 decrease < 35 ml: negative
• Failed test: inconclusive
Interpretation of test results and subsequent management
• Positive: step-up in treatment
• Negative: maintain treatment
• Inconclusive: eventually repeat test or management determined by other clinical considerations
Including undesirable consequences (burden, side effects, societal costs) related to management
• Side effects of drug treatment
Impact on people-important outcomes
• TP: correct step-up in treatment → reduced COPD deterioration → drug side effects
• FP: incorrect step-up in treatment → unchanged risk of COPD deterioration → drug side effects
• TN: correct unchanged treatment, no impact on people important outcomes
• FN: incorrect unchanged treatment → deterioration of COPD

(d).

Follow-up testing

Follow-up testing may serve various objectives, such as maintaining disease control, detecting potential adverse drug effects, identifying early disease recurrence, or monitoring drug compliance. Examples include questionnaires to detect drug side effects, and mammography in women with a history of breast cancer.

We illustrated this with a test-management pathway for the use of spirometry in smokers with COPD (*figure 2d*) [15]. A positive test result prompts intensifying pharmaceutical treatment, whereas a negative test result may not lead to alterations, except for general smoking cessation advise in all cases. Repeated measurements and identifying possible reasons for suboptimal results (e.g., limited drug adherence) could be useful. Therefore, it is important to carefully consider all potential test consequences, to provide effective healthcare aimed at improving people-important outcomes. The example presented may thus be too narrowly focused, since issues such as feasibility of spirometry in a frail population, lifestyle measures, and drug adherence may also be important.

Conclusion

Designing test-management pathways can help formulate specific health questions about the use of testing as essential first step in guideline development [17]. These questions should then be answered by systematically reviewing and analysing the consequences of the different test results (TP, FP, TN, FN, inconclusive) on people-important outcomes, and considering other aspects, such as patient values, costs, and feasibility [18-20]. To illustrate, in the context of breast cancer, the question is not merely whether MRI-testing is more accurate than mammography in detecting breast cancer. Rather, the question is what is the net benefit of MRI-testing in comparison to traditional mammography in terms of people-important outcomes such as disease-free survival, taking into account all aspects, including patient burden, overdiagnosis and costs. This can be determined by formal modelling, in which evidence for the various steps of the test-management pathway is integrated into a decision analysis. An alternative is informal modelling, in which assumptions are made about the effects of different test results on people-important outcomes. In addition, further considerations are required to move from evidence to recommendations. These include certainty of the evidence, values, balance between the desirable and undesirable effects, resource use, equity, acceptability, and feasibility [19].

The elaborated test-management pathways in this paper serve as examples that can be used to explain the concept of test-management pathways.

To further facilitate the process of designing test-management pathways, an online tool for drafting such a pathway could be developed, which could be integrated in existing guideline development software [21]. Future research could evaluate the use of test-management pathways and its impact on guideline quality and guideline implementation.

References

1. Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, et al. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PLoS One. 2015;10(10):e0134553. doi:10.1371/journal.pone.0134553.
2. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014;67(7):760-8. doi:10.1016/j.jclinepi.2014.01.006.
3. Gopalakrishna G, Leeflang MM, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6(9):e010549. doi:10.1136/bmjopen-2015-010549.
4. Tuut MK, de Beer JJA, Burgers JS, van de Griendt EJ, van der Weijden T, Langendam MW. Applying GRADE for diagnosis revealed methodological challenges: an illustrative example for guideline developers. J Clin Epidemiol. 2020;131:123-32. doi:10.1016/j.jclinepi.2020.11.021.
5. Mustafa RA, Wiercioch W, Cheung A, Prediger B, Brozek J, Bossuyt P, et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 2: a review of methodological and practical challenges. J Clin Epidemiol. 2017;92:18-28. doi:10.1016/j.jclinepi.2017.09.003.
6. Korenstein D, Harris R, Elshaug AG, Ross JS, Morgan DJ, Cooper RJ, et al. To Expand the Evidence Base About Harms from Tests and Treatments. J Gen Intern Med. 2021;36(7):2105-10. doi:10.1007/s11606-021-06597-9.
7. Hoffmann TC, Del Mar C. Clinicians' Expectations of the Benefits and Harms of Treatments, Screening, and Tests: A Systematic Review. JAMA internal medicine. 2017;177(3):407-19. doi:10.1001/jamainternmed.2016.8254.
8. Hoffmann TC, Del Mar C. Patients' expectations of the benefits and harms of treatments, screening, and tests: a systematic review. JAMA internal medicine. 2015;175(2):274-86. doi:10.1001/jamainternmed.2014.6016.
9. Tuut MK, Burgers JS, van der Weijden T, Langendam MW. Do clinical practice guidelines consider evidence about diagnostic test consequences on patient-relevant outcomes? A critical document analysis. Journal of evaluation in clinical practice. 2021 doi:10.1111/jep.13619.
10. Tuut MK, Burgers JS, De Beer JJA, Bindels PJE, Bossuyt PMM, Cals JW, et al. Required knowledge for guideline panel members to develop healthcare related testing recommendations - a developmental study. J Clin Epidemiol. 2024;173 doi:10.1016/j.jclinepi.2024.111438.
11. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schünemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
12. Tuut MK, Langendam MW, Burgers JS, Van der Weijden T. Workshop: Training guideline panel members involved in developing recommendations about healthcare tests. Guideline International Network; 2023; Glasgow.
13. National Institute for Health and Care Excellence. Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer. 2023. Available from: https://www.nice.org.uk/guidance/cg164.
14. Dutch College of General Practitioners. NHG Standaard Acuut hoesten [Acute cough]. 2024. Available from: https://richtlijnen.nhg.org/standaarden/acuut-hoesten.
15. Dutch College of General Practitioners. NHG-Standaard COPD. 2021. Available from: https://richtlijnen.nhg.org/standaarden/copd.
16. Jacobsen LM, Bocchino LE, Lum JW, Kollman C, Barnes-Lomen V, Sulik M, et al. Accuracy of Three Commercial Home-Use Hemoglobin A1c Tests. Diabetes Technol Ther. 2022;24(11):789-96. doi:10.1089/dia.2022.0187.
17. Tuut MK, Gopalakrishna G, Leeflang MM, Bossuyt PM, Weijden Tvd, Burgers JS, et al. Specifying the test-management pathway to formulate focused guideline questions about healthcare related tests: a step-by-step guide. BMC Med Res Methodol. 2024 [submitted]
18. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.

19. Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
20. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta. 2014;427:49-57. doi:10.1016/j.cca.2013.09.018.
21. Evidence Prime. GRADEpro. Available from: https://www.gradepro.org/product.

Chapter 6.

Co-creation of a step-by-step guide for specifying the test-management pathway to formulate focused guideline questions about healthcare related tests

Mariska Tuut
Gowri Gopalakrishna
Mariska Leeflang
Patrick Bossuyt
Trudy van der Weijden
Jako Burgers
Miranda Langendam

BMC Medical Research Methodology [submitted]

Abstract

Background: Guideline development on testing is known to be difficult for guideline developers. It requires consideration of various aspects, such as accuracy, purpose of testing, and consequences on management and people-important outcomes. This can be outlined in a test-management pathway. We aimed to create and user-test a step-by-step guide for guideline developers for designing a test-management pathway.

Methods: Developmental design with a co-creative strategy. We created a draft step-by-step guide, that was user tested in a workshop with 19 experts, and by interviewing 7 guideline panel members.

Results: Our proposed guide consists of five blocks of signalling questions: patients/population, index test(s), current practice/comparison/control, people-important outcomes, and the link between testing and outcome(s). The user testing led to refinement of the signalling questions, the use of inclusive terminology, and addition of a test-management pathway figure with detailed explanation.

Conclusions: The step-by-step guide for formulating focused guideline questions regarding healthcare related testing can help in identifying relevant characteristics of the population, tests, and outcomes and to create a test management pathway. This should facilitate the formulation of evidence-based guideline recommendations about healthcare related testing.

Funding: The creation and expert user testing of the step-by-step guide were part of the DECIDE project, a 5-year project which ran from January 2011 to December 2015, co-funded by the European Commission under the Seventh Framework Programme. The funder was not involved in the design, conduct, interpretation or reporting of the study. The research work done to adjust and test the step-by-step guide was done without specific financial funding.

Keywords: healthcare related testing, guidelines, methodology, education in guideline methods, guidance

Background

Guidelines aim to support decision making in healthcare practice[1]. The ultimate goal of a guideline is to improve or sustain health outcomes that are considered important from the perspective of the target population of the guidelines, so-called people-important outcomes, such as mortality and quality of life. A set of key questions define the scope of the guideline. Answers to such questions are based on systematic reviews of the evidence, combined with clinical expertise, and patients' or consumers' values and preferences. These are subsequently translated to guideline recommendations by the guideline panel [2]. The key questions include specific components, such as the population of interest, the intervention of interest and people-important outcomes [3].

In healthcare, tests can provide additional information about the past, current or future state of a person. The information may be relevant for diagnostic, prognostic, screening, monitoring, treatment (options), or other purposes [4]. Testing in itself usually has no direct effect on a patient/person's health status. In addition, healthcare related tests are rarely used in isolation. They are typically included in a test-management pathway in which the information from testing is used to guide further actions [5-7].

The incremental value of test information will depend on population characteristics (e.g., features, symptoms, context and setting), test characteristics (e.g. sensitivity and specificity), management options following the test result and their downstream consequences on people-important outcomes [8]. The chain of these elements, starting from the testing process and ending in people-important outcomes is called the test-management pathway.

Different terms can be used for pathways that link testing to further clinical actions and people-important outcomes, such as test-treatment pathway, diagnostic pathway, clinical pathway, and analytical framework. In our study, we use the term 'test-management pathway' to be as inclusive as possible. Additionally, we prefer to use the term 'guidelines', rather than 'clinical practice guidelines' so as to also include the public health domain. We also use the term 'test', 'instead of 'diagnostic test', to include other purposes and settings of testing and test strategies [9].

Currently, the dominant source of evidence about testing most often comes from studies evaluating test performance, such as diagnostic accuracy [10]. Consequently, most guideline recommendations on testing are based on evidence concerning test accuracy only [11]. While having the best available estimates of a test's clinical sensitivity and specificity is desirable, it is not sufficient for deciding whether testing should be recommended for use. Accuracy measures can help in estimating how many

false positive and false negative results one may expect with testing but this information should be put into context. For instance, the clinical performance of a test may differ in public health compared to a clinical setting due to factors such as the pre-test probability of the population being tested, previous tests conducted, and the resulting management decisions.

To develop recommendations about testing, guideline developers need to consider (a) the purpose of testing, (b) the desired downstream consequences of the test, in terms of minimal important changes in people-important outcomes, and (c) the link between test results, (healthcare) actions, and these outcomes [9, 12-14]. In addition, feasibility of the test (including sustainability), test burden (e.g. pain, time, discomfort), resources and costs need to be considered.

The aim of testing is to improve people-important outcomes. A test-management pathway provides a visual representation of the essential steps required to move from testing to people-important outcomes, which is crucial in guideline development [15]. If guideline developers do not oversee and consider the consequences of testing, they cannot balance the relevant benefits and harms of testing. Relying on test accuracy solely may overestimate the added value of a test and may lead to overtesting, overdiagnosis and overtreatment.

Several agencies refer to the identification of test-management pathways in the evaluation of healthcare related tests and in drafting testing recommendations [13, 16-18]. These organisations mention the development of such pathways as part of the scoping process of a guideline, or as part of developing focused questions for systematic literature review. Studies in the guideline development community also support the integration of pathways in diagnostic test evaluation [19].

Identifying and outlining the elements of a test-management pathway in time and formulating focused questions about healthcare relating testing is not an easy task [20, 21]. Guideline developers have acknowledged that the inclusion of people-important outcomes in guideline development regarding testing is necessary but currently lacking. The formulation of key questions has been identified as a challenging aspect of this process, and there is consensus that education can play a crucial role in addressing this challenge [19]. Guideline developers therefore need support to formulate focused questions about testing at the start of a guideline development process.

Currently, a practical guide for the development of a test-management pathway is not available. Our group aimed to create, and user test a step-by-step guide on how to

design such a test-management pathway aimed at guideline developers. The intention was that such a guide would assist guideline developers in formulating focused questions and evidence-based recommendations on testing.

Methods

General methodology

This project was based on a developmental design with a co-creative strategy. The initial creation of the step-by-step guide and the first phase of user testing were part of the DECIDE project, a 5-year project from January 2011 to December 2015, co-funded by the European Commission under the Seventh Framework Programme. Its objective was to build on the work of the GRADE Working Group to develop and evaluate methods for the dissemination of guidelines, including the evaluation of evidence and the development of recommendations about healthcare related tests [22]. Finalisation of the step-by-step guide and additional user testing was conducted in 2023. The authors who participated in the pilot testing and user testing sessions are all researchers in the field of test evaluation and/or guideline development. They do not currently hold any active healthcare provider roles.

Firstly, the project team drafted a number of signalling questions per PICO element. Secondly, the step-by-step guide was co-created with two experts in the field and underwent user-testing with experts in the field and guideline panel members. This approach was selected to ensure comprehensive consideration of all relevant aspects. The Standards for Reporting Qualitative Research (SRQR) have been used to guide reporting of the research [23].

Development of the step-by-step guide for creating a test-management pathway

The initial project team (GG, MML, PMB, MWL) selected the Population – Index test – Comparator – Outcome (PICO) elements as a starting point [24]. Using these elements and handbooks as basis (Agency for Healthcare Research and Quality, US Preventive Services Task Force, Cochrane handbook (for diagnostic test accuracy), GRADE for Diagnosis), the project team proposed a number of signalling questions for each PICO element, also based on their own expertise and experience in guideline development and study design [13, 16-18]. The aim of these questions was to facilitate guideline panel members in identifying issues that may need consideration when positioning the test of interest in its proposed pathway. The draft step-by-step guide was co-created in 2014 with one diagnostic test accuracy systematic reviewer and one guideline methodologist (MKT) within the project team. With these experts, the test-management

pathways for their topic of interest was drawn and their feedback was incorporated into the draft step-by-step guide.

User testing workshop with experts

Workshop participants were healthcare professionals and researchers with expertise and/or interest in guideline development who participated in the DECIDE Conference in Edinburgh in June 2014. We provided the participants with a 15-minute introduction on the relevance of creating a test-management pathway in developing testing recommendations and presented our proposed approach.

Then, test-management pathways were drafted using the step-by-step guide for two example questions: (1) B-type Natriuretic Peptide (BNP) testing for heart failure in elderly patients, and (2) CT-scanning in children with head injury who present at the emergency department. These topics were proposed by two volunteer participants. The test-management pathways were drafted through a collaborative effort between one researcher (PMB) and these volunteers in the presence of the other participants. Another researcher (MML) documented the process on a whiteboard. Two other project team members (GG and MWL) observed the process and took minutes.

Participants of the workshop gave input on these pathways, could ask questions and provided feedback. At the end of the workshop, participants completed a questionnaire about the usefulness and perceived challenges of the process used in the step-by-step guide (*Appendix 1*). The responses to these questionnaires were used to inform potential improvements to the step-by-step guide, including the wording of the steps.

User testing with guideline panel members

In this phase, conducted in 2023, we used a before-after approach, in which we asked guideline panel members to formulate a guideline question on testing without and then with the use of the step-by-step guide. We selected a purposeful sample of at least five guideline panel members from an unspecified number of guideline panels, relying on our own network in the Netherlands. To be eligible, guideline panel members had to be involved at the start or in the development process of a guideline on testing. Guideline panel members were invited to participate per email. We provided the participants with a brief description of the project and planned two interviews with each participant to collect data.

The interviews were conducted by te first author of this study using the interview guides in *Appendix 2*. In the first interview conducted online, participants were asked to formulate a key question concerning the added value of a test for their guideline topic

of interest. Then we sent our step-by-step guide, asked the participant to read this guide carefully and to note any questions, if the guide was not sufficiently clear. For this part of the study, we updated the step-by-step guide using inclusive terminology and translated it into Dutch (see *Appendix 3*).

In the second interview, conducted face-to-face, participants were asked to draw the test-management pathway for their test of interest using the step-by-step guide and answered any questions they had in the process. Then, participants were asked to adjust the originally formulated key question, if needed, and to provide feedback on the step-by-step guide and its use for this purpose.

All interviews were video recorded for note-taking and for incorporating feedback in the final version of the step-by-step guide.

Results

Development of the step-by-step guide for creating a test-management pathway

We created a guide consisting of five blocks of signalling questions concerning: (1) (P) patients/population, (2) (I) index test(s), (3) (C) current practice/comparison/control, (4) (O) people-important outcomes, and (5) link between testing and outcome(s). Pilot-testing of the draft step-by-step guide on diagnosis of eosinophilia in asthma and breast cancer screening resulted in refinement of the guide and the conclusion that the order in which the questions are addressed could vary, depending on the clinical question or topic. As an illustrative case, the pilot on breast cancer screening is reported in *Appendix 4*. The draft step-by-step guide is shown in *Appendix 5*.

User testing with experts

Nineteen participants provided feedback on the step-by-step guide by completing the questionnaire (see *Appendix 6* for detailed results). All agreed that drafting a test-management pathway is useful or even essential. Key issues raised were that more than one test-management pathway is likely for each guideline or key question and that all relevant stakeholders, such as healthcare professionals and consumers, should be involved in drafting the test-management pathway.

About half of the participants did not immediately see a direct link between the test-management pathway and derivation of relevant key questions. The participants who saw a link, valued the inclusion of people-important outcomes in the pathway and mentioned that making these outcomes explicit facilitates inferring changes in people-

important outcomes when considering alternative testing in the test-management pathway.

Participants had different opinions about the ordering of the questions, the use of PICO, and the way the guidance was set up. People wondered why we chose a particular order in some cases (such as IPCO) and preferred sticking to the original PICO-order. One participant mentioned that setting should be explicitly included as an element in addition to the PICO. Some participants would have liked to see harms and patients' values and preferences added to the outcome section as well. Following the user testing conducted in this phase of the study, no significant amendments were made to the step-by-step guide. However, a number of refinements have been incorporated.

All participants, except one, would consider using the test-management pathway in their guideline work if step-by-step user guidance would be available. About half of the participants preferred an open question format for the guide, while others favoured a checklist format. One participant suggested producing software that could help in the visualization of the pathway.

Besides knowledge about tests, diagnostic research, and evidence-based medicine, participants indicated that they would value training in interviewing skills and in moderating discussions involving the guideline panel. This training could have different formats, such as video tutorials, hands-on practicing, online training, and/or a more detailed step-by-step checklist.

User testing with guideline panel members

During the final round of user testing, seven guideline panel members from two Dutch panels on the topics secondary care for people with autoimmune haemolytic anaemia and primary care for women with dysmenorrhea were included. The participants included two clinical chemists, one haematologist, one general practitioner, and three patient representatives. In the first online interview, all interviewees were able to formulate an initial testing question. Prior to the second interview, six participants had reviewed the step-by-step user guide that was provided after the initial interview. During the second interview, all participants were able to create a test-management pathway for their question of interest, by using the step-by-step guide and instructions provided by the interviewer.

After drafting the test-management pathway for their test of interest, six participants adjusted their original question. These adjustments included:

- Refining the population of interest (such as adding information about the setting and earlier tests performed)
- Specifying the purpose of the test and its place in the test-management pathway
- Addressing practical aspects of testing, such as difficulties in performing the test adequately
- Defining test burden
- Adding the impact of testing in terms of impact on people-important outcomes

Participants found the step-by-step guide helpful for structuring questions and defining the purpose and impact of the test of interest. They also found the examples provided useful and intended to use the guide in a guideline panel setting. Suggestions for improvement included the need for instruction for usage, a figure/example of a test-management pathway, and the explanation of terminology for patient representatives.

Final step-by-step guide

In the final version of the step-by-step guide, we added an introduction, instructions, and a figure with the test-management pathway. The final version of the step-by-step guide is presented in *box 1*.

Discussion

This study presents a step-by-step guide for guideline panels to formulate focused questions regarding healthcare related testing. The guide can aid in creating a test-management pathway by identifying relevant characteristics of the population, tests, and outcomes of interest when developing clinical practice guidelines or public health guidelines.

The formulation of focused rather than broad questions allows explicit consideration of factors beyond test accuracy. These include feasibility, timing, test burden, management effectiveness and impact on people-important outcomes. Furthermore, the step-by-step guide offers the possibility of distinguishing between different patient subgroups. It is assumed that this approach will result in recommendations that are better balanced and that are explicitly aimed at improving people-important outcomes. This may lead to less overtesting, overdiagnosis and subsequent overtreatment, which would be beneficial from a patient's perspective as well as from a public health perspective.

Box 1. Final step-by-step guide for developing a test-management pathway

This step-by-step guide aims to assist guideline panels in formulating focused questions about healthcare related testing through drafting a test-management pathway. *Figure 1* shows an example format of such a pathway.

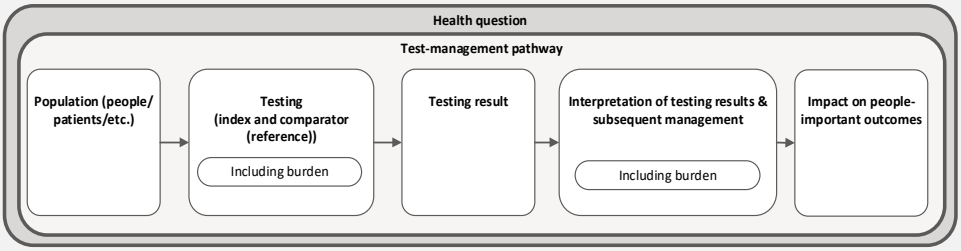


Figure 1. Illustration of a test-management pathway

Guideline panels can use this guide to define focused questions by selecting signalling questions for their topic of interest. The user can change the order of the steps and questions.

Steps	Signalling questions
People (Setting & Timing)	**For whom is testing considered?**
▪ For which persons is testing considered? ▪ Define healthcare setting	*Consider personal characteristics, setting, referral patterns, previous test results.* ▪ Are we interested in a particular age, sex or gender? ▪ Have the persons been referred from another setting? Were other tests performed? ▪ In what setting will the persons be tested? (population screening program, general practitioners practice, physiotherapy practice, hospital, etc.) ▪ Should subgroups be considered?
Index test	**Which test or testing strategy is considered?**
▪ Define measurand ▪ Primary purpose of the index test ▪ Define measurement platform or assay(s)	*The guideline panel will have to be specific enough in the description of the test that is considered.* ▪ What is the measurand (the physical quantity or property that is being measured)? ▪ What is the primary purpose of testing (screening, diagnostic, prognostic, predictive, monitoring, etc.) ▪ What is the role of the test relative to other tests (triage, replacement, add-on, parallel/combined) ▪ Is a combination of tests or specific testing strategy considered? (multimarker score, sequence of tests, etc.) ▪ What is the burden associated with the test (efforts to undergo the test, adverse effects, complications, costs, etc.) ▪ Are there any feasibility considerations? (resource requirements, training, storage, transport, etc.) ▪ Are there any acceptability considerations? (patients values and preferences, equity, costs, etc.) ▪ What platform or which kind of assay is used for the measurand?

Outcome(s) of interest	**What is the ultimate goal to achieve, avoid or simplify in people in whom testing is considered?**
▪ Define the anticipated or desired impact of testing on downstream (people-important) outcomes ▪ Define the how the index test results can guide (clinical) management decisions	*Guideline panels will likely need an introduction on how to define these outcomes.* ▪ What are the (crucial and important) people-important outcomes that ultimately matter? ▪ How may the index test help to improve, avoid, simplify or these outcome(s)?
Linking outcomes to testing	**How will testing guide further healthcare actions or patient management?**
▪ Link (positive, negative, failed, inconclusive, continuous) test results to management options and people-important outcomes	*Testing in itself rarely leads to the desired outcomes.* ▪ What management options are available after testing, to achieve, avoid, or simplify the people-important outcomes mentioned under c? ▪ What management options may follow the following test results: - For dichotomized test results: positive test result - For dichotomized test results: negative test result - For continuous test results: actual test results - Failed tests - Inconclusive test results ▪ What is the target condition or target event? (this may be a disease or disease stage) ▪ What are the consequences of false positive and false negative tests results on people-important outcomes?
Comparator	**What is the alternative to testing?**
▪ Define the existing pathway or the one that would be in place if the index test under (b) was not available	*This refers to the 'C' in the PICO framework, the comparator. The comparator may be the standard of care.* ▪ What is currently being done to achieve, avoid or simplify the people-important outcome(s) mentioned under c? ▪ What type of information guides or would guide management if we did not or do not have the index test results?

Explicit step-by-step guidance on how to actually derive such pathways is limited in the existing guidance. So far, we have not been able to identify studies reporting on the experience of users applying these approaches. Test-management pathways and concepts have been presented earlier as a tool for setting the scene and framing the question(s) in a guideline development or test accuracy review process [13, 16-18]. The AHRQ and the USPSTF refer to the development of such pathways as a guide to help in formulating specific key questions [16, 17]. Both organizations use the term 'Analytical Framework', which they use both for intervention related questions and for test-related questions as a way of going from a more ambiguous initial claim to a specific answerable guideline or review question. The Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy uses the term 'Clinical pathway' to outline how

patients might present, when they would be considered for testing and the role of the test [18].

Limitations of this study

The first user testing was done with participants experienced in test research and/or guideline development. Most participants in these sessions had prior knowledge about evaluation of tests and development of guideline recommendations about tests. It is therefore questionable whether the results of these user testing phases are applicable to guideline panels with less methodological expertise and experience. It is known that guideline panels are quite familiar with treatment guidelines and have limited initial understanding of the link between testing, downstream management, and people-important outcomes [19]. Thus, more guidance might be needed. We therefore also user tested the step-by-step guide with guideline panel members.

We tested the step-by-step guide in a limited number of persons and panels. They were recruited as a purposeful sample of experienced and less experienced guideline developers and reviewers with varying expertise and experience in test research. During user testing with guideline panel members, we observed that participants considered reformulating their initial test question after using the step-by-step guide. The instructions on creating questions for a guideline reflect a natural situation, as the development of guideline questions typically involves a group process led by a guideline methodologist.

Some data were collected almost a decade ago. Therefore, we adjusted the step-by step guide using terminology that is inclusive and more widely accepted by today's standards. We believe that our findings are still relevant after adjustment and by adding a test-management pathway visualization and instructions for use.

The primary focus for our step-by-step guide is to raise awareness on people-important outcomes. Recommendations on tests can be focused on other aims as well, such as simplifying or streamlining the healthcare process, to reduce costs, to increase efficiency, or to reduce painful procedures. We agree that these considerations matter but in all cases the people-important outcomes should also be considered.

Implications for practice

The step-by-step guide is meant to be used in a flexible manner. During the user testing sessions, there was some debate about where to start in the process: with the 'P' for people or population, or with the 'I' for index test. We think this may depend on the overall question to be answered. For example, if an index test is central in the question, such as 'Should we use this test in these patients?', then starting with the 'I' seems to

result in a more focused process. On the other hand, if the question is about whether a test should be recommended in a particular setting, then first describing the 'P' and setting may be more helpful.

We suggest to utilize the step-by-step guide in the guideline panel process during the stage of (PICO) question generation [25, 26]. Drafting a test-management pathway will often be an iterative process. Further modifications of the pathway during guideline development may be needed. Our step-by-step guide can help in this process. Although using the step-by-step guide in the guideline development process may require some time, it is anticipated that this will facilitate the specification of more focused questions. We expect that this might reduce the time required at a later stage in the guideline development process and will enable the development of targeted and more balanced recommendations.

Though our focus was on guideline development, we have experienced that the user guide can also be useful in other areas of decision making. As authors, we have used it in developing recommendations about coverage in a healthcare benefits package. We have also used it when designing clinical trials and deciding on the proper performance measures. Within the recently introduced European Union In Vitro Diagnostics Regulation, clinical performance should be informative about the clinical utility of the test, reflecting the purpose of testing in the intended use setting and population.

Users expressed that a digital tool that is both intuitive and flexible would be helpful for drawing test-management pathways, and to document the iterations it goes through. We suggest developing an online tool, for example as a feature in software such as RevMan and/or GRADEpro.

Implications for research

In developing a test-management pathway we encourage further evaluations of the step-by-step guide in guideline panels. This could result in additional tools and instruments to facilitate the development of recommendations about tests and testing.

Once the pathway is defined, research evidence to support assumptions made in the pathway can be sought. One could also use the test-management pathway to decide on minimally acceptable performance of the tests, and to evaluate limitations in the applicability of research findings.

Conclusion

We have developed a step-by-step guide, for guideline developers, to create a test-management pathway, which can be helpful in formulating focused questions

regarding healthcare related testing. The guide facilitates guideline developers in defining structured questions by identifying relevant characteristics of the population, tests, and outcomes of interest. This is an essential step in the development of informed, evidence-based, guideline recommendations for healthcare related testing.

References

1. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Clinical Practice Guidelines We Can Trust. In: Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US). Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011 isbn: doi:10.17226/13058.
2. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. (eds.). Evidence-based medicine: How to practice and teach EBM. New York: Churchill-Livingstone; 2000. isbn:978-0443062407.
3. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011;64(4):395-400. doi:10.1016/j.jclinepi.2010.09.012.
4. Bossuyt PM, Deeks JJ, Leeflang MM, Takwoingi Y, Fleming E. Preface. 2023 In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 20 (update July 2023). Cochrane. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
5. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ. 2012;344:e686. doi:10.1136/bmj.e686.
6. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem. 2012;58(12):1636-43. doi:10.1373/clinchem.2012.182576.
7. Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ. 2011;343:d4684. doi:10.1136/bmj.d4684.
8. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jama. 1994;271(5):389-91. doi:10.1001/jama.271.5.389.
9. Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, et al. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PLoS One. 2015;10(10):e0134553. doi:10.1371/journal.pone.0134553.
10. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol. 2012;65(3):282-7. doi:10.1016/j.jclinepi.2011.07.003.
11. Tuut MK, Burgers JS, van der Weijden T, Langendam MW. Do clinical practice guidelines consider evidence about diagnostic test consequences on patient-relevant outcomes? A critical document analysis. Journal of evaluation in clinical practice. 2021 doi:10.1111/jep.13619.
12. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106-10. doi:10.1136/bmj.39500.677199.AE.
13. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. Evidence-based medicine. 2008;13(6):162-3. doi:10.1136/ebm.13.6.162-a.
14. Mustafa RA, Wiercioch W, Cheung A, Prediger B, Brozek J, Bossuyt P, et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 2: a review of methodological and practical challenges. J Clin Epidemiol. 2017;92:18-28. doi:10.1016/j.jclinepi.2017.09.003.
15. Tuut MK, Burgers JS, De Beer JJA, Bindels PJE, Bossuyt PMM, Cals JW, et al. Required knowledge for guideline panel members to develop healthcare related testing recommendations - a developmental study. J Clin Epidemiol. 2024;173 doi:10.1016/j.jclinepi.2024.111438.
16. U.S. Preventive Services Task Force. U.S Preventive Services Task Force Procedure Manual. 2021. Available from: https://www.uspreventiveservicestaskforce.org/uspstf/sites/default/files/inline-files/procedure-manual-2022.pdf.
17. Chang SM, Matchar DB. Methods Guide for Medical Test Reviews. 2012. Available from: https://effectivehealthcare.ahrq.gov/products/collections/methods-guidance-tests.
18. Deeks JJ, Bossuyt PM. Chapter 2: Evaluating medical tests. 2023 In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 20 (updated July 2023). Cochrane. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.

19. Gopalakrishna G, Leeflang MM, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6(9):e010549. doi:10.1136/bmjopen-2015-010549.
20. Tuut MK, de Beer JJA, Burgers JS, van de Griendt EJ, van der Weijden T, Langendam MW. Applying GRADE for diagnosis revealed methodological challenges: an illustrative example for guideline developers. J Clin Epidemiol. 2020;131:123-32. doi:10.1016/j.jclinepi.2020.11.021.
21. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol. 2014;67(7):760-8. doi:10.1016/j.jclinepi.2014.01.006.
22. DECIDE Collaboration. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based Evidence (DECIDE). Available from: http://www.decide-collaboration.eu.
23. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med. 2014;89(9):1245-51. doi:10.1097/ACM.0000000000000388.
24. World Health Organization. WHO handbook for guideline development, 2nd ed. 2014. Available from: https://apps.who.int/iris/handle/10665/145714.
25. GIN-McMaster. GIN-McMaster Guideline Development Checklist (GDC). 2014. Available from: https://cebgrade.mcmaster.ca/guidelinechecklistprintable.pdf.
26. Piggott T, Baldeh T, Akl EA, Junek M, Wiercioch W, Schneider R, et al. Supporting effective participation in health guideline development groups: The Guideline Participant Tool. J Clin Epidemiol. 2021;130:42-8. doi:10.1016/j.jclinepi.2020.07.022.

Appendix 1. Questionnaire used to get feedback at DECIDE the user testing workshop

1a. Do you think the approach we presented today can be useful in understanding the context and place of a test?

1b. Does it give insight into the different types of evidence needed?

2. What do you think of the structure (PICO) and flow of questions in this demo?

3. Would you consider using this approach in developing a guideline about diagnostic tests?

4. What level training would you need to successfully define such a pathway for a guideline?

5. What kind of training would you prefer?
   - ☐ Hands-on workshop
   - ☐ Online training / webinar
   - ☐ Structured guidance (step-by-step user guide)
   - ☐ Video-taped examples
   - ☐ Other, please specify

6. Would you prefer an open (semi structured interview) or structured (checklist) approach?

7. Do you have any suggestions for improvement?

Appendix 2. Interview guides for the user testing sessions with guideline panel members

[translated from Dutch]

First interview (online)

Item
• Intro
• Explanation of the context and purpose of the study (guideline development on healthcare related testing, study to test a step-by-step guide aimed at facilitating the formulation of key questions)
START VIDEO RECORDING
• Would you please formulate a key question on the use of one or more tests for the guideline that you are involved in?
• Thank participant
STOP VIDEO RECORDING
• Explanation of the remainder of the study: you will receive the step-by-step guide by email with a request to read it critically and note questions/clarifications
• Schedule appointment for second (on site) interview
• Thank participant again and close session

Second interview (on site)

Item
START VIDEO RECORDING
• Repeat key question formulated in the first (online) interview
• Concerning the step-by-step guide: - Did you manage to read through/study the step-by-step guide? - Do you have any preliminary questions/comments regarding the step-by-step guide?
• Going through the manual (participant in the lead, interviewer can adjust and possibly provide clarification): - Complete test-management pathway for the same testing situation - Possibly reformulate initial key question - Determine whether and in what way the key question has been changed
• Ask for feedback on the step-by-step guide
STOP VIDEO RECORDING

Appendix 3. Updated step-by-step guide for developing a test-management pathway

Steps	Signalling questions
People (Setting & Timing)	**In whom is testing considered?**
▪ Define the eligibility criteria: in which persons is testing considered? ▪ Define healthcare setting	*Consider personal characteristics, setting, referral patterns, previous test results.* ▪ Are we interested in a particular age, sex or gender? ▪ Have the persons been referred from another setting? Were other tests performed? ▪ In what setting will the persons be tested? (population screening program, general practitioners practice, physiotherapy practice, hospital, etc.) ▪ Should subgroups be considered?
Index test	**Which test or testing strategy is considered?**
▪ Define measurand ▪ Primary purpose of the index test ▪ Define measurement platform or assay(s)	*The guideline panel will have to be specific enough in the description of the test that is considered.* ▪ What is the measurand (the physical quantity or property that is being measured)? ▪ What is the primary purpose of testing (screening, diagnostic, prognostic, predictive, monitoring, etc.) ▪ What is the role of the test relative to other tests (triage, replacement, add-on, parallel/combined) ▪ Is a combination of tests or specific testing strategy considered? (multimarker score, sequence of tests, etc.) ▪ What is the burden associated with the test (efforts to undergo the test, adverse effects, complications, costs, etc.) ▪ Are there any feasibility considerations? (resource requirements, training, storage, transport, etc.) ▪ Are there any acceptability considerations? (patients values and preferences, equity, costs, etc.) ▪ What platform or which kind of assay is used for the measurand?
Outcome(s) of interest	**What is the ultimate goal to achieve, avoid or simplify in people in whom testing is considered**
▪ Define the anticipated or desired impact of testing on downstream (people-important) outcomes ▪ Define the how the index test results can guide (clinical) management decisions	*Guideline panels will likely need an introduction on how to define these outcomes.* ▪ What are the (crucial and important) people-important outcomes that ultimately matter? ▪ How may the index test help to improve, avoid, simplify or these outcome(s)?
Linking outcomes to testing	**How will testing guide further healthcare actions or patient management?**
▪ Link (positive, negative, failed, inconclusive, continuous) test results to management options and people-important outcomes	*Testing in itself rarely leads to the desired outcomes.* ▪ What management options are available after testing, to achieve, avoid, or simplify the people-important outcomes mentioned under c? ▪ What management options may follow the following test results: - For dichotomized test results: positive test result

	- For dichotomized test results: negative test result - For continuous test results: actual test results - Failed tests - Inconclusive test results ▪ What is the target condition or target event? (this may be a disease or disease stage) ▪ What are the consequences of false positive and false negative tests results on people-important outcomes?
Comparator	**What is the alternative to testing?**
▪ Define the existing pathway or the one that would be in place if the index test under (b) was not available	*This refers to the 'C' in the PICO framework, the comparator. The comparator may be the standard of care.* ▪ What is currently being done to achieve, avoid or simplify the people-important outcome(s) mentioned under c? ▪ What type of information guides or would guide management if we did not or do not have the index test results?

Translation in Dutch

Stappen	Vragen
Populatie (Setting & Timing) (P)	**Bij wie wordt testen overwogen?**
▪ Beschrijf de in- en exclusiecriteria: bij welke personen wordt testen overwogen? ▪ Beschrijf de gezondheidszorg setting	*Overweeg persoonskenmerken, setting, verwijzingen, voorgaande testresultaten* ▪ Zijn we geïnteresseerd in een bepaalde leeftijd, geslacht of sekse? ▪ Zijn de personen doorverwezen vanuit een andere omgeving? Zijn er andere tests uitgevoerd? ▪ In welke setting worden de personen getest? (bevolkingsonderzoek, huisartsenpraktijk, fysiotherapiepraktijk, ziekenhuis, etc.) ▪ Moeten subgroepen worden overwogen?
Indextest (I)	**Welke test of teststrategie wordt overwogen?**
▪ Definieer meetgrootheid ▪ Primair doel van de indextest ▪ Leg meetsysteem of assay(s) vast	*De richtlijnwerkgroep moet specifiek genoeg zijn in de beschrijving van de test die overwogen wordt* ▪ Wat is de te meten grootheid (de fysieke grootheid of eigenschap die gemeten wordt)? ▪ Wat is het primaire doel van de test (screening, diagnostisch, prognostisch, voorspellend, monitoring, etc.) ▪ Wat is de rol van de test ten opzichte van andere tests (triage, vervanging, aanvulling, parallel/gecombineerd)? ▪ Wordt een combinatie van tests of een specifieke teststrategie overwogen? (multimarker score, volgorde van testen, etc.) ▪ Wat is de belasting van de test (moeite om de test te ondergaan, bijwerkingen, complicaties, kosten, enz.) ▪ Zijn er implementatieknelpunten? (benodigde middelen, training, opslag, transport, etc.) ▪ Zijn er knelpunten met betrekking tot aanvaardbaarheid? (waarden en voorkeuren van patiënten, rechtvaardigheid, kosten, etc.) ▪ Welk platform of welk soort assay wordt gebruikt voor de te meten grootheid?

Uitkomsten (O)	**Wat is het uiteindelijke doel om te bereiken, vermijden of vereenvoudigen bij mensen bij wie testen wordt overwogen?**
▪ Beschrijf de verwachte of gewenst impact van testen of patiëntrelevante uitkomstmaten ▪ Beschrijf hoe de resultaten van de indextest (klinisch) beleid kunnen bepalen	*Richtlijnwerkgroepen hebben mogelijk een uitleg nodig over het bepalen van patiëntrelevante uitkomstmaten* ▪ Wat zijn de (cruciale en belangrijke) patiëntrelevante uitkomstmaten die uiteindelijk van belang zijn? ▪ Hoe kan de indextest helpen om deze uitkomst(en) te verbeteren, vermijden of vereenvoudigen?
Koppelen van uitkomsten aan testen	**Hoe kan het testen van invloed zijn op het beleid bij de patiënt?**
▪ Koppel testresultaten (positief, negatief, mislukt, inconclusief, continu) aan beleid en patiëntrelevante uitkomstmaten	*Testen zelf leidt zelden tot de gewenste uitkomsten.* ▪ Welke beleidsopties zijn beschikbaar na testen, om de onder c genoemde patiëntrelevante uitkomstmaten te verbeteren, vermijden of vereenvoudigen? ▪ Welke beleidsopties kunnen volgen op de volgende testresultaten: - Voor dichotome testresultaten: positief testresultaat - Voor dichotome testresultaten: negatief testresultaat - Voor continue testresultaten: actuele testresultaten - Mislukte testen - Inconclusieve testen ▪ Wat is de beoogde conditie of gebeurtenis waarop de test is gericht (dit kan bijv. een ziekte of stadium zijn) ▪ Wat zijn de gevolgen van fout-positieve en fout-negatieve testresultaten op patiëntrelevante uitkomstmaten?
Vergelijking (C)	**Wat is het alternatief voor testen?**
▪ Beschrijf de bestaande test-management strategie of de test-management strategie die van toepassing zou zijn als de indextest onder b) niet beschikbaar zou zijn	*Dit verwijst naar de C in de PICO, de controle/vergelijking. Dit kan standaardzorg zijn.* ▪ Wat is de huidige test-management strategie om patiëntrelevante uitkomsten (genoemd bij c) te bereiken, vermijden of vereenvoudigen? ▪ Op basis van welke informatie wordt het beleid (of zou het worden) bepaald als de indextest niet beschikbaar zou zijn?

Appendix 4. Illustrative case of pilot testing

Table 1 describes how the initially broad key question brought in the pilot was clarified by the interview process leading to the identification of specific issues to be clarified. These were included as part of the test-management pathway (*Figure 1*) that were originally not explicitly identified by the user prior to applying this approach.

Table 1. Example of an initially ambiguous key question clarified through test-management pathway development

General topic	**MRI to replace mammography in breast cancer screening**
Initially ambiguous non-specific / broad key question	Will using MRI instead of mammography to screen for breast cancer lead to a higher survival of women at high risk of developing breast cancer?
Key concerns identified through the step-by-step approach of pathway development	
Patients, Setting, Timing (P)	Naïve, high-risk women (i.e. with a family history of breast cancer) below the age of 50 years, identified through general practitioners and then referred to secondary care where they will enter the pathway
Index test(s) (I)	MRI is the replacement test being considered, although not all hospitals will have this facility
Comparison or Existing test/strategy (C)	Mammography is the existing test/strategy and the available treatment options identified were combinations of radiotherapy/chemotherapy/ surgery/immunotherapy depending on tumour type and stage with varying prognosis
Outcomes of interest (O)	▪ Increased (breast-cancer specific / disease-free) survival among high-risk women through early detection ▪ Reduced anxiety of disease and hence improve quality of life through reassurance ▪ Impact on clinical management decisions would involve re-testing after a year for true negatives and false negatives; follow up in six months for false positives
Linking outcomes to test accuracy	▪ True positives: will go through biopsy and receive treatment depending on risk category ▪ True negatives: will be re-tested a year later ▪ False positives: unnecessary biopsy and anxiety; will be monitored in approximately 6 months' time ▪ False negatives: wrongly reassured and may only be diagnosed a year later

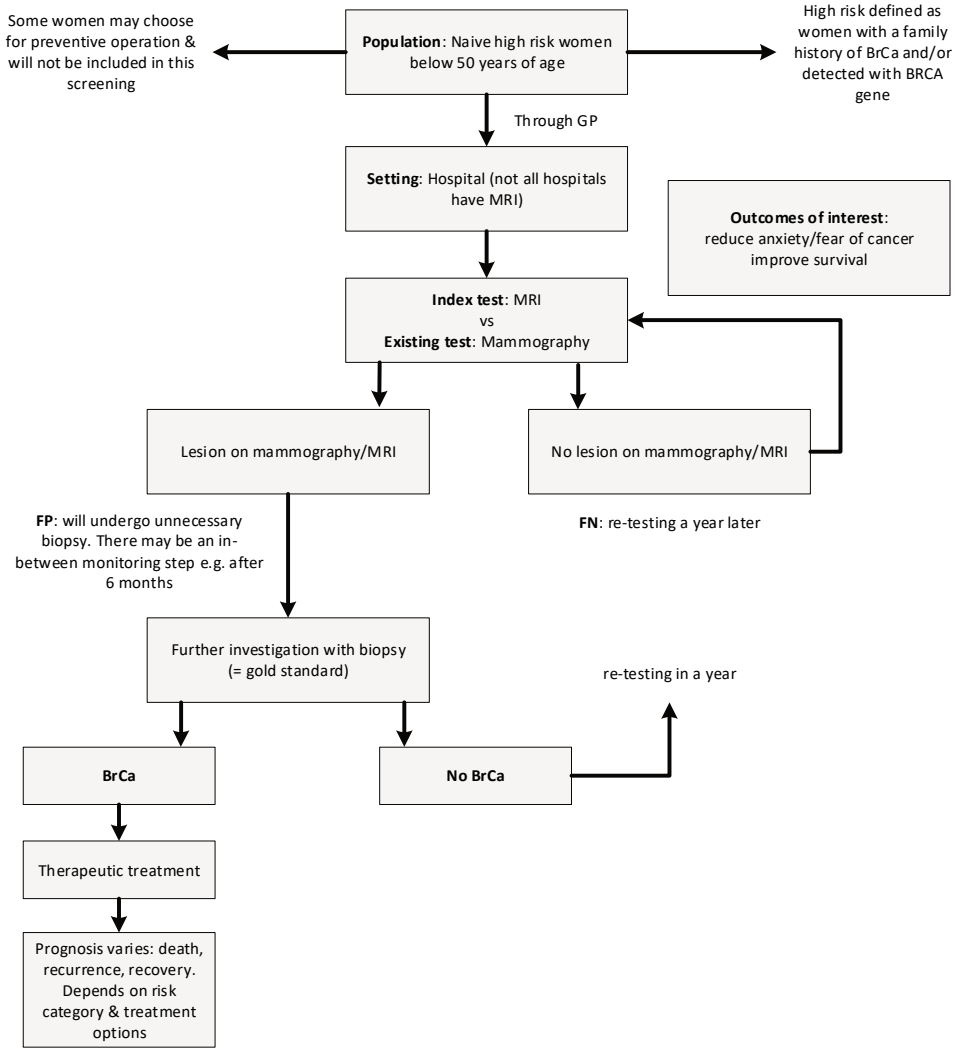


Figure 1. Illustrative example of a test-management pathway developed from user testing: In women with a high risk of developing breast cancer, will using MRI instead of mammography to screen for breast cancer lead to a higher survival of these women

In this example, MRI (the index test) is more sensitive than mammography (the comparator), but it is unclear whether it also leads to a better survival in this patient group. Randomized controlled studies exploring the differences between MRI and mammography on people-important outcomes such as mortality are lacking. The test-management pathway will generally be the same for both tests. Any difference in survival between the two groups (MRI versus mammography) will therefore be mainly driven by differences in test performance (such as sensitivity and specificity) in results between the two tests. However, a more accurate test does not necessarily lead to a

survival benefit. For example, women at high risk of developing breast cancer, because of hereditary factors, may develop a more aggressive tumour leading to a higher mortality, independent of the stage in which it is diagnosed. In that case, screening may not improve survival and a test with a higher accuracy may not have any net health benefit above a test with a lower accuracy.

Appendix 5. Draft step-by-step guide for developing a test-management pathway

Steps	Trigger questions
Patients (Setting & Timing)	
▪ Define patient characteristics ▪ Define the target condition ▪ Define prior tests & setting	**What kind of patients are being considered?** *Consider patient characteristics, setting, referral patterns* **Trigger questions:** ▪ Are you interested in a particular age, gender, etc? ▪ What is the disease or disease stage that the index test is intended to identify? ▪ Have the patients been referred from another place and have other tests been done there? ▪ In what situation will the patients be tested? Is it a screening situation, or at the GP's, or somewhere else? ▪ What is the healthcare setting in which the index test will be applied: community, primary, or secondary care?
Index test	
▪ Role & purpose of the index test ▪ Point in the pathway where the index test might be considered ▪ Test variations, if relevant ▪ Test specifications	**What is the test or tests of interest?** *The guideline development group must have a clear idea of where the new test may be placed in the pathway. There may be a need to go back to this item once the pathway is better defined.* **Trigger questions:** ▪ What is the purpose of the test i.e. diagnostic, prognostic, monitoring etc? ▪ What is the role of a test (i.e. triage, replacement, add-on in comparison to existing test(s))? ▪ What are the test variations i.e. are there different manufacturers of a test, who will be operating and interpreting the test, and is there more than one threshold to be considered? ▪ What are the test specifications to need to be considered (i.e. resource requirements, training, translations, specialized equipment or conditions etc)?
Comparison or Existing test(s)/strategy	
▪ Define the pathway that would be in place if the index test was not available.	**What is the comparison or existing test strategy to avoid/achieve the outcome of interest?** *This refers to the 'C' in the PICO framework, the comparator. The comparator may be standard care.* **Trigger questions:** ▪ What is currently being done to avoid/achieve outcome(s)? ▪ What would we do if we do not use the index test? ▪ What will alternatively guide clinical decision-making? ▪ What treatment options are available?

Outcome(s) of interest to avoid or achieve

- Define the impact of the index test on downstream (patient) outcomes
- Define the impact of the index test on clinical management decisions

What are we trying to avoid, achieve or simplify in patients?
Guideline development groups will likely need an introduction on how to define patient outcomes, and its link to medical testing. There may be different outcomes for different settings.
Trigger questions:
- How may the introduction of the index test help to avoid, simplify or improve these (patient) outcome(s)?
- What is the potential impact of the index test on clinical management decisions e.g. decisions involving referral for further investigation or treatment options etc.?

Linking outcomes to the testing

- What actions follow after the different test results?
- How do these actions impact the four test accuracy categories (TP, TN,FP, FN) and inconclusive test results?
- If possible, one could provide weights to different downstream outcomes

Trigger questions:
For true positives (TP) and false positives (FP):
- Will patients with a positive test be: referred to a specialist, referred for subsequent testing, treated for the condition?
- What should ideally be done in those with the target condition?
- Is there effective treatment available?
- What will be possible outcomes for patients who do not have the target condition and test positive?

For false negatives (FN) and true negative (TN):
- What will happen to those patients who test negative?
- Is it assumed that these patients are 'healthy' or will they probably have another disease than the target condition?
- Will the patients be re-tested in due time?
- How likely is it that this condition will be missed, or will there be a delayed diagnosis in the false negatives?
- What is the prognosis of patients with the target condition if treatment is being withheld?

For inconclusive test results:
- What will happen to these patients, will they be re-tested and within what duration and the number of times

Appendix 6. Detailed feedback on the step-by-step guide from DECIDE workshop participants

Question	Agreement	Feedback
Do you think the approach we presented today is useful in understanding the context and place of a test? (yes/no)	19/19	▪ The pathway definition does not necessarily need to be the first step in guideline development ▪ There may be more than one pathway for a given test strategy ▪ Some pathways maybe challenging due to variation in practice ▪ A multidisciplinary team may be needed going across primary, secondary, and tertiary care in order to establish accurate, comprehensive, factual information on the pathway
Does it give insight into the different types of evidence needed? (yes/no)	14/17	More information and guidance on defining clear inclusion and exclusion criteria for the evidence search is needed
What do you think of the structure (PICO) and flow of questions in this approach?		▪ The visualisation was helpful to get a picture of missing information ▪ The approach is good as an initial starting point to developing the pathway ▪ It is difficult to establish a smooth flow of questioning since discussions tended to go off tangent into individual benefits versus population benefits and/or harms or if an entire panel is involved ▪ There was expectation of a software to support this approach ▪ Starting and ending with the patient outcomes maybe more appropriate ▪ The Interviewer must have some background information about the health problem ▪ The approach should be more about Bayesian steps of pre-test probability estimates derivation and sequence of diagnostic ins- and rule-outs. ▪ Defining the clinical outcomes was difficult
Would you consider using this approach in developing a guideline? (yes/no)	18/19	▪ The process should engage the full guideline panel ▪ Different strategies could be used for getting different perspectives: e.g. focus group or one to one interviewing with an inductive approach ▪ These pathways are not only useful for diagnostic tests but also for treatments ▪ It seems time-consuming with a need for healthcare professionals and guideline panels to be "trained" or "used to" this approach ▪ The approach was not explicit enough on how it can be used in practice to help guideline development / make recommendation ▪ Discussions about the reference standard can cause discussions to go off tangent and to reach consensus

	▪ Starting with defining the patients and not the index test might be more logical
What level of training/support would you need to successfully define such a pathway for a guideline?	▪ Basic understanding of diagnostic test research ▪ Search strategies and selection criteria for studies ▪ Flow of the questions: how to select the most important questions to keep the workload feasible for the review/guideline group ▪ How to best visualize the final output and capture the iterations in between Interviewing and group facilitation skills ▪ Knowledge in Evidence Based Medicine and guideline development ▪ Hands-on one-to-one training or another 'live demonstration'
What kind of training would you prefer?	▪ Hands-on workshop (n=11) ▪ Structured guidance (step-by-step guide) (n=10) ▪ Online training / webinar (n=4) ▪ Video-taped examples (n=4) ▪ Other (journal series, workshops like this, real examples) (n=3)
Would you prefer an open, semi structured or structured (checklist) approach?	▪ Structured approach (n=3) ▪ Semi-structured interview as so many variations and it allows for clarifications (n=9) ▪ Both could be helpful, so that you could choose, depending on the topic and issues addressed, structured approach as starting point with most important questions plus additional questions for anything else at the interview) (n=4) ▪ Unsure (n=1) ▪ No reply (n=2)
Do you have any suggestions for improvement?	▪ Possible harmful effects (that are not foreseen in the initial hypothesis/key question) deserve more attention ▪ Give multiple examples of pathways and how to visualize them effectively ▪ Illustrate pros, cons and practicalities of using this approach using real examples ▪ More focus on contextualizing the question

Chapter 7.

General discussion

General discussion

In this discussion chapter I first summarise the main findings, followed by a reflection on the strengths and limitations of the thesis. In the section 'reflections' I put the findings in perspective, to subsequently come to interpretations considering the strengths and limitations and relevant recent literature, concerning the topics test-management pathway concept, from guideline to practice including the diagnostic process and patient/public information, and overdiagnosis. The discussion ends with a description of the conclusions and recommendations for practice and further research.

This thesis aimed to facilitate and improve guideline development concerning healthcare related testing. The objectives were to explore current practices and challenges, identify required knowledge, and develop and test a tool to formulate appropriate guideline questions. The following research questions were addressed:

1. What are challenges and possible solutions when assessing the certainty of evidence of a test-management pathway?
2. Which types of evidence (diagnostic accuracy, burden of the test, natural course, treatment effectiveness, link between test result and administration of treatment) are used to support guideline recommendations about testing?
3. What is the minimum knowledge required for guideline panel members involved in developing recommendations about testing?
4. Can a step-by-step guide aid guideline developers in formulating key questions about testing?

Main findings

This thesis emphasises the importance of the test-management pathway concept (*figure 1*) in guideline development. The rationale for this concept is that healthcare related tests are typically not used in isolation, and that testing in itself regularly has no direct impact on people-important outcomes.

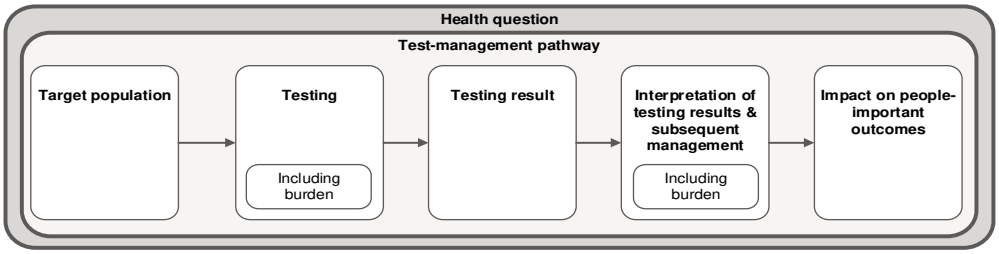


Figure 1. Test-management pathway concept

In order to explore challenges in processing the evidence for tests and testing strategies and suggest solutions, we conducted systematic reviews of all elements of a test-management pathway of an illustrative example using GRADE. Therefore, we analysed the evidence for the clinical question: what is the value of specific immunoglobulin E (sIgE) blood testing as an add-on test to history taking (I) compared to history taking alone (C) in patients suspected of having allergic rhinitis (P), with a relief of nasal or ocular symptoms as critical outcomes, and concentration, sleep problems, work/school absence, and quality of life as important outcomes (O). This study identified challenges and suggested solutions. One major challenge was the lack of high quality evidence in all elements of the test-management pathway, including test burden, natural course of the condition of interest, and the link between test results and people-important outcomes. It is therefore not possible to draw any firm conclusions. Conducting more relevant studies while using the GRADE downgrading factors, such as risk of bias and imprecision as a guidance, is a potential long-term solution in a broader perspective. In the context of guideline development, a broadening of the scope can be a solution, for example, by shifting the focus of test burden from specific to more general. Input from patient advocates could be considered, particularly regarding experiences with test burden. However, it is important to note that this input can not be used to increase the certainty of the evidence. Furthermore, the description of the natural course of the disease of interest lacked transparency, for which we suggest downgrading for indirectness. There was also no evidence found about linking test results to subsequent management. To address this, we suggest concentrating on disease-specific details, such as treatment adherence and difficulties, and discussing these with the guideline panel. Additionally, qualitative evidence could be included. The lack of evidence led to the inability to determine the overall certainty of evidence, as some elements of the test-management pathway were missing. To determine the overall certainty of the evidence, we recommend considering the critical elements for decision making, as suggested by the guideline panel. Finally, the critical appraisal of all elements of the test-management pathway is substantially more time consuming than merely evaluating test accuracy. We propose to focus the discussion on those elements of the test-management pathway that are critical for the decision of whether or not to recommend a test, and to conduct systematic reviews for those elements.

In light of these challenges, we aimed to evaluate the evidence-base of current guidelines about healthcare related testing. We therefore conducted a systematic document analysis of published guidelines for three test questions that diverged in terms of invasiveness, purpose, disease of interest, and costs: C-reactive protein

(CRP) to increase the likelihood of pneumonia, colonoscopy to detect colon cancer, and fractional exhaled nitric oxide (FeNO) to diagnose (severe) asthma. We analysed fifteen publicly available national and international guidelines published between 2016 and 2020 and in force at that time. Ten of these fifteen guidelines assessed the accuracy of the test, but only four of these supported the assessment with a systematic review of the literature, including an evaluation of the certainty of the evidence. The remaining elements of the test-management pathway (i.e. test burden, natural course of the disease of interest, management effectiveness, and linked evidence) were hardly considered in a transparent way.

As critical appraisal of the evidence beyond test accuracy appeared challenging and current practice is suboptimal, we were keen to facilitate the process of guideline development about healthcare related testing. However, existing competency-based frameworks for guideline developers do not adequately address the expertise required for test evaluation. Thus, we conducted a developmental study to determine the knowledge required for guideline panel members to effectively contribute to the development of healthcare related testing recommendations. Based on literature review and nine semi-structured interviews with international experts on the topic, we compiled a list of 26 knowledge components across seven domains: health question, test-management pathway, target population, test, test result, interpretation of test results & subsequent management, and impact on people-important outcomes. For each knowledge component, we defined the necessary level of knowledge. The key component appeared to be understanding and insight into the concept of the test-management pathway, which helps to focus on people-important outcomes. The other required knowledge components, such as the formulation of the purpose and role of a test and the interpretation of false positive and false negative test results in terms of people-important outcomes, fit seamlessly in this concept. In a separate manuscript, we have provided examples of test-management pathways for different test scenarios. These examples can be used by guideline methodologists, guideline panel chairs and trainers to facilitate the understanding of the test-management pathway concept by guideline panel members.

Since the test-management pathway concept is a crucial knowledge component, we conducted a study to develop and test a step-by-step guide for formulating focused questions about healthcare related testing through drafting such a test-management pathway. This study was already initiated over a decade ago as part of the DECIDE-project, an EU-funded project aimed at developing and evaluating methods for disseminating guidelines, including evaluating evidence and developing recommendations for healthcare related tests [1]. During the DECIDE-project, which

we developed the draft version of the guide and tested it among experts in a workshop. The guide was recently refined and updated, and then tested among seven guideline panel members. As a result, an introduction, instructions and a visualisation of the test-management pathway were added. The final step-by-step guide assists guideline developers in formulating structured questions by identifying important characteristics of the population, the test(s), people-important outcomes, and the link between testing and these outcomes.

Strengths and limitations

The case study on evaluating sIgE for diagnosing allergic rhinitis highlights challenges faced by guideline methodologists and offers practical solutions. However, we did not consult guideline panels, which could have been more efficient, for example in determining people-important outcomes and discussing the critical elements of the test-management pathway. Although only one case was systematically analysed, the results of the study reflect a wide range of guidelines, as the challenges identified are common and the proposed solutions are considered feasible based on own experience.

Our document analysis study, which demonstrated that guidelines on testing are not transparently based on evidence for all parts of the test-management pathway, showed consistent results across the three different tests, supporting the generalisability of the results. However, if we had been able to include tests with other than diagnostic purposes, this might have widened the scope. For example, recommendations for monitoring tests may place more emphasis on, for example, test burden (e.g. frequent visits) and management effectiveness. Another weakness of this study is the lack of information about the dynamics in the guideline panels. It is therefore unclear whether a guideline panel considered test consequences, such as test burden, management effectiveness, and linked evidence, or based their recommendations solely on information about test accuracy. It is possible that guideline panels did consider these consequences but chose not to include them in the published guideline. This may be the case because they did not feel it was necessary or because they deemed some parts of the test-management pathway were irrelevant to the topic of interest.

Our research on the knowledge required for developing healthcare related testing guideline recommendations was conducted through a purposeful combination of literature review and interviews with nine international opinion leaders from various countries and perspectives. These perspectives include test evaluation, guideline development about testing, and consumer involvement. A limitation of this study is

that it only focused on the cognitive domain of required competencies. In guideline development, skills and attitudes are also important. However, compared to other domains of guidelines, such as treatment guidelines, the required skills and attitudes are not expected to be significantly different for guidelines on testing. Competencies for guideline development are described in various tools such as the Guideline Participation Tool and the Checklist for Guideline panel Chairs, and incorporated in the GIN-McMasters Guideline Development Checklist and include committing to the process, being clear on roles, familiarising with guideline methodology, preparing for meetings, contributing to discussions in a fair and equitable manner, maintain confidentiality, and being respectful [2-5].

In our study on the step-by-step guide for formulating focused questions, we tested the guide among guideline experts and regular guideline panel members, which supports its validity. However, it is important to note that some of the data collected was conducted over a decade ago, which may now be outdated. Therefore, we have incorporated new insights into the final guide. Furthermore, the guide was tested in controlled settings rather than in a real guideline panel. Implementing the guide in guideline panels may require additional attention.

Overall, this thesis combines various study designs, including literature and document analyses, as well as developmental studies. Insights from the updated framework for developing and evaluating complex interventions from the Medical Research Council (MRC) were utilised [6]. The MRC framework is based on dynamic, iterative, and creative principles. When planning the development of an intervention, the first step is to have a thorough understanding of the problem and the potential for an intervention [7]. This involves analysing the problem and using collaborative and user-centered approaches to customise interventions to the context of guideline development [8]. The research presented fills a gap in knowledge and addresses a practical need. All studies were conducted with the aim of international applicability, requiring no local adaptation (except for possible translations).

This research focuses on guideline panel members, including healthcare professionals and patient representatives, and guideline panels as a whole, particularly with regard to the knowledge components and the step-by-step guide. This approach aligns with competency-based frameworks for guideline development and training courses, such as the INGUIDE Certified Guideline Panel Member Course [9, 10]. In the interview study, where we identified required knowledge components to develop guideline recommendations for healthcare related testing, there was some debate about whether the requirements for patient representatives would be the same as for

healthcare professionals in a guideline panel. This was not systematically analysed, however, all interviewees agreed that the test-management pathway concept should be understood by everyone in the guideline panel. Additionally, it would have been beneficial to investigate systematically if the support needs of patient representatives differ from those of healthcare professionals in a guideline panel. The GIN Public Toolkit provides practical advice for guideline developers on involving patients and public in guideline activities. It was developed by the Guidelines International Network and offers for example best practices and tips for successful patient and public involvement [11]. It may be worth considering whether these tools should be adapted for guidelines that include key questions about healthcare related tests. Furthermore, it is important to note that other roles within the guideline panel, such as guideline methodologists and guideline panel chairs, may require different knowledge and tools to optimize their work. These competencies and tools are not included in this thesis.

During the first phase of this thesis, we identified challenges and suboptimal practices. In the second phase, we provided knowledge components and a tool to facilitate the development of guidelines for healthcare related testing. It is unclear whether implementing the required knowledge, such as through training, and following the step-by-step guide will actually enhance the guideline development process and result in better recommendations for healthcare related testing to improve people-important outcomes. Further studies will be required to determine whether and how these instruments improve guideline panel processes and future guideline recommendations.

This thesis presents a series of studies that concentrate on creating guidelines for healthcare related testing. The evaluation of evidence regarding the benefits and harms of testing in a test-management pathway is a major focus, with particular emphasis on people-important outcomes. This is a crucial aspect of guideline development, highlighting the challenges involved, which is a strength of the thesis. However, it is important to acknowledge that the process of moving from evidence to recommendations may present additional challenges. A weakness of this thesis is that these considerations were not included in the research. It is unclear whether such considerations vary between treatment guidelines and testing guidelines. The GRADE working group has identified several factors that should be considered when formulating guideline recommendation, including values and preferences, resource use, cost-effectiveness, equity, acceptability, and feasibility [12-19]. These factors are addressed in the evidence-to-decision framework (EtD) and apply to both testing and treatment recommendations [1]. However, it is unclear whether addressing these factors in the development of testing recommendations requires specific knowledge

or tools. For instance, developing screening recommendations may introduce additional issues when accounting for variability in values and preferences, as well as costs. For example, individuals should have the ability to make an informed decision regarding whether or not to undergo screening tests, based on their preferences, such as a desire to be fully informed or a wait-and-see approach when experiencing symptoms. This could impact the considerations that need to be made when moving from evidence to a recommendation, as well as the formulation of recommendations, which may require testing specific knowledge or tools.

The author of this thesis is an experienced independent self-employed guideline methodologist, who has worked for various Dutch organisations, in guideline development, training, coaching, improving and facilitating guideline development methods for over 25 years. This experience strengthens the thesis by aiding in problem structuring, identifying relevant research questions as well as potential interventions, collecting data, and implementing results. However, over-engagement could also lead to conflicts of interest, blind spots, and selection bias in data collection. To reduce these risks, the studies conducted in this thesis involved a critical supervisory team and several authors with diverse perspectives.

Reflections

Test-management pathway concept

This thesis highlights the relevance of the test-management pathway concept (*figure 1*). When reflecting on this concept over the years, a few things emerge.

First, the language has become more inclusive. This is consistent with the trend to use ‘guidelines’ as an umbrella term, rather than clinical practice guidelines or medical guidelines, to include, for example, public health and non-clinical professionals. An example of this more inclusive language is the term test-management pathway itself, which has previously been called test-treatment pathway, test-treatment strategy, management pathway, care pathway or clinical pathway [20]. The term test-management pathway is more neutral and includes populations other than patients, such as public or consumers, as well as other actions following test results rather than treatment, such as further testing or watchful waiting. The same principle applies to the target population and people-important outcomes, which were previously referred to as patients, and patient important outcomes or patient relevant outcomes, respectively. To be as inclusive as possible, one could also argue for ‘outcomes important to people (who receive the test offered), relatives (for example, in genetic testing for inherited diseases), society (to include public impact, for example, relevant

to infection control measures, and forensic medicine and because of resources), and environment (to include sustainability)'. As an alternative, these aspects may also be considered when moving from evidence to decision (e.g. in considerations of acceptability and resources). Finally, the term 'test' is now referred to as 'testing', as this can include, for, example, a cascade of testing procedures, different frequencies of testing, and tests from different manufacturers. In terms of test performance, there has been a shift from diagnostic accuracy to test accuracy and clinical performance.

Secondly, there has been a recurring debate regarding the definition of test burden. From an individual patient or consumer perspective (e.g. in the consultation room), burden can be considered as the practical demands that come with undergoing the process of a test that people may dislike, such as taking medication to prepare for a test or visiting the hospital [17]. In a broader sense, stress and costs associated with undergoing the test procedure can also be considered. From a population perspective, burden can also be defined as any undesirable aspect and consequence of testing, including adverse effects and complications related to the test. However, adverse effects and complications may also be considered people-important outcomes, possibly depending on their severity and prevalence. Agreeing clear definitions with all relevant stakeholders can solve any potential confusion. Thirdly, there is a noticeable discrepancy between the various purposes of testing and the evidence that has been published. The vast majority of literature on test evaluation and guideline development methods concerning testing focuses on diagnosing diseases [19, 21-23]. However, in healthcare practice and guideline development, it is also necessary to clarify the value of testing for other purposes. This includes among others prognostic and follow-up tests, such as a test to predict the likelihood of a particular event such as stroke, or a test to monitor the course of certain diseases, such as lung function in COPD. Additionally, it is possible for a single test to serve multiple purposes. For instance, mammography can be used as a screening instrument in women aged 50-75 years, as a diagnostic step in women with signs of breast cancer, and as a follow-up measure in women who have undergone breast cancer treatment. It is important for guideline developers to acknowledge the various possible purposes of testing.

And fourth, when assessing the added value of tests, it is noteworthy that both test evaluation methods and guideline development methods, as well as this thesis, place great emphasis on measures of test accuracy, implying a dichotomy. However, it is important to note that test results can fall into different data types: they can be binary (e.g. a pregnancy test), ordinal (e.g. BI-RADS assessment categories ranging from 0 to 6), counts (e.g. complete blood count), or continuous (e.g. body temperature). Furthermore, it should be realised that in healthcare practice, a test is usually

integrated into a testing pathway, such as a diagnostic pathway that includes patient history and physical examination, rather than standing on its own. It is important to acknowledge that testing evaluation in a scientific or guideline development manner occurs in a simplified version of reality. Guideline developers should be aware of this and incorporate real-life practical aspects into their considerations and recommendations. This may be increasingly important when looking to the future, particularly as algorithms and genomics-based personalised medicine emerge.

From guidelines to practice

Moving from evaluating the test-management pathway in the context of guideline development to the impact of guideline recommendations on healthcare related testing on people-important outcomes in practice, effective implementation strategies are crucial. Guidelines can only be effective if they are able to change the behaviour of healthcare professionals and consumers. The implementation of guidelines involves the use of various tools, that target different aspects of implementation, such as dissemination, understanding, adoption, and putting into practice [24]. Two aspects that deserve special attention are highlighted in the following sections: the diagnostic process and patient/public information.

Diagnostic process

In healthcare practice, the diagnostic process is an empirical and iterative process [25]. It involves both inductive and deductive elements, based on Bayes' theorem [26]. Generally inductive processes, such as routine testing, can be seen as hypothesis generation. Additionally, deductive processes, such as specific testing, can be seen as hypothesis testing, to confirm or rule out a specific diagnosis. This entire diagnostic process in the consultation room is known as the hypothetico-deductive method [27, 28].

The diagnostic process involves uncertainty, including the interpretation and integration of information, the formulation of diagnoses, and communication with patients. However, it is important to note that patients and healthcare professionals experience different aspects of uncertainty. Patients are often unaware of uncertainties in the diagnostic process [29]. To manage diagnostic uncertainty, healthcare professionals frequently use patient-centred communication strategies, such as empathy, and diagnostic reasoning strategies, such as exclusion of serious diagnoses. Patient reactions and experiences related to diagnostic uncertainty are mixed, indicating variable tolerance for uncertainty [30].

Patient/public information

Guidelines can be used to inform shared decision making. Patient versions of guidelines, which translate recommendations into simple language, are commonly used to inform patients and the public about information in guidelines that is important to them. While the development of guidelines follows strict criteria, there is a heterogeneous methodology for developing patient versions [31, 32]. Patient decision aids are additional tools derived from guidelines. These are intended to assist patients in making an informed decision about a specific preference sensitive recommendation from a guideline. The use and effectiveness of patient decision aids have primarily been evaluated in the context of management decisions, such as oncological or orthopaedic treatment. However, evidence for decision aids regarding testing is limited, although promising studies are being undertaken [33]. This is important, since research has shown that the current practices for communicating the downsides of testing are suboptimal. For example, decision aids used to support shared decision-making on prostate cancer screening often lack information on possible overdiagnosis [34]. Additionally, evidence suggests that the application of shared decision-making can be improved, particularly among people with limited health literacy [35].

It is important to raise awareness about the downsides of testing not only in scientific and official healthcare publications but also in lay press. For-profit testing centers tend to avoid communicating the harms and other negative consequences of testing, and information about the downsides of testing has not yet reached a wide audience of healthcare consumers [36, 37]. Additional publications in public media could raise awareness and comprehension of the significance of the advantages and disadvantages of testing, promoting rational testing and potentially decreasing both under- and overtesting.

Overdiagnosis

In healthcare, there is a growing recognition that excessive healthcare interventions may result in avoidable costs and potentially harm for patients and other healthcare consumers. The Choosing Wisely initiative, which encourages discussions between healthcare professionals and patients regarding commonly used treatments and tests lacking strong supporting evidence, has raised awareness of this issue [38]. However, its implementation is lagging behind [39].

As awareness of low-value care has increased, so has attention to overdiagnosis. This refers to the identification of problems that are not causing harm or the medicalisation of ordinary events or results through expanded definitions of diseases. This can cause more harm than benefit [40]:

- Overdetection is the identification of abnormalities that are unlikely to cause harm, for instance because of absent or slow progress, or spontaneous recovery. Examples include full-body scanning, which may reveal non-progressive tumours (known as 'incidentalomas') or growing cultures of saliva in self-limiting upper airway tract infections.
- Overdefinition occurs when the threshold for a risk factor is lowered without evidence of net benefit, or when the definition of a disease is expanded to include people with ambiguous or very mild symptoms. Examples include the definition of hypertension (which is <130/80 according to the American College of Cardiology and the American Heart Association, and <140/90 according to the European Society of Hypertension) [41, 42], diagnosing pre-diabetes [43], or Alzheimer's disease [44].

Overdiagnosis can lead to labelling (including stigmatisation) and overtreatment, which can have negative physical, mental, social, and financial impact on patients [40]. Rates of overdiagnosis exist for various conditions:

- In Australia, estimates suggest that overdiagnosis occurs in 18% of all cancer diagnoses in women and 24% in men. The most commonly affected types are renal, thyroid, melanoma, breast, and prostate cancer [45].
- A recent meta-analysis found that using the LLN (lower limit of normal) definition resulted in overdiagnosis of COPD in an average of 48% of cases, with outliers above 60% in primary care. However, when using the GOLD criteria, the prevalence of COPD overdiagnosis was significantly lower. Overdiagnosis was also found to be associated with inappropriate treatment [46].
- Approximately 5% of adult patients who self-report a penicillin allergy are truly allergic to penicillins [47]. Overreporting of penicillin allergy leads to the prescription of non-first choice antibiotics, which can contribute to the development of antimicrobial resistance.

Implementing the test-management pathway in healthcare policymaking may reduce overtesting, overdiagnosis, and consequently overtreatment. This involves evaluating the net benefit of testing on people-important outcomes in guideline development. This approach aligns with updated guidance for systematic reviewers and guideline developers [19, 21, 48, 49].

Conclusions

The research conducted in this thesis identified challenges and proposed suggestions to overcome these challenges, including the suggestion to focus on those elements of the test-management pathway that drive the decision of whether or not to recommend

a test. In order to facilitate effective and efficient guideline development for healthcare related testing, it appeared crucial to transparently present all considerations. In addition, this thesis determined the knowledge required to adequately develop healthcare related testing guideline recommendations and provided a tool to facilitate the specification of the test-management pathway. This facilitates considering people-important outcomes when formulating guideline questions on healthcare related testing.

If 'good guidelines can only make you better' [50], 'guidelines would be as transparent as possible in revealing dilemmas and uncertainties' [51], and 'the challenge of scientific research is to formulate the question' [52], then this thesis adds evidence to the growing pile of knowledge about appropriate guideline development on healthcare related testing. It contributes to the goal of providing healthcare professionals and consumers with trustworthy guidance on testing, with the aim of reducing or preventing overtesting and undertesting, and associated overtreatment and undertreatment, and improving people-important outcomes.

Recommendations for practice

The aim of this thesis is to facilitate and improve guideline development concerning healthcare related testing. To achieve this, the results of the research conducted should be integrated into practice. The key message is to incorporate the test-management pathway concept into the development process of guidelines on healthcare related testing. This can be accomplished through two main activities:

- To improve the knowledge of guideline developers regarding the importance of the test-management pathway, including related knowledge, guidance on guideline development should be updated, and this topic should be included in the training of guideline panel members and methodologists.
- To facilitate the specification of a tests-management pathway during the question formulation phase by guideline panels, an online tool should be created. This tool should preferably be integrated into existing guideline development software.

These recommendations for practice are discussed further in the impact chapter of this thesis.

Recommendations for research

Further research is necessary to continuously improve and facilitate the development of guidelines for healthcare related testing. This includes identifying the required knowledge for guideline methodologists and panel chairs to develop appropriate recommendations on testing, in addition to the knowledge components for guideline

panel members. An identical study design could be used, with input from international experts in this field. This would enable and facilitate the implementation of the test-management pathway concept into guideline development. Furthermore, there is a lack of evidence regarding the needs of guideline panel members and other people involved in guideline development. Determining these needs would enable the development and delivery of appropriate training. However, it is possible that there may be a wide range of needs, for example depending on educational background and role of guideline panel members, necessitating tailored training. Therefore, to design such training, methods from educational design research seem indicated. Purposes of such training ('why') should be linked to actions ('how'), supported by arguments. Such methods includes collaborative, user-centered, and iterative approaches to customise interventions to the context of guideline development [8].

Additionally, it is recommended to evaluate the effectiveness of the step-by-step guide for specifying the test-management pathway in 'living labs' rather than in a controlled setting with a single guideline panel member. This involves assessing the use of the guide, as well as the factors that facilitate or impede its implementation, in real-world guideline development settings, specifically within guideline panels. Ideally, this evaluation should be conducted using an online version of the tool, which is expected to be more efficient and interactive.

Finally, in addition to the step-by-step guide for specifying the test-management pathway, it may be beneficial to develop and test additional tools to facilitate guideline developing on healthcare related testing. For example, tools could be created to aid in the interpretation of false positive, false negative, and inconclusive test results, or for interpreting evidence related to the elements of the test-management pathway, and tested in a user-centered design. Furthermore, it would be useful to investigate whether particular instruments are necessary to move from evidence to recommendations when creating guidelines for healthcare related testing.

References

1. DECIDE Collaboration. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based Evidence (DECIDE). Available from: http://www.decide-collaboration.eu.
2. GIN-McMaster. GIN-McMaster Guideline Development Checklist (GDC). 2014. Available from: https://cebgrade.mcmaster.ca/guidelinechecklistprintable.pdf.
3. Piggott T, Baldeh T, Akl EA, Junek M, Wiercioch W, Schneider R, et al. Supporting effective participation in health guideline development groups: The Guideline Participant Tool. J Clin Epidemiol. 2021;130:42-8. doi:10.1016/j.jclinepi.2020.07.022.
4. Schunemann H, Akl E. Checklist for Guideline Panel Chairs, v 3.1. 2017. Available from: https://macgrade.mcmaster.ca/resources/checklist-for-guideline-panel-chairs/.
5. Schunemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2014;186(3):E123-42. doi:10.1503/cmaj.131237.
6. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061. doi:10.1136/bmj.n2061.
7. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open. 2019;9(8):e029954. doi:10.1136/bmjopen-2019-029954.
8. McKenney S, Reeves TC. Educational design research: Portraying, conducting, and enhancing productive scholarship. Med Educ. 2021;55(1):82-92. doi:10.1111/medu.14280.
9. Sultan S, Morgan RL, Murad MH, Falck-Ytter Y, Dahm P, Schünemann HJ, et al. A Theoretical Framework and Competency-Based Approach to Training in Guideline Development. J Gen Intern Med. 2020;35(2):561-7. doi:10.1007/s11606-019-05502-9.
10. INGUIDE. International Guideline Training and Certification Program. Available from: www.inguide.org.
11. Guidelines International Network. GIN Public Toolkit: patient and public involvement in guidelines. 2024. Available from: https://g-i-n.net/wp-content/uploads/2024/01/Toolkit-combined_revJAN24.pdf.
12. Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032.
13. Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. J Clin Epidemiol. 2019;111:94-104. doi:10.1016/j.jclinepi.2018.01.013.
14. Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66(7):726-35. doi:10.1016/j.jclinepi.2013.02.003.
15. Zhang Y, Coello PA, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. J Clin Epidemiol. 2019;111:83-93. doi:10.1016/j.jclinepi.2018.05.011.
16. Pottie K, Welch V, Morton R, Akl EA, Eslava-Schmalbach JH, Katikireddi V, et al. GRADE equity guidelines 4: considering health equity in GRADE guideline development: evidence to decision process. J Clin Epidemiol. 2017;90:84-91. doi:10.1016/j.jclinepi.2017.08.001.
17. GRADE Working Group. GRADE Handbook. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013. Available from: https://gdt.gradepro.org/app/handbook/handbook.html.
18. Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. BMJ. 2016;353:i2089. doi:10.1136/bmj.i2089.
19. Schunemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82. doi:10.1016/j.jclinepi.2019.02.003.

20. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21-35. doi:10.1016/s0749-3797(01)00261-6.
21. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (eds.). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023). Cochrane; 2023. Available from: https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
22. Hultcrantz M, Mustafa RA, Leeflang MMG, Lavergne V, Estrada-Orozco K, Ansari MT, et al. Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper. J Clin Epidemiol. 2020;117:138-48. doi:10.1016/j.jclinepi.2019.05.002.
23. El Mikati IK, Morgan RL, Murad MH, Sultan S, Falck-Ytter Y, Mustafa RA. Testing guidelines during times of crisis: challenges and limitations of developing rapid and living guidelines. Clin Microbiol Infect. 2023;29(4):424-8. doi:10.1016/j.cmi.2023.01.020.
24. Wensing M, Grol R (eds.). Implementatie. Effectieve verbetering van de patiëntenzorg. Achtste, herziene druk. 8 ed.: Bohn Stafleu van Loghum; 2023. isbn:9789036829083.
25. Norman G, Barraclough K, Dolovich L, Price D. Iterative diagnosis. BMJ. 2009;339:b3490. doi:10.1136/bmj.b3490.
26. Wulff HR. (eds.). Principes van klinisch denken en handelen; Nederlandse bewerking. Utrecht: Bohn, Scheltema & Holkema; 1980. isbn:90 313 0399 2.
27. Brush JE, Jr., Sherbino J, Norman GR. Diagnostic reasoning in cardiovascular medicine. BMJ. 2022;376:e064389. doi:10.1136/bmj-2021-064389.
28. Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. BMJ. 2002;324(7339):729-32. doi:10.1136/bmj.324.7339.729.
29. Meyer AND, Giardina TD, Khawaja L, Singh H. Patient and clinician experiences of uncertainty in the diagnostic process: Current understanding and future directions. Patient Educ Couns. 2021;104(11):2606-15. doi:10.1016/j.pec.2021.07.028.
30. Dahm MR, Cattanach W, Williams M, Basseal JM, Gleason K, Crock C. Communication of Diagnostic Uncertainty in Primary Care and Its Impact on Patient Experience: an Integrative Systematic Review. J Gen Intern Med. 2023;38(3):738-54. doi:10.1007/s11606-022-07768-y.
31. Meyer N, Hauprich J, Breuing J, Hellbrecht I, Wahlen S, Konsgen N, et al. Barriers and facilitators in developing patient versions of clinical practice guidelines - qualitative interviews on experiences of international guideline producers. BMC health services research. 2024;24(1):78. doi:10.1186/s12913-023-10524-5.
32. Meyer N, Hellbrecht I, Breuing J, Hauprich J, Wahlen S, Konsgen N, et al. Heterogeneous methodology in the development of patient versions of clinical practice guidelines: a scoping review. J Clin Epidemiol. 2023;161:53-64. doi:10.1016/j.jclinepi.2023.07.005.
33. Linden I, Wolfs C, Perry M, Metsemakers J, van der Weijden T, de Vugt M, et al. Implementation of a diagnostic decision aid for people with memory complaints and their general practitioners: a protocol of a before and after pilot trial. BMJ Open. 2021;11(6):e049322. doi:10.1136/bmjopen-2021-049322.
34. Pathirana TI, Pickles K, Riikonen JM, Tikkinen KAO, Bell KJL, Glasziou P. Including Information on Overdiagnosis in Shared Decision Making: A Review of Prostate Cancer Screening Decision Aids. MDM Policy Pract. 2022;7(2):23814683221129875. doi:10.1177/23814683221129875.
35. Van der Weijden T, Van der Kraan J, Brand PLP, Van Veenendaal H, Drenthen T, Schoon Y, et al. Shared decision-making in the Netherlands: Progress is made, but not for all. Time to become inclusive to patients. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2022;171:98-104. doi:10.1016/j.zefq.2022.04.029.
36. Prescan. Prescan. In één dag inzicht in je gezondheid. Available from: www.prescan.nl.
37. Srivastava R. Next time your doctor orders a scan, know the benefits but don't forget to ask about the harm. The Guardian. 2024 30-01-2024.
38. ABIM Foundation. Choosing Wisely. Philadelphia 2023. Available from: https://www.choosingwisely.org/.
39. Tian EJ, Nguyen C, Chung L, Morris C, Kumar S. The Effectiveness of Public Awareness Initiatives Aimed at Encouraging the Use of Evidence-Based Recommendations by Health Professionals: A Systematic Review. J Patient Saf. 2024 doi:10.1097/PTS.0000000000001202.
40. Brodersen J, Schwartz LM, Heneghan C, O'Sullivan JW, Aronson JK, Woloshin S. Overdiagnosis: what it is and what it isn't. BMJ Evid Based Med. 2018;23(1):1-3. doi:10.1136/ebmed-2017-110886.

41. Whelton PK, Carey RM, Aronow WS, Casey DE, Jr., Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2018;138(17):e426-e83. doi:10.1161/CIR.0000000000000597.
42. Mancia G, Kreutz R, Brunstrom M, Burnier M, Grassi G, Januszewicz A, et al. 2023 ESH Guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension: Endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA). J Hypertens. 2023;41(12):1874-2071. doi:10.1097/HJH.0000000000003480.
43. Fukunaga A, Inoue Y, Nakagawa T, Honda T, Yamamoto S, Okazaki H, et al. Diabetes, Prediabetes, and the Risk of a Composite Outcome of Long-term Sickness Absence and Pre-retirement Death Due to Physical Disorders. J Epidemiol. 2024;34(3):105-11. doi:10.2188/jea.JE20220245.
44. Widera E. Who gets to decide on what it means to have Alzheimer's disease? J Am Geriatr Soc. 2024 doi:10.1111/jgs.18793.
45. Glasziou PP, Jones MA, Pathirana T, Barratt AL, Bell KJ. Estimating the magnitude of cancer overdiagnosis in Australia. The Medical journal of Australia. 2020;212(4):163-8. doi:10.5694/mja2.50455.
46. Fiore M, Ricci M, Rosso A, Flacco ME, Manzoli L. Chronic Obstructive Pulmonary Disease Overdiagnosis and Overtreatment: A Meta-Analysis. J Clin Med. 2023;12(22) doi:10.3390/jcm12226978.
47. Stichting Werkgroep Antibiotica Beleid. The Dutch Working Party on Antibiotic Policy (SWAB) guideline for the approach to suspected Antibiotic Allergy. 2022. Available from: https://swab.nl/nl/exec/file/download/192.
48. Schunemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020 doi:10.1016/j.jclinepi.2019.12.020.
49. Schunemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. 2020;122:142-52. doi:10.1016/j.jclinepi.2019.12.021.
50. Burgers JS. Quality of clinical practice guidelines. Nijmegen: Catholic University Nijmegen; 2002.
51. Van der Weijden T. Richtlijnen in de spreekkamer, van dogma naar dans. Maastricht: Maastricht University; 2010.
52. Langendam MW. The impact of harm reduction-based methadone treatment on HIV infection and mortality. Amsterdam: University of Amsterdam; 2000.

Impact

Impact

The aim of this thesis is to facilitate and improve guideline development concerning healthcare related testing. To achieve this goal, the research findings presented in this thesis should be implemented in practice to have an impact on guideline development, and ultimately, on healthcare quality and people-important outcomes. This chapter discusses the valorisation potential of the research conducted in this thesis.

The thesis examines the challenges of developing guidelines for healthcare related testing and proposes solutions to overcome these challenges. Additionally, it defines the knowledge required for developing guideline recommendations on healthcare related testing and provides a tool to facilitate the specification of the test-management pathway to achieve impact on people-important outcomes. The results described in this thesis may be of relevance to various groups including 1) healthcare professionals and healthcare consumers, especially those participating in guideline panels, 2) guideline methodologists and chairs, and 3) guideline trainers.

Healthcare professionals and healthcare consumers

Healthcare professionals and consumers can benefit from the research conducted in this thesis, either directly through their participation in guideline panels, or indirectly in healthcare practice. The thesis focuses on guideline development, which occurs in guideline panels. Guideline panel members will be better able to fulfil their role if they are equipped with the necessary knowledge. For example, if guideline panel members are fully aware that the clinical effectiveness of testing is determined by evaluating the test-management pathway and that guideline panel members are able to interpret false positive and false negative test results in terms of people-important outcomes, they may be less likely to rely on test accuracy results solely. Additionally, the step-by-step guide for specifying the test-management pathway can assist guideline panel members in formulating focused questions about healthcare related testing. An online tool could further facilitate this process. The creation of such a tool is a priority in projects aimed at facilitating the implementation of methods. This tool could be integrated in software, such as the guideline development tool (GRADEpro by McMaster University and Evidence Prime).

In healthcare practice, healthcare professionals and consumers could also benefit from the knowledge generated in this thesis. If implemented properly, guidelines on healthcare related testing would be more transparent about the net benefits of testing, based on the evaluation of the evidence throughout the entire test-management pathway. This could result in more detailed guideline recommendations concerning testing, which may result in more awareness about the benefits and harms of testing

among healthcare professionals and consumers, the end-users of the guidelines. This in turn would have an impact on the quality of healthcare. For example through ultimately supporting informed and shared decision-making about testing in healthcare practice, and possibly reducing overdiagnosis and subsequent overtreatment.

Guideline methodologists and chairs

The results of this thesis could increase awareness among guideline methodologists and guideline panel chairs of the additional challenges involved in developing guidelines on healthcare-related testing beyond developing guidelines in general. Furthermore, it is important for them to recognise that guideline recommendations on healthcare related testing often fail to consider important factors necessary for adequate development, such as consequences of testing.

Methodologists and chairs could use the defined knowledge components in their instructions to guide panel members in developing proper guidelines on healthcare related testing. The examples provided in this thesis can facilitate the uptake of the test-management pathway concept in this educational process. Furthermore, guideline methodologists and guideline panel chairs can use the step-by-step guide for specifying a test-management pathway. This will help identifying focused questions about healthcare related testing, in collaboration with guideline panel members.

As previously mentioned, an online tool could aid in this process and could be integrated into guideline development software that is available on international level, such as the guideline development tool (GRADEpro). On a national level, initiatives are being taken to implement the required knowledge components for guideline panel members to adequately develop guideline recommendations about healthcare related testing and the step-by-step guide to specify the test-management pathway. Both topics are on the agenda for a Dutch Guideline Network thematic meeting (GENEVER). GENEVER is a networking community, within ‘Richtlijnen Netwerk Nederland’ (Dutch Guideline Network) that is easily accessible to professionals interested and/or experienced in guideline development and/or implementation. The bi-annual GENEVER meetings are well-attended by guideline methodologists and other professionals working in guideline development from various Dutch guideline organisations. Additionally, this thesis provides new knowledge that could be incorporated into the Dutch GRADE manuals and tools for developing guidelines on healthcare related testing [1, 2]. These reports have been developed by the Dutch GRADE Network, a formal entity of the international GRADE working group. Moreover, the new insights from this thesis could be embedded in the update of the ‘AQUA-

Leidraad', the Dutch 'guideline for guidelines', which is regularly updated by Dutch guideline developers from multiple organisations, brought together in the 'Richtlijnen Netwerk Nederland' (Dutch Guideline Network) [3].

Guideline trainers

The research conducted in this thesis can be used by guideline trainers to educate and train guideline panel members, guideline methodologists, and guideline panel chairs. The defined knowledge components required to adequately develop guideline recommendations on healthcare related testing can serve as learning objectives in course and training material development. The examples presented in this thesis can aid in the adoption of the test-management pathway concept. The step-by-step guide for specifying a test-management pathway can be used to practice.

Initiatives that are being explored include incorporating the gained knowledge of this thesis project in the GRADE for Diagnosis course of the Dutch GRADE Network and developing add-on testing modules in the International Guideline Training and Certification Program INGUIDE. INGUIDE is a joint partnership of Guidelines International Network (GIN) and McMaster University's Department of Health Research Methods, Evidence, and Impact. Currently, add-on modules for certified guideline panel members, as well as for guideline methodologists, and eventually lead guideline developers and chairs, are being considered.

In addition, it is explored if the knowledge required for developing guideline recommendations on healthcare related testing, as well as the step-by-step guide to aid the specification of the test-management pathway and facilitate the formulation of focused questions, can be integrated in the ZonMw funded project 'Learning platform for guideline development: future-proof and sustainable'. This learning platform is a joint collaboration between the Care and Public Health Research Institute (CAPHRI) and the Maastricht School of Health Professions Education (SHE) at Maastricht University, the Academic Center of Epileptology Kempenhaeghe Maastricht UMC, and the Knowledge Institute of the Dutch Association of Medical Specialists.

Lastly, the results of this thesis will be incorporated into guideline training and coaching for various guideline developing organisations in the Netherlands.

The proposed initiatives are likely to succeed due to the networks and collaborations among all researchers involved in this thesis, both in the Netherlands and internationally.

References

1. Kuijpers T, Langendam MW, Tuut MK, De Beer JJA, Van der Wees P. Tool GRADE voor diagnostiek. 2018. Available from: https://nl.gradeworkinggroup.org/docs/Tool+GRADE+voor+diagnostiek+20180630.pdf.
2. Tuut MK, Langendam MW, De Beer JJA, Kuijpers T. GRADE-methodiek voor diagnostische tests en teststrategieën. 2015. Available from: https://nl.gradeworkinggroup.org/docs/GRADE-methodiek-voor-diagnostische-tests-en-teststrategieen-2015.pdf.
3. Adviesgroep Kwaliteitsstandaarden Zorginstituut Nederland. AQUA-Leidraad. Zorginstituut Nederland; 2021. Available from: https://www.zorginzicht.nl/binaries/content/assets/zorginzicht/ontwikkeltools-ontwikkelen/aqua-leidraad.pdf.

Summary

Summary

This thesis describes research in the field of guideline development, more specifically the development of guidelines for healthcare related testing, with the aim of facilitating and improving the process of developing guidelines recommendations about testing. This summary outlines the separate chapters of the thesis and highlights the conclusions.

Chapter 1 provides the general introduction to the thesis. It sets out the rationale for the thesis by introducing the topic and its components, emphasizing their importance and challenges, and defining the aim and research questions.

Guidelines, including clinical practice and public health guidelines, are documents that provide recommendations to enhance healthcare. The development of guidelines follows a clear process that includes systematic reviewing of available evidence and analysis of the benefits and harms of alternative care options, within a guideline panel of experts and representatives from key affected groups. Many organisations worldwide have adopted the GRADE approach, which emphasises the importance of certainty of evidence for clinically relevant differences in people-important outcomes. This approach pays specific attention to guideline development on healthcare related testing, taking into account the indirect link between testing and people-important outcomes, and emphasising the importance of consideration of false positive, false negative and inconclusive test results on people-important outcomes. Although the general competencies and knowledge required for guideline development are known, specific knowledge for creating testing guidelines has, to our knowledge, not yet been established.

The purpose of testing is to improve or prevent deterioration of people-important outcomes. People-important outcomes are components of people's (health) status following an intervention, and are used to assess effectiveness. Unlike treatment, testing usually does not have an immediate impact on people-important outcomes, although there are some exceptions. This implies that a series of steps, such as treatment, must be taken to move from testing to people-important outcomes. Testing in healthcare can serve various purposes, including screening, surveillance, risk classification, diagnosis, staging, treatment triage, prognosis, and follow-up. To assess the value of a test, various aspects should be considered. These include the analytic performance, clinical performance, clinical effectiveness, cost-effectiveness, and the broader impact of the test. Defining the role of a new test relative to existing tests, such as triage or add-on, is also critical.

In practice, both overuse and underuse of tests are common, and this can have a significant impact. For example, laboratory diagnostics accounts for approximately 2% of healthcare spending, yet it influences 64-67% of clinical decisions. Incorrect testing can result in high healthcare costs, unnecessary test burden, and anxiety.

Developing guidelines on healthcare related testing presents several challenges. These include formulating key questions that incorporate people-important outcomes, searching and synthesising evidence, interpreting test accuracy measures, and formulating recommendations. This thesis focuses on challenges and solutions in the development of guideline recommendations about healthcare related testing, with specific attention to the required knowledge for developing these recommendations and tools to facilitate this process. The aim of this thesis is to facilitate and improve guideline development concerning healthcare related testing. This has led to the following research questions:

1. What are challenges and possible solutions when assessing the certainty of evidence of a test-management pathway?
2. Which types of evidence (diagnostic accuracy, burden of the test, natural course, treatment effectiveness, link between test result and administration of treatment) are used to support guideline recommendations about testing?
3. What is the minimum knowledge required for guideline panel members involved in developing recommendations about testing?
4. Can a step-by-step guide aid guideline developers in formulating key questions about testing?

Chapter 2 addresses the first research question. This chapter analyses the added value of a test in an illustrative example. Specifically, it examines the net benefit of specific immunoglobulin E (sIgE) blood testing as an add-on test to history taking compared to history taking alone in patients suspected of having allergic rhinitis in primary care. The critical outcomes examined are relief of nasal or ocular symptoms, while the important outcomes include concentration, sleep problems, work/school absence, and quality of life. By using GRADE for diagnosis, we systematically assessed the available evidence on the elements of the test-management pathway, including test accuracy, test burden, management effectiveness, natural course, and the link between test results and management. Throughout this process, we identified challenges and proposed solutions to address them.

The lack of high certainty evidence for the various elements of the test-management pathway is a major challenge in interpreting the evidence and assessing the net benefit of a test. Another major challenge is the time required to systematically evaluate the

complete test-management pathway. To save time, consulting panel members, including patient representatives, may be a practical solution for selecting critical elements of the pathway for which a systematic review of the evidence should be undertaken. For less critical elements, the guideline panel may then refer to other guidelines, grey literature, professional expertise, and professional and consumer experience. The guideline panel can provide recommendations on the methodological approach for each element of the test-management pathway.

Chapter 3 addresses the second research question. This chapter evaluates the extent to which evidence-based guidelines on tests cover all elements of the test-management pathway. Specifically, it examines publicly accessible guidelines on three common tests: C-reactive protein (CRP) to estimate the likelihood of pneumonia, colonoscopy to detect colon cancer, and fractional exhaled nitric oxide (FeNO) to diagnose (severe) asthma in a systematic document analysis. Fifteen national and international guidelines published between 2016 and 2020 were analysed. The guidelines' methodological quality was evaluated using AGREE-II domain methodology, and it varied from poor to excellent.

Test accuracy was considered in the development of ten out of fifteen guideline recommendations, with four of them being based on a systematic review and rating of the certainty in the evidence. None of the guidelines included an evaluation of all steps of the test-treatment pathway. Three guidelines included consideration of test burden and two of natural course, but without a systematic review of the evidence. Of the three guideline recommendations that included consideration of management effectiveness, one based this on a systematic review and rating of the certainty in the evidence. The link between test results and management was not considered in any of the guidelines. Reporting issues and challenging methodology may explain the lack of transparent consideration of all elements of the test-management pathway.

Chapter 4 addresses the third research question. This is a developmental study, in which we determined the minimum knowledge required for guideline panel members involved in developing recommendations on healthcare related testing. We determined a draft set of knowledge components based on literature review. Subsequently, semi-structured interviews were conducted with nine internationally respected experts in testing in healthcare, test evaluation, guideline development including GRADE for tests, public involvement in guideline development, and training in guideline development on healthcare testing. The knowledge components were modified based on feedback from the interviewees and approved by all study participants.

The list of knowledge components required for guideline panel members to adequately develop recommendations on healthcare related testing consists of 26 items. These items cover the topics health question, test-management pathway, target population, test, test result, interpretation of test results & subsequent management, and impact on people-important outcomes. The required level of knowledge for each component is also defined. Understanding the test-management pathway concept appears to be the key knowledge component, linking all other essential knowledge components.

Chapter 5 provides four practical examples of test-management pathways for test scenarios in various settings, purposes, and roles. For each test-management pathway example concrete details are meticulously described, for educational purpose. The need for such examples became apparent during the interviews in chapter 4 and in academic presentations on this topic. The scenarios include various types of tests: self-testing, screening, diagnostic testing, and follow-up testing. These examples can be used by guideline methodologists, guideline panel chairs, and trainers to help guideline panel members understand and adopt the test-management pathway concept.

Chapter 6 addresses the fourth research question. In this developmental study, we created a step-by-step guide for guideline developers to specify a test-management pathway using a co-creative design. The draft guide underwent user testing in a workshop with nineteen healthcare professionals and researchers who have expertise and/or interest in guideline development. The adjusted step-by-step guide was subsequently user-tested in a before-after approach. Seven guideline panel members were asked to formulate a guideline question on testing, first without and subsequently with the use of the step-by-step guide.

The step-by-step guide for specifying a test-management pathway consists of five blocks with signalling questions, which emphasise people (including setting and timing), the index test, outcomes of interest, linking outcomes to testing, and comparator. The user can change the order of the steps and questions. Participants found the step-by-step guide helpful for structuring questions and defining the purpose and impact of the test of interest, and were intended to use the guide in a guideline panel setting. The guide should facilitate guideline developers in defining guideline questions on healthcare related testing by identifying relevant elements, which is an essential step in guideline development.

Chapter 7 provides an overview of the results presented in this thesis and a general discussion based on these findings, including a general reflection on methodological strengths and limitations. The thesis highlights the challenges of developing guideline

recommendations on healthcare related testing, including the frequent lack of evidence for critical elements of a test-management pathway, and the time required to adequately evaluate the evidence. The thesis highlights the significance of the test-management pathway concept in guideline development on healthcare related testing. This is crucial to understand for guideline panel members when developing guideline recommendations on healthcare related testing. The thesis also provides examples of test-management pathways and a step-by-step guide for specifying such pathways. These can help to understand the importance of the test-management pathway concept and facilitate the formulation of key questions about healthcare related tests. The research focuses on evaluating the evidence and facilitating guideline panel members in the guideline development process. It does not cover the process of moving from evidence to decision and the roles of guideline methodologists and guideline panel chairs.

In addition to the previous described results, the research has prompted reflections on the concept of test-management pathways. These include the use of more inclusive language over time, as well as a recurring debate regarding the definition of test burden. Furthermore, in published evidence, there is a great focus on diagnostic tests and dichotomous test results, whereas other purposes and test results are less discussed. It is acknowledged that test evaluation in guideline development occurs in a simplified version of reality. Guideline developers should be aware of these insights. Additionally, it is important to raise awareness about the potential downsides of testing, not only in scientific and guideline development environments, but also in the context of shared decision-making. Implementing the test-management pathway in healthcare policymaking could potentially reduce overtesting, overdiagnosis (including overdetection and overdefinition), and subsequent overtreatment. This involves evaluating the net benefit of testing on people-important outcomes in guideline development.

Recommendations for practice include emphasising the importance of the test-management pathway concept when updating guidance on guideline development, incorporating this concept into training of guideline panel members and methodologists, and creating an online tool to specify the test-management pathway by guideline panels. Recommendations for research include identifying the required knowledge for guideline methodologists and guideline panel chairs to develop recommendations on testing, evaluating the step-by-step guide for specifying the test-management pathway in guideline panel settings, and developing and testing educational strategies and tools to facilitate guideline development on healthcare related testing.

Samenvatting

Samenvatting

Dit proefschrift beschrijft onderzoek op het gebied van ontwikkeling van richtlijnen voor de gezondheidszorg, en dan specifiek richtlijnaanbevelingen over testen, met als doel om het proces van richtlijnontwikkeling over de inzet van testen te faciliteren en verbeteren. Deze samenvatting geeft een overzicht van de afzonderlijke hoofdstukken en de conclusies.

Hoofdstuk 1 bevat de algemene inleiding van dit proefschrift. Het beschrijft de rationale van het proefschrift, het definieert het onderwerp en belangrijke componenten hierin, inclusief de uitdagingen, en eindigt met het doel van het onderzoekt en de onderzoeksvragen.

Richtlijnen in de gezondheidszorg zijn documenten die aanbevelingen bevatten om de gezondheidszorg te verbeteren. De ontwikkeling van richtlijnen gaat volgens een vastomlijnd proces. Dat proces omvat een systematische beoordeling van het beschikbare bewijs en een analyse van de voor- en nadelen van de verschillende opties voor interventie in de praktijk, in dit geval testen. Dit wordt gedaan door richtlijnwerkgroepen met daarin vertegenwoordigers van de belangrijkste betrokken beroepsgroepen en patiënten-/consumentengroepen. De competenties en kennis die nodig zijn voor het ontwikkelen van richtlijnen in het algemeen zijn bekend.

Wereldwijd hanteren veel organisaties de GRADE aanpak binnen hun richtlijnontwikkeling. GRADE legt de nadruk op de zekerheid van bewijs voor relevante verschillen in belangrijke uitkomstmaten, ook wel patiëntrelevante uitkomstmaten genoemd. De GRADE aanpak besteedt specifiek aandacht aan richtlijnontwikkeling over testen, met aandacht voor de impact van terecht- en fout-positieve, terecht- en fout-negatieve en niet-conclusieve testresultaten op relevante uitkomsten. Desondanks is de specifieke kennis die nodig is voor het ontwikkelen van richtlijnen over testen niet eerder vastgesteld.

Het doel van testen is het verbeteren en/of voorkómen van verslechteren van relevante uitkomsten. Met relevante uitkomsten worden componenten van de gezondheid van mensen bedoeld, die worden gebruikt om de effectiviteit van interventies te beoordelen. In tegenstelling tot behandeling heeft testen doorgaans geen directe impact op deze relevante gezondheids-uitkomsten, alhoewel er enkele uitzonderingen zijn. Dit betekent dat een aantal stappen moet worden uitgevoerd, zoals behandeling, om van testen naar relevante uitkomsten te gaan. Testen in de gezondheidszorg kan meerdere doelen dienen, zoals screening, surveillance, risicostratificatie, diagnostiek, stadiëring, prognosebepaling en follow-up.

De inleiding beschrijft verschillende concepten die de waarde van een test bepalen: de analytische prestatie, de klinische prestatie (ook wel bekend als diagnostische accuratesse of testaccuratesse), de klinische effectiviteit, de kosteneffectiviteit en de brede impact van een test. Daarnaast is het van cruciaal belang om de rol van een nieuwe test (bijvoorbeeld triage of vervanging) ten opzichte van bestaande testen te benoemen.

In de praktijk komt verkeerd gebruik van testen frequent voor. Daarbij is zowel sprake van te weinig (*under use*) als te veel gebruik van testen (*over use*), met een aanzienlijke impact tot gevolg. Laboratoriumonderzoek bijvoorbeeld, dat ongeveer 2% van de uitgaven in de gezondheidszorg behelst, beïnvloedt 46-67% van de klinische besluitvorming. Verkeerd gebruik van testen kan leiden tot onnodige ongerustheid, onnodige bijwerkingen als gevolg van testen, onnodige behandeling en vermijdbare hoge kosten voor de gezondheidszorgen.

Het ontwikkelen van richtlijnen over testen in de gezondheidszorg brengt verschillende uitdagingen met zich mee, zoals het formuleren van uitgangsvragen gericht op verbeteren van relevante uitkomsten, het zoeken en evalueren van wetenschappelijk bewijs, het interpreteren van testaccuratesse en het formuleren van aanbevelingen. Dit proefschrift richt zich op knelpunten en oplossingen bij het ontwikkelen van richtlijnaanbevelingen over testen, met specifieke aandacht voor de kennis die nodig is om deze aanbevelingen te ontwikkelen en hulpmiddelen voor dit proces. Het doel van dit proefschrift is het faciliteren en verbeteren van richtlijnontwikkeling over testen in de gezondheidszorg. Dit heeft geleid tot de volgende onderzoeksvragen:

1. Wat zijn knelpunten en mogelijke oplossingen bij het beoordelen van de zekerheid van bewijs van een test-managementstrategie?
2. Welke typen bewijs (testaccuratesse, nadelige aspecten gerelateerd aan een test (*test burden*), natuurlijk beloop, effectiviteit van behandeling en link tussen testresultaat en behandeling) worden bij richtlijnontwikkeling gebruikt als bewijs voor aanbevelingen over testen?
3. Wat is de minimaal vereiste kennis voor richtlijnwerkgroepleden om richtlijnen te ontwikkelen over testen?
4. Kan een stap-voor-stap handleiding richtlijnontwikkelaars helpen bij het formuleren van gespecificeerde uitgangsvragen over testen?

Hoofdstuk 2 gaat in op de eerste onderzoeksvraag. In dit hoofdstuk is de toegevoegde waarde van een test geanalyseerd, als een illustratief voorbeeld. Het gaat om het onderzoeken van het voordeel van specifiek immunoglobuline E (sIgE) bloedtesten als aanvullende test op de anamnese bij patiënten die verdacht worden van allergische

rhinitis in de eerstelijns gezondheidszorg. Als cruciale uitkomstmaten zijn hierbij vermindering van oculaire en nasale symptomen geëvalueerd, terwijl concentratie, slaapproblemen, absentie van school of werk en kwaliteit van leven als belangrijke uitkomstmaten zijn beschouwd.

Door gebruik te maken van de GRADE aanpak voor testen hebben we het beschikbare bewijs voor alle elementen van de test-managementstrategie systematisch beoordeeld. Dit betrof de testaccuratesse, nadelige aspecten gerelateerd aan de test (*test burden*), effectiviteit van behandeling, natuurlijk beloop en de link tussen testresultaten en behandeling. Tijdens het beoordelingsproces hebben we knelpunten in het proces geïdentificeerd en oplossingen voorgesteld om deze uitdagingen aan te pakken.

Een grote uitdaging voor het beoordelen van de toegevoegde waarde van een test betreft de interpretatie van het bewijs. Bewijs voor de verschillende bouwstenen van de test-managementstrategie is vaak afwezig of van lage of zeer lage zekerheid. Een ander groot knelpunt betreft de tijd die gemoeid gaat met het systematisch evalueren van de gehele test-managementstrategie. Een praktische oplossing om tijd te besparen is het selecteren van de kritische elementen van de test-managementstrategie die naar verwachting van de richtlijnwerkgroep (inclusief patiëntvertegenwoordigers) de richting en sterkte van de aanbeveling bepalen. Hierop kan dan het systematisch literatuuronderzoek dat nodig is voor richtlijnontwikkeling worden gericht. Voor minder kritieke elementen kan de richtlijnwerkgroep eventueel gebruik maken van andere richtlijnen, grijze literatuur, professionele expertise en/of ervaring van professionals en zorgconsumenten.

Hoofdstuk 3 beschrijft in hoeverre in evidence-based richtlijnen over testen alle elementen van de test-managementstrategie zijn geëvalueerd (de tweede onderzoeksvraag). Daarbij zijn in een systematische documentanalyse openbaar toegankelijke richtlijnen over drie veel voorkomende testen onderzocht, namelijk C-reactief proteïne (CRP) om het risico op een pneumonie in te schatten, colonoscopie om een coloncarcinoom op te sporen en bepaling van de fractie stikstofoxide in de uitgeademde lucht (FeNO) om (ernstig) astma te diagnostiseren.

In totaal zijn vijftien nationale en internationale richtlijnen, gepubliceerd tussen 2016 en 2020, geanalyseerd. De methodologische kwaliteit van de richtlijnen werd geëvalueerd met behulp van het domein methodologie van het AGREE-II instrument. Deze kwaliteit varieerde van slecht tot uitstekend.

Bij de ontwikkeling van tien van de vijftien bestudeerde richtlijnaanbevelingen werd de testaccuratesse overwogen, waarbij dat in vier van deze tien gebaseerd was op systematisch literatuuronderzoek met beoordeling van de zekerheid van bewijs. In geen enkele van de geïncludeerde richtlijnen werd de volledige test-managementstrategie geëvalueerd. In drie richtlijnen werden de nadelige aspecten gerelateerd aan testen (*test burden*) overwogen en in twee richtlijnen het natuurlijk beloop, maar in alle gevallen was dat zonder een systematische beoordeling van het bewijs. Van de drie richtlijnaanbevelingen die de effectiviteit van de behandeling beschouwden, was dit in slechts één aanbeveling gebaseerd op een systematische review van de literatuur en beoordeling van de zekerheid van bewijs. De link tussen testresultaten en behandeling werd in geen van de geïncludeerde richtlijnen overwogen. Het gebrek aan transparante overweging van alle elementen van de test-managementstrategie wordt mogelijk verklaard door gebrek aan transparante rapportage of door complexe ontwikkelmethodologie.

Hoofdstuk 4 beschrijft een ontwikkelstudie, waarin we hebben vastgesteld welke kennis minimaal vereist is voor richtlijnwerkgroepleden die betrokken zijn bij de ontwikkeling van aanbevelingen over testen (de derde onderzoeksvraag). Hierbij hebben we eerst op basis van literatuuronderzoek een voorlopige set met kenniscomponenten beschreven. Daarna hebben we semigestructureerde interviews gehouden met negen internationaal gerespecteerde experts. Het betrof experts op het gebied van testen in de gezondheidszorg, wetenschappelijke testevaluatie, richtlijnontwikkeling inclusief GRADE voor testen, publieke betrokkenheid bij richtlijnontwikkeling en/of training in richtlijnontwikkeling over testen. De kenniscomponenten zijn aangepast op basis van feedback van de geïnterviewden en goedgekeurd door alle deelnemers aan het onderzoek.

De lijst met kenniscomponenten beschrijft 26 items die vereist zijn voor richtlijnwerkgroepleden om adequaat aanbevelingen te kunnen ontwikkelen over testen. Deze items hebben betrekking op de thema's uitgangsvraag, test-managementstrategie, doelpopulatie, test, testresultaat, interpretatie van testresultaat & daaropvolgend management en impact op relevante uitkomsten. Voor elk item is het vereiste kennisniveau vastgesteld. Het begrijpen van het concept test-managementstrategie is de belangrijkste kenniscomponent, die alle andere essentiële kenniscomponenten met elkaar verbindt.

Hoofdstuk 5 beschrijft vier uitgewerkte test-managementstrategieën voor testscenario's met verschillende doelen en rollen en in verschillende settings als praktische en educatieve voorbeelden die aansluiten bij het onderzoek naar benodigde

kennis (de derde onderzoeksvraag). Ter educatie hebben we in elk voorbeeld concrete details nauwkeurig beschreven. De behoefte aan dergelijke uitgewerkte voorbeelden bleek tijdens de interviews die gehouden werden met de experts (zie hoofdstuk vier) en bij wetenschappelijke presentaties over dit onderwerp. De volgende testscenario's zijn uitgewerkt: zelftesten, screening, diagnostische testen en follow-up testen. Deze voorbeelden kunnen gebruikt worden door richtlijnmethodologen, voorzitters van richtlijnwerkgroepen en richtlijntrainers om het begrip en de toepassing van het concept test-managementstrategie te faciliteren.

Hoofdstuk 6 beschrijft een stap-voor-stap handleiding voor richtlijnontwikkelaars om een test-managementstrategie te specificeren ontwikkeld met behulp van co-creatie (de vierde onderzoeksvraag). De concept handleiding is getest onder gebruikers in een workshop met negentien zorgprofessionals en onderzoekers met expertise en/of interesse in richtlijnontwikkeling. De aangepaste handleiding hebben we vervolgens getest onder richtlijnwerkgroepleden in een voor-na design. Daarbij hebben we zeven richtlijnwerkgroepleden gevraagd om een uitgangsvraag over testen te formuleren, eerst zónder en daarna mét gebruik van de stap-voor-stap handleiding.

De stap-voor-stap handleiding voor het specificeren van een test-managementstrategie bestaat uit vijf blokken met vragen, gericht op populatie (inclusief zorgsetting en timing van de beoogde test), de indextest, relevante uitkomsten, de link tussen testen en uitkomsten, en de controletest. De gebruiker van de handleiding kan de volgorde van de stappen en vragen naar eigen inzicht aanpassen. Deelnemers vonden de stap-voor-stap handleiding nuttig voor het structuren van de uitgangsvragen en het definiëren van het doel en de impact van de test. Zij gaven aan de handleiding te willen gebruiken in een richtlijnwerkgroep. De stap-voor-stap handleiding kan richtlijnontwikkelaars helpen bij het definiëren van uitgangsvragen over testen door het identificeren van relevante elementen. Dit is een essentiële stap in richtlijnontwikkeling.

Hoofdstuk 7 bevat een overzicht van de resultaten van dit proefschrift en de algemene discussie van de resultaten, inclusief sterke en zwakke punten in de onderzoeksaanpak. Het proefschrift benadrukt de uitdagingen bij het ontwikkelen van richtlijnaanbevelingen over testen, inclusief het gebrek aan bewijs voor cruciale elementen van een test-managementstrategie en de tijd die nodig is om de evidence adequaat te kunnen beoordelen. Begrip van het concept test-managementstrategie bij het ontwikkelen van richtlijnen over testen is cruciaal voor richtlijnwerkgroepleden. Daarnaast voorziet dit proefschrift in uitgewerkte voorbeelden van test-managementstrategieën en een stap-voor-stap handleiding om test-

managementstrategieën te specificeren. Deze kunnen helpen om het belang van het concept test-managementstrategie te begrijpen en uitgangsvragen over testen te formuleren.

Naast de concrete resultaten van de verschillende onderzoeken, heeft het onderzoek ook geleid tot kritische reflecties op het concept test-managementstrategieën. Deze omvatten het gebruik van meer inclusieve taal en de definitie van nadelige aspecten gerelateerd aan testen (*test burden*). Daarnaast viel op dat de bestaande literatuur veel aandacht besteedt aan diagnostische testen en dichotome testresultaten, terwijl andere doelen van testen (bijvoorbeeld follow-up of stadiëring) en andersoortige testresultaten (bijvoorbeeld continue) minder worden belicht. Ook wordt erkend dat testevaluatie in het kader van richtlijnontwikkeling gebeurt in een versimpelde versie van de werkelijkheid. Richtlijnontwikkelaars moeten zich hiervan bewust zijn. Daarnaast is het van belang om het bewustzijn van potentiële nadelen van testen te vergroten, niet alleen in de wetenschap en bij richtlijnontwikkeling, maar ook in de context van samen beslissen. Het implementeren van het concept van de test-managementstrategie in beleidsontwikkeling in de gezondheidszorg kan mogelijk overmatig gebruik van testen, overdiagnostiek (zowel overmatige detectie als herdefinitie van aandoeningen) en daaropvolgende overbehandeling beperken. Dit heeft uiteraard ook betrekking op het evalueren van het netto voordeel van testen op relevante uitkomsten in het kader van richtlijnontwikkeling.

Aanbevelingen voor de praktijk betreffen onder andere het benadrukken van het belang van het concept test-managementstrategie bij het actualiseren van handleidingen voor richtlijnontwikkeling, het meenemen van dit concept in training van richtlijnwerkgroepleden en andere richtlijnontwikkelaars en het ontwikkelen van een online tool waarmee richtlijnwerkgroepleden de test-managementstrategie eenvoudig kunnen specificeren. Aanbevelingen voor nader onderzoek zijn onder andere het identificeren van de benodigde kennis voor richtlijnmethodologen en richtlijnwerkgroepvoorzitters om aanbevelingen over testen te ontwikkelen, het evalueren van de stap-voor-stap handleiding voor het specificeren van de test-managementstrategie in richtlijnwerkgroepen en het ontwikkelen en testen van educatieve strategieën en hulpmiddelen om richtlijnontwikkeling over testen te faciliteren.

Publiekssamenvatting

Publiekssamenvatting

Dit proefschrift gaat over de ontwikkeling van richtlijnen over testen in de gezondheidszorg.

Richtlijnen helpen zorgverleners (zoals dokters, verpleegkundigen en fysiotherapeuten) en zorgontvangers (patiënten/cliënten/burgers) bij keuzes in de zorg. Aanbevelingen in richtlijnen worden gebaseerd op wetenschappelijke kennis en op expertise en ervaringen van zorgprofessionals en ervaring van zorgontvangers. Richtlijnen worden ontwikkeld om de gezondheidszorg te verbeteren; zorggebruikers moeten er dus beter van worden.

Met **'testen'** worden procedures bedoeld waarmee een gezondheids- of ziektetoestand bij mensen kan worden gemeten. Voorbeelden zijn: bloedonderzoek, vragenlijstonderzoek, beeldvormend onderzoek, functietesten (bijvoorbeeld een ECG of een longfunctietest) en weefselonderzoek. Zulke testen kunnen om verschillende redenen worden gedaan, zoals het stellen of juist uitsluiten van een diagnose, screening of vroege opsporing, keuze van behandeling of controle.

Testen zijn meestal niet 100% accuraat. Een voorbeeld: een test kan als uitslag geven dat iemand een bepaalde ziekte heeft, terwijl dat in werkelijkheid niet zo is. Iemand kan dan behandeld worden voor een ziekte die diegene niet heeft. Andersom kan ook: een test kan als uitslag geven dat iemand een bepaalde ziekte niet heeft, terwijl dat in werkelijkheid wél zo is. Iemand kan dan geen behandeling krijgen terwijl die wel nodig is.

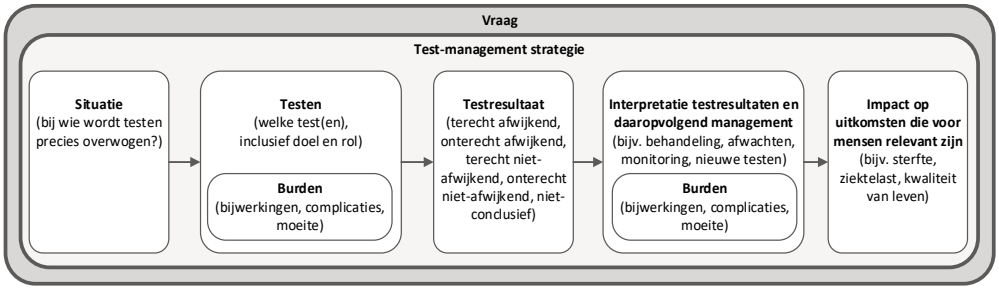
Bij de ontwikkeling van richtlijnen over testen wordt vaak gekeken naar **testaccuratesse**, zoals het aantal terechte en onterechte testresultaten. Maar, als richtlijnen over testen ervoor moeten zorgen dat zorggebruikers er beter van worden, dan moet ook gekeken worden naar de voor- en nadelen van de test en naar de consequenties voor het beleid of de behandeling. Immers, van alleen testen worden mensen niet beter.

Om te bepalen of een bepaalde test in een bepaalde situatie van toegevoegde waarde is, zijn de volgende overwegingen van belang:

- Wat is de **situatie**? Hiermee worden bijvoorbeeld de doelpopulatie van de test (bij welke mensen wordt de test overwogen?), het doel van de test (bijvoorbeeld diagnostiek of screening) en de rol van de test ten opzichte van bestaande testen (bijvoorbeeld vervanging van een bestaande test) bedoeld.
- Meet de test wat deze moet meten? Hiermee wordt de **accuratesse** bedoeld.

- Wat zijn de negatieve aspecten gerelateerd aan de test? Hiermee worden bedoeld: de **belasting** voor diegene die de test moet ondergaan en mogelijke **bijwerkingen** en **complicaties** van de test (bijvoorbeeld een kijkonderzoek van de dikke darm met voorbereidend laxeren en het risico op darmperforatie).
- Wat zijn de **consequenties** van de testresultaten? Hiermee wordt bijvoorbeeld geruststelling, behandeling (inclusief bijvoorbeeld bijwerkingen en complicaties van behandeling) en monitoring bedoeld.
- Wat zijn de belangrijke uitkomsten voor degenen die de test ondergaan? Dit worden **relevante uitkomstmaten** genoemd. Hiermee worden uitkomsten bedoeld waarvan zorggebruikers vinden dat ze beter worden, bijvoorbeeld minder kans op sterfte, minder ziektelast of betere kwaliteit van leven.

Al deze overwegingen hebben invloed op de vraag of testen in een bepaalde situatie zinvol is. Dit wordt schematisch weergegeven in een zogenoemde **test-managementstrategie** *(Figuur 1)*.

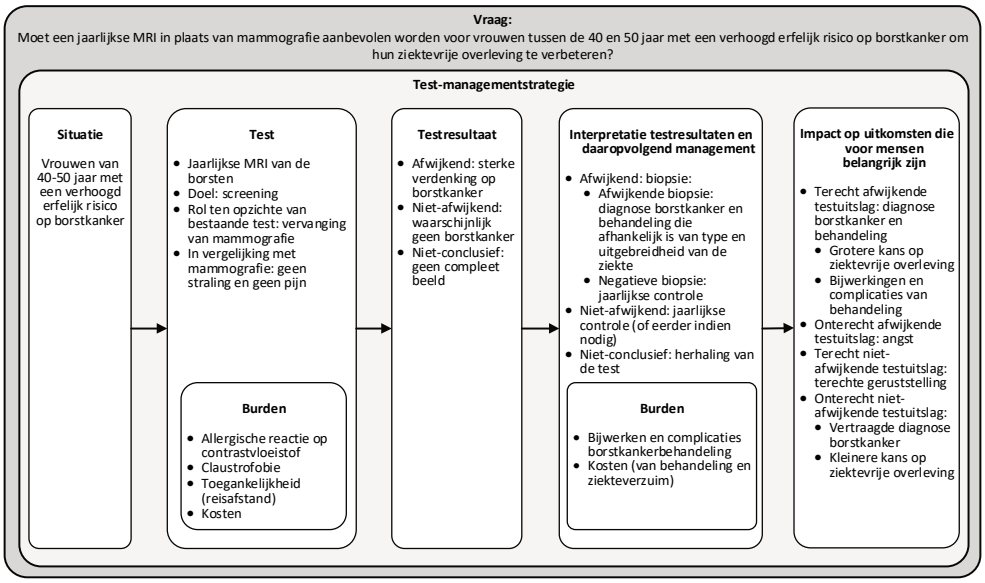


Figuur 1. Schematische weergave van de test-managementstrategie

Er is vaak weinig wetenschappelijk bewijs dat direct antwoord geeft op de vraag of testen in een bepaalde situatie bijdraagt aan relevante uitkomsten. Daarom moeten voor de ontwikkeling van een richtlijnen meestal alle stappen apart geanalyseerd worden. Dat kost veel tijd.

Bovendien is specifieke kennis nodig voor het ontwikkelen van richtlijnen over testen. We hebben een handleiding ontwikkeld die daarbij kan helpen. En om een test-managementstrategie goed te kunnen begrijpen, hebben we een aantal voorbeelden uitgewerkt. Een van die voorbeelden is weergegeven in *figuur 2*.

De belangrijkste boodschap van dit proefschrift is dat richtlijnontwikkelaars het belang van de test-managementstrategie inzien. Een goede testaccuratesse alleen is niet genoeg; het gaat om de impact van testen op relevante uitkomsten. Alleen dan kunnen mensen beter worden van testen.



Figuur 2. Voorbeeld uitgewerkte test-managementstrategie

Bibliography

Publications in scientific journals

- Tuut MK, Cals JWL, Jansen J, Burgers JS. Developing guideline recommendations about tests: educational examples of test-management pathways. Evid Based Med [submitted].
- Tuut MK, Gopalakrishna G, Leeflang MM, Patrick PM, Van der Weijden T, Burgers JS, Langendam MW. Co-creation of a step-by-step guide for specifying the test-management pathway to formulate focused guideline questions about healthcare related tests. BMC Med Res Methodol [submitted].
- Van Glansbeek-Schutijser B, Alizadeh M, Tuut M, Pelle T, De Hoop I, Van der Putten G-J, et al. Oral health, motivation and resistance to daily oral care: recommended approaches for (informal) caregivers of care dependent people - a systematic review. J Geront Nurs [submitted].
- Tuut MK, Burgers JS, De Beer JJA, Bindels PJE, Bossuyt PMM, Cals JW, et al. Required knowledge for guideline panel members to develop healthcare related testing recommendations - a developmental study. J Clin Epidemiol. 2024;173 doi:10.1016/j.jclinepi.2024.111438.Tuut MK, Burgers JS, Van der Weijden T, Langendam MW. Do clinical practice guidelines consider evidence about diagnostic test consequences on patient-relevant outcomes? A critical document analysis. J Eval Clin Pract. 2022;28(2):278-87. doi:10.1111/jep.13619.
- Vissink A, Tuut MK, Benedictus J, Hoeksema AR, Janssen MJEJ, Parunovac M, et al. Medicatiegeïnduceerde hyposialie en xerostomie. QP Tandheelkunde. 2021;16(6):28-32.
- Tuut MK, De Beer JJA, Burgers JS, Van de Griendt EJ, Van der Weijden T, Langendam MW. Applying GRADE for diagnosis revealed methodological challenges: an illustrative example for guideline developers. J Clin Epidemiol. 2021;131:123-32. doi:10.1016/j.jclinepi.2020.11.021.
- Hemmelder MH, van Balen J, Scherpbier N, Schenk PW, Tuut MK, Gansevoort RT. [Chronic kidney damage guidelines revision]. Ned Tijdschr Geneeskd. 2018;162
- Van de Griendt EJ, Tuut MK, De Groot H, Brand PLP. Applicability of evidence from previous systematic reviews on immunotherapy in current practice of childhood asthma treatment: a GRADE (Grading of Recommendations Assessment, Development and Evaluation) systematic review. BMJ Open. 2017;7(12):e016326. doi:10.1136/bmjopen-2017-016326.
- Van de Griendt EJ, Tuut MK, Bindels PJ. Nieuwe richtlijnen voor de diagnostiek en behandeling van astma bij kinderen in de tweede lijn. Ned Tijdschr Allergie & Astma. 2015;15:15-21.
- Smeele IJ, Barnhoorn MJM, Broekhuizen BD, Chavannes NH, In 't Veen JCCM, Van der Molen T, et al. NHG-Standaard Astma bij volwassenen (derde herziening). Huisarts Wet. 2015;58(3):142-54.
- Geijer RM, Tuut MK, in't Veen JC, Broekhuizen BD, Chavannes NH, Smeele IJ. [The NHG guidelines 'Adult asthma' and 'COPD']. Ned Tijdschr Geneeskd. 2015;159:A9076.
- Bindels PJE, Van de Griendt EJ, Grol MH, Hensbergen W, Steenkamer TA, Uijen JH, et al. NHG-Standaard Astma bij kinderen (derde herziening). Huisarts Wet. 2014;57:70-80.
- Bindels PJ, Van de Griendt EJ, Tuut MK, Steenkamer TA, Uijen JH, Geijer RM. [Dutch College of General Practitioners' practice guideline 'Asthma in children']. Ned Tijdschr Geneeskd. 2014;158:A7935.
- Oudejans JJ, Tuut MK. Richtlijn 'Primaire tumor onbekend'. Ned Tijdschr Oncol. 2013;10:117-21.
- Hoorn EJ, Tuut MK, Hoorntje SJ, van Saase JL, Zietse R, Geers AB. Dutch guideline for the management of electrolyte disorders--2012 revision. Neth J Med. 2013;71(3):153-65.
- Thio HB, Balak DM, Meilof JF, Stegeman CA, Voskuyl AE, Van Everdingen JJ, Tuut MK. [Guideline 'Diagnostics of small-vessel vasculitis']. Ned Tijdschr Geneeskd. 2012;156(21):A4317.
- Borgonjen RJ, Van Everdingen JJ, Bik CM, Tuut MK, Spuls PI, Van de Kerkhof PC. Prospective comparison of three guideline development methods for treatment of actinic keratosis. BMJ Qual Saf. 2011;20(10):832-41. doi:10.1136/bmjqs.2010.050443.
- Van Everdingen JJ, Dreesens DH, Tuut MK. [Advice on guidelines: planning for guideline development in the Netherlands]. Ned Tijdschr Geneeskd. 2010;154:A1599.
- Zonderland HM, Tuut MK, den Heeten GJ, Asperen CJ, de Bock GH, Rutqers EJ, et al. [Revised practice guideline 'Screening and diagnosis of breast cancer']. Ned Tijdschr Geneeskd. 2008;152(43):2336-9.
- Struikmans H, Nortier JW, Rutgers EJ, Zonderland HM, Bontenbal M, Elkhuizen PH, et al. [Guideline 'Treatment of breast cancer 2008' (revision)]. Ned Tijdschr Geneeskd. 2008;152(46):2507-11.
- Dekhuijzen PN, Broeders ME, Tuut MK, Grol MH. [Practice guideline 'Medical treatment of COPD']. Ned Tijdschr Geneeskd. 2008;152(26):1465-8.
- Werkgroep antiretrovirale behandeling van de Nederlandse Vereniging van Aids Behandelaren. [Revised guideline "Antiretroviral Treatment"]. Ned Tijdschr Geneeskd. 2005;149(43):2399-405.

- Van Tulder MW, Tuut M, Pennick V, Bombardier C, Assendelft WJ. Quality of primary care guidelines for acute low back pain. Spine (Phila Pa 1976). 2004;29(17):E357-62. doi:10.1097/01.brs.0000137056.64166.51.
- Spuls PI, Tuut MK, van Everdingen JJ, de Rie MA, Werkgroep Psoriasis van de Nederlandse Vereniging voor Dermatologie en V. [The practice guideline 'Photo(chemo)therapy and systemic therapy in severe chronic plaque-psoriasis']. Ned Tijdschr Geneeskd. 2004;148(43):2121-5.
- Koes BW, Sanders RJ, Tuut MK, Kwaliteitsinstituut voor de Gezondheidszorg CBO. [The Dutch Institute for Health Care Improvement (CBO) guideline for the diagnosis and treatment of aspecific acute and chronic low back complaints]. Ned Tijdschr Geneeskd. 2004;148(7):310-4.
- Van Tienhoven G, Tuut MK, Nortier JW, Rutgers EJ. [Dutch Institute for Healthcare Improvement guideline, "Treatment of breast cancer": an important document, but keep doors to new 'evidence' open]. Ned Tijdschr Geneeskd. 2003;147(8):362, author reply 3.
- Rutgers EJ, Nortier JW, Tuut MK, van Tienhoven G, Struikmans H, Bontenbal M, et al. [Dutch Institute for Healthcare Improvement guideline, "Treatment of breast cancer"]. Ned Tijdschr Geneeskd. 2002;146(45):2144-51.
- Tuut M, Hense HW. Smoking, other risk factors and fibrinogen levels. evidence of effect modification. Ann Epidemiol. 2001;11(4):232-8. doi:10.1016/s1047-2797(00)00226-x.
- Rutgers EJ, Tuut MK. [CBO guideline 'Breast cancer: screening and diagnosis']. Ned Tijdschr Geneeskd. 2001;145(3):115-9.
- Grobbee DE, Tuut MK, Hoes AW. [CBO guideline 'High blood pressure' (revision)]. Ned Tijdschr Geneeskd. 2001;145(43):2071-6.
- Tuut MK. Medische richtlijnen van het Kwaliteitsintituut voor de Gezondheidszorg CBO. Ned Tijdschr Med Microbiol. 2000;8(4):121-3.
- Stam J, Koudstaal PJ, Franke CL, Kappelle LJ, Boiten J, Tuut MK. [CBO guideline 'Stroke' (revision)]. Ned Tijdschr Geneeskd. 2000;144(34):1653-4.
- Limburg M, Tuut MK. [CBO guideline 'Stroke' (revision) Dutch Institute for Healthcare Improvement]. Ned Tijdschr Geneeskd. 2000;144(22):1058-62.

Conference contributions

- Van der Weijden T, Langendam MW, Burgers JS, Tuut MK. Training guideline panel members involved in developing recommendations about healthcare tests. [Workshop]. Guidelines International Network; Glasgow, 2023. https://g-i-n.net/wp-content/uploads/2023/10/Abstract-Book-Final.pdf.
- Tuut MK, Langendam M, Burgers JS, Van der Weijden T. Required knowledge to develop guideline recommendations about healthcare related tests. [Oral presentation]. Guidelines International Network; Glasgow, 2023. https://g-i-n.net/wp-content/uploads/2023/10/Abstract-Book-Final.pdf.
- Tuut MK, Burgers JS, Van der Weijden T, Langendam MW. Clinical practice guidelines about diagnostic tests hardly consider evidence about test consequences. [Oral presentation]. WEON; Online, 2021.
- Tuut MK, Langendam MW, Burgers JS, Van der Weijden T. Do diagnostic recommendations in clinical practice guidelines consider evidence about test consequences? [Oral presentation by Langendam MW]. Guidelines International Network; Adelaide, 2019. https://journals.lww.com/ijebh/Fulltext/2019/12000/Abstracts_for_the_GIN_and_JBI_Conference,_Adelaide.2.aspx.
- Tuut MK, Burgers JS, Van der Weijden T, Langendam MW. Diagnostic recommendations in clinical practice guidelins (CPGs): based on quicksand? [Poster presentation]. CAPHRI Research Day; Valkenburg, 2019.
- Tuut MK, De Beer JJA, Burgers JS, Van de Griendt EJ, Sijbom M, Van der Weijden T, Langendam MW. Application of GRADE for test-treatment strategies: challenges and possible solutions. [Poster]. Guidelines International Network; Manchester, 2018.
- Tuut MK, Van de Griendt EJ, Brand PLP. Systematic review of effectiveness of immunotherapy for pediatric asthma: the importance of applicability of research evidence. [Oral presentation by Brand PLP]. Evidence Live; Oxford, 2016.
- Tuut MK, Van Everdingen JJ, Borgonjen RJ, Oudejans JJ. Evidence-based guideline development using pressure-cook sessions: study and experience. [Oral presentation]. Guidelines International Network; Berlin, 2012. doi:10.3205/12gin072.

- Burgers JS, Kersten S, Rosenbrand K, Van Barneveld TA, Franx G, Hulshof C, et al. Stories and lessons of 30 years guideline development in The Netherlands. [Oral presentation by Burgers, J.S.]. Guidelines International Network; 2012, 2012. doi:10.3205/12gin084.
- Tuut MK, Van Croonenborg J, Rutgers EJ, Struikmans H, Nortier JW. Experiences with “living guidelines” in the Netherlands. [Poster]. Guidelines International Network; Lyon, 2005. https://g-i-n.net/wp-content/uploads/2022/01/GIN-2005-Abstract-Book.pdf.
- Van Everdingen JJ, Liberati A, Rosier P, Tuut MK. Evidence based guidelines as a toolbox for clinical improvement. [Oral presentation by Liberati, A.]. 6th European Forum on Quality improvement in Health Care; Bologna, 2001.
- Tuut MK, Spuls PI. The Dutch guideline on the teratment of psoriasis: a combination of the analysis of the literature, practical aspects and patients' opinions. [Oral presentation by Spuls PI]. 9th International Cochrane Colloquium; Lyon, 2001.
- Wittenberg J, Tuut MK, Assendelft WJ. Evaluating the quality of scientific medical literature. [Oral presentation by Wittenberg J]. European Forum on Quality Improvement in Health Care; Amsterdam, 2000.
- Wittenberg J, Tuut M, Van Everdingen JJ. Transparency of the guideline development process by using a new guideline structure. [Poster]. Cochrane Colloquium; Cape Town, 2000.
- Tuut MK, Limburg M. The national guideline on stroke 1999: evaluation of new methods. [Oral presentation]. European Forum on Quality Improvement in Health Care; Amsterdam, 2000.
- Tuut MK, Assendelft WJ, Sanders RJ, Van Everdingen JJ. Existing guidelines as a basis for a multidisciplinary guideline on nonspecific low back pain. [Oral presentation]. The Fourth International Forum for Primary Care Research on Low Back Pain; Eilat, 2000.
- Tuut M, Wittenberg J, Rutjes A, Assendelft WJ. Instruments used for evaluation of medical literature. [Poster]. Cochrane Colloquium; Cape Town, 2000.

Book contributions and methodological reports

- Fleuren M, Van Enst A, Tuut M. Evaluatie en monitoring kwaliteitsstandaarden verpleegkundigen en verzorgenden. Utrecht: Verpleegkundigen & Verzorgenden Nederland; 2022. https://www.venvn.nl/media/e4dbko4v/evaluatie-en-monitoring-richtlijnen.pdf.
- Nederlands Huisartsen Genootschap, Nederlandse Vereniging voor Medische Microbiologie, Nederlandse Vereniging voor Klinische Chemie en Laboratoriumgeneeskunde, SAN Centra voor Medische Diagnostiek, Federatie Medisch Coördinerende Centra. Handleiding voor de opzet en uitvoering van het Diagnostisch Toetsoverleg (DTO) in de huisartsenzorg. 2018. https://www.nhg.org/wp-content/uploads/2022/11/diagnostisch_toetsoverleg_dto_2018_web.pdf.
- Kuijpers T, Langendam MW, Tuut MK, De Beer JJA, Van der Wees P. Tool GRADE voor diagnostiek. Dutch GRADE Network; 2018. https://nl.gradeworkinggroup.org/docs/Tool+GRADE+voor+diagnostiek+20180630.pdf.
- Tuut MK, Langendam MW, De Beer JJA, Kuijpers T. GRADE-methodiek voor diagnostische tests en teststrategieën. Dutch GRADE Network; 2015. https://nl.gradeworkinggroup.org/docs/GRADE-methodiek-voor-diagnostische-tests-en-teststrategieen-2015.pdf.
- AQUA. AQUA-Leidraad / Leidraad voor Kwaliteitsstandaarden. Diemen; 2014-2021. https://www.zorginzicht.nl/binaries/content/assets/zorginzicht/ontwikkeltools-ontwikkelen/aqua-leidraad.pdf.
- Kremer LCM, Burgers JS, Tuut MK. Criteria voor goede richtlijnen. In: Van Everdingen JJ, Dreesens DH, Burgers JS, Swinkels JA, Van Barneveld TA, Van der Weijden T, editors. Handboek evidence-based richtlijnontwikkeling Een leidraad voor de praktijk. 2. Houten: Bohn Stafleu van Loghum; 2014 isbn:
- Tuut MK, De Beer JJA, Van Everdingen JJ, Van den Eerenbeemt A. (eds.). Pinkhof Zakwoordenboek 'evidence-based' geneeskunde. Houten: Bohn Stafleu van Loghum; 2011. isbn:978-90-313-8804-2.
- Scholten RPJM, Tuut MK, Kremer LCM, Assendelft WJ. Beoordelen van de kwaliteit van medisch-wetenschappelijk onderzoek. In: Everdingen JJE, Burgers JS, Assendelft WJ, Swinkels JA, van Barneveld TA, Van de Klundert JLM, editors. Evidence-based richtlijnontwikkeling Een leidraad voor de praktijk. 1. Houten: Bohn Stafleu van Loghum; 2004 isbn: 9789031342099.

Guidelines (and other publications to improve the quality of care in practice)

- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn spirometrie. SAN Centra voor Medische Diagnostiek; 2023. https://de-san.nl/Praktijkrichtlijnen/.
- Verpleegkundigen & Verzorgenden Nederland. Richtlijn Signalering en preventie van zorginfecties. Utrecht: Verpleegkundigen & Verzorgenden Nederland; 2022. https://www.venvn.nl/media/25gjm1ee/v-vn-richtlijn-zorginfecties-versie-20-januari-2022.pdf.
- Verpleegkundigen & Verzorgenden Nederland. Richtlijn Decubitus. Utrecht: Verpleegkundigen & Verzorgenden Nederland; 2021. https://www.venvn.nl/media/adujx1ja/20210224-richtlijn-decubitus.pdf.
- Schellevis FG, Burgers JS, Tuut MK, Kleinnibbelink G. Generieke module Comorbiditeit. Utrecht: Nederlands Huisartsen Genootschap; 2021. https://www.ggzstandaarden.nl/uploads/pdf/project/project_7d44b913-5e59-4ed9-90cb-4c122345f2d5_comorbiditeit__authorized-at_20-06-2018.pdf.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn Addendum uitvoering longfunctieonderzoeken tijdens COVID-19. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn audiometrie. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn echografie. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn fundusfotografie. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn Röntgenonderzoek. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn 30-minuten bloeddrukmeting. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn 24-uursbloeddrukmeting. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn prostaatechografie. SAN Centra voor Medische Diagnostiek; 2021. https://de-san.nl/Praktijkrichtlijnen/.
- Nederlandse Vereniging voor Kindergeneeskunde, Nederlands Huisartsen Genootschap, Verpleegkundigen & Verzorgenden Nederland, Longfonds. Richtlijn astma bij kinderen. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2021. https://richtlijnendatabase.nl/richtlijn/astma_bij_kinderen/startpagina_-_astma_bij_kinderen.html.
- Kennisinstituut Mondzorg. Richtlijn Mondzorg voor jeugdigen - preventie en behandeling. Utrecht: Kennisinstituut Mondzorg; 2021. https://www.hetkimo.nl/wp-content/uploads/2021/02/2020.12.31-KPR-MvJ-preventie-en-behandeling-caries-DEF.pdf.
- Kennisinstituut Mondzorg. Richtlijn Xerostomie en hyposialie gerelateerd aan medicatie en polyfarmacie. Utrecht: Kennisinstituut Mondzorg; 2021. https://www.hetkimo.nl/richtlijnen/xerostomie-en-hyposialie-gerelateerd-aan-medicatie-en-polyfarmacie/introductie/.
- Commissie Leidraad Mondzorg Corona. Leidraad Mondzorg Corona. 2020-2023.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn OSAS. SAN Centra voor Medische Diagnostiek; 2020. https://de-san.nl/Praktijkrichtlijnen/.
- Nederlandse Vereniging voor Kindergeneeskunde, Nederlandse Vereniging voor Obstetrie en Gynaecologie, Koninklijke Nederlandse Organisatie van Verloskundigen, Stichting Kind en Ziekenhuis. Richtlijn Postnatale zorg in de algemene kindergeneeskunde. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2020. https://richtlijnendatabase.nl/richtlijn/postnatale_zorg_in_de_algemene_kindergeneeskunde/startpagina_-_postnatale_zorg_in_de_algemene_kindergeneeskunde.html.
- Nederlandse Federatie voor Nefrologie, Nederlandse Internisten Vereniging, Nederlandse Vereniging voor Heelkunde, Verpleegkundigen & Verzorgenden Nederland, Nierpatiënten Vereniging Nederland, Landelijk Overleg Nier Transplantatie, et al. Richtlijn Zorg bij eindstadium nierfalen. Utrecht: Nederlandse Internisten Vereniging; 2020.

https://richtlijnendatabase.nl/richtlijn/zorg_bij_eindstadium_nierfalen/startpagina_-_zorg_bij_eindstadium_nierfalen.html.
- Kennisinstituut Mondzorg. Richtlijn Derde molaar. Utrecht: Kennisinstituut Mondzorg; 2020. https://www.hetkimo.nl/richtlijnen/derde-molaar/introductie/.
- Kennisinstituut Mondzorg. Richtlijn Antitrombotica. Utrecht: Kennisinstituut Mondzorg; 2019. https://www.hetkimo.nl/richtlijnen/antitrombotica/introductie/.
- Kennisinstituut Mondzorg. Richtlijn wortelcariës bij ouderen. Utrecht: Kennisinstituut Mondzorg; 2019. https://www.hetkimo.nl/richtlijnen/wortelcaries-bij-ouderen/introductie/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn rust-ECG. SAN Centra voor Medische Diagnostiek; 2018. https://de-san.nl/Praktijkrichtlijnen/.
- Nederlandse Vereniging voor Kindergeneeskunde, Nederlands Huisartsen Genootschap, Nederlandse Vereniging van Ziekenhuisapothekers, Jeugdreuma Vereniging Nederland. Richtlijn Juveniele idiopathische artritis (JIA). Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2018. https://richtlijnendatabase.nl/richtlijn/juveniele_idiopathische_artritis_jia/startpagina_-_jia.html.
- Nederlandse Internisten Vereniging, Nederlandse vereniging van Maag-Darm-Leverartsen, Nederlandse Vereniging voor Radiologie, Vereniging Klinische Genetica Nederland, Nederlandse Vereniging voor Klinische Chemie en Laboratoriumgeneeskunde, Hemochromatose Vereniging Nederland. Richtlijn Hereditaire hemochromatose. Utrecht: Nederlandse Internisten Vereniging; 2018. https://richtlijnendatabase.nl/richtlijn/hereditaire_hemochromatose_hh/startpagina.html.
- Nederlands Huisartsen Genootschap, Nederlandse Internisten Vereniging. Richtlijn chronische nierschade. Utrecht: Nederlands Huisartsen Genootschap, Nederlandse Internisten Vereniging; 2018. https://richtlijnendatabase.nl/richtlijn/chronische_nierschade_cns/startpagina_-_chronische_nierschade_cns.html.
- De Grauw W, De Leest K, Schenk PW, Scherpbier-De Haan N, Tjin-A-Ton J, Tuut MK, Van Balen J. NHG-Standaard Chronische nierschade. Utrecht: Nederlands Huisartsen Genootschap; 2018. https://richtlijnen.nhg.org/standaarden/chronische-nierschade.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn gynaecologische echografie. SAN Centra voor Medische Diagnostiek; 2017. https://de-san.nl/Praktijkrichtlijnen/.
- Nederlandse Vereniging voor Neurologie, Nederlandse Vereniging voor Kindergeneeskunde, Nederlandse Vereniging voor Kinderneurologie. Richtlijn Migraine. Utrecht: Nederlandse Vereniging voor Neurologie; 2017. https://richtlijnendatabase.nl/richtlijn/hoofdpijn/migraine.html.
- Nederlandse Vereniging voor Kindergeneeskunde, Nederlandse Vereniging voor Medische Microbiologie, Nederlandse Vereniging voor Obstetrie en Gynaecologie, Nederlandse Vereniging van Ziekenhuisapothekers, Nederlandse Vereniging voor Klinische Chemie en Laboratoriumgeneeskunde, Patiëntenfederatie Nederland, et al. Richtlijn Preventie en behandeling van early-onset neonatale infecties. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2017. https://richtlijnendatabase.nl/richtlijn/preventie_en_behandeling_van_early-onset_neonatale_infecties/informatie_en_ondersteuning_bij_early-onset_neonatale_infecties.html.
- Nederlandse Vereniging voor Kindergeneeskunde. Richtlijn follow-up van kinderen na opname op een intensive care. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2017. https://www.nvk.nl/themas/kwaliteit/richtlijnen/richtlijn?componentid=6881283&tagtitles=Intensive%252bCare.
- Nederlandse Internisten Vereniging, Nederlandse Vereniging van Spoedeisende Hulp Artsen, Nederlandse Vereniging voor Anesthesiologie, Nederlandse Vereniging voor Cardiologie, Nederlandse Vereniging voor Psychiatrie, Nederlandse Vereniging van Ziekenhuisapothekers, et al. Richtlijn Intoxicaties: eerste opvang in het ziekenhuis. Utrecht: Nederlandse Internisten Vereniging; 2017. https://richtlijnendatabase.nl/richtlijn/intoxicaties_eerste_opvang_in_het_ziekenhuis/startpagina_intoxicaties.html.
- Alliantie Transgenderzorg. Kwaliteitsstandaard Psychische transgenderzorg. Amsterdam: Alliantie Transgenderzorg; 2017. https://www.transvisie.nl/wp-content/uploads/2018/01/kwaliteitsstandaard-transgenderzorg-18122017-geautoriseerd.pdf.
- Recommendations in laparoscopic and robotic surgery in urology. Utrecht: Nederlandse Vereniging voor Urologie; 2017. https://portal.nvu.nl/WebserviceWordpress/qgws.asmx/nvu_get_document?id=F7BBADE9-EC12-E711-9438-005056B31E13.

- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn echocardiografie. SAN Centra voor Medische Diagnostiek; 2016. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn ambulante ECG-registratie. SAN Centra voor Medische Diagnostiek; 2016. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn inspannings-ECG. SAN Centra voor Medische Diagnostiek; 2016. https://de-san.nl/Praktijkrichtlijnen/.
- Nederlandse Vereniging voor Kindergeneeskunde, Nederlands Huisartsen Genootschap, Stichting Kind en Ziekenhuis. Evidence-based richtlijn diagnostiek naar onderliggende aandoeningen bij kinderen met recidiverende luchtweginfecties. Rotterdam: Sophia Kinderziekenhuis; 2016. https://kinderinfectieziekten.nl/wp-content/uploads/2016/09/12-07-2016-Diagnostiek-recidiverende-luchtweginfecties.pdf.
- Nederlandse Vereniging voor Kindergeneeskunde. Richtlijn Coeliakie. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2015.
- Nederlandse Vereniging voor Kindergeneeskunde. Richtlijn diagnostiek en behandeling van ongecompliceerde pneumonie bij kinderen in de tweede en derde lijn. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2015. https://www.nvk.nl/themas/kwaliteit/richtlijnen/richtlijn?componentid=7864320&tagtitles=Infectieziekten%252ben%252bImmunologie%2cIntensive%252bCare%2cLongziekten.
- Nederlandse Internisten Vereniging, Nederlandse Vereniging voor Klinische Geriatrie, Vereniging van Specialisten Ouderengeneeskunde, Verpleegkundigen & Verzorgenden Nederland. Richtlijn Slaapproblemen bij acuut opgenomen ouderen. Utrecht: Nederlandse Internisten Vereniging; 2015. https://richtlijnendatabase.nl/richtlijn/slaapproblemen_bij_acuut_opgenomen_ouderen/slaapproblemen_-_korte_beschrijving.html.
- Nederlands Huisartsen Genootschap. NHG-Standaard COPD. Utrecht: Nederlands Huisartsen Genootschap; 2015.
- Nederlands Huisartsen Genootschap. NHG-Standaard Astma bij Volwassenen. Utrecht: Nederlands Huisartsen Genootschap; 2015.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn ureum ademtest. SAN Centra voor Medische Diagnostiek; 2014.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn MRI. SAN Centra voor Medische Diagnostiek; 2014. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn DEXA. SAN Centra voor Medische Diagnostiek; 2014. https://de-san.nl/Praktijkrichtlijnen/.
- SAN Centra voor Medische Diagnostiek. Praktijkrichtlijn enkel-arm index. SAN Centra voor Medische Diagnostiek; 2014. https://de-san.nl/Praktijkrichtlijnen/.
- Nederlandse Vereniging voor Urologie, Nederlandse Internisten Vereniging, Nederlandse Vereniging voor Radiologie, Nederlandse Vereniging voor Klinische Chemie en Laboratoriumgeneeskunde. Richtlijn Nierstenen. Utrecht: Nederlandse Vereniging voor Urologie; 2014. https://richtlijnendatabase.nl/richtlijn/nierstenen/startpagina_richtlijn_nierstenen_2023.html.
- Nederlandse Vereniging voor Kindergeneeskunde. Richtlijn Bronchopulmonale dysplasie (BPD). Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2014.
- Nederlands Huisartsen Genootschap. NHG-Standaard Astma bij kinderen. Utrecht: Nederlands Huisartsen Genootschap; 2014.
- Long Alliantie Nederland, Koninklijke Nederlandse Maatschappij ter bevordering der Pharmacie, Longfonds, Nederlands Huisartsen Genootschap, Nederlandse Vereniging van Artsen voor Longziekten en Tuberculose, Nederlandse Vereniging voor Kindergeneeskunde, et al. Multidisciplinaire richtlijn astma. Amersfoort: Long Alliantie Nederland; 2014. https://www.longalliantie.nl/content/Multidisciplinaire_richtlijn_astma_STATUS_GEAUTORISEERD_28012014_def.pdf.
- Nederlandse Vereniging voor Kindergeneeskunde, Nederlandse Vereniging voor Radiologie, Nederlandse Vereniging van Ziekenhuisapothekers. Richtlijn Koorts bij kinderen. Utrecht: Nederlandse Vereniging voor Kindergeneeskunde; 2013. https://richtlijnendatabase.nl/richtlijn/koorts_bij_kinderen/koorts_bij_kinderen_-_startpagina.html.
- Nederlandse Vereniging voor Dermatologie en Venereologie, Nederlandse Internisten Vereniging, Nederlandse Vereniging voor Heelkunde, Nederlandse Vereniging voor Medische Microbiologie. Richtlijn Cellulitis-Erysipelas onderste extremiteiten. Utrecht: Nederlandse Vereniging voor

Dermatologie en Venereologie; 2013. https://richtlijnendatabase.nl/richtlijn/cellulitis-erysipelas_onderste_extremiteiten/cellulitis_en_erysipelas_-_korte_beschrijving.html.
- Nederlandse Vereniging voor Pathologie, Nederlandse Internisten Vereniging, Nederlandse Vereniging van Artsen voor Longziekten en Tuberculose, Nederlandse vereniging van Maag-Darm-Leverartsen, Nederlandse Vereniging voor Heelkunde, Nederlandse Vereniging voor Keel-Neus-Oorheelkunde en Heelkunde van het Hoofd-Halsgebied, et al. Richtlijn Primaire tumor onbekend. Utrecht: Nederlandse Vereniging voor Pathologie; 2012. https://richtlijnendatabase.nl/richtlijn/primaire_tumor_onbekend/primaire_tumor_onbekend_-_startpagina.html.
- Nederlandse Vereniging van Hoofdpijnpatiënten, Nederlands Huisartsen Genootschap, Nederlands Instituut van Psychologen, Nederlands Oogheelkundig Gezelschap, Nederlandse Maatschappij tot bevordering der Tandheelkunde, Nederlandse Vereniging voor Anesthesiologie, et al. Richtlijn Chronische aangezichtspijn. Utrecht: Nederlandse Vereniging van Hoofdpijnpatiënten; 2012. https://www.neurologie.nl/wp-content/uploads/2021/10/2012-11-21-Richtlijn-Chronische-Aangezichtspijn.pdf.
- Nederlandse Internisten Vereniging. Richtlijn Elektrolytstoornissen. Utrecht: Nederlandse Internisten Vereniging; 2012. https://publicatie.internisten.nl/wp-content/uploads/2022/12/richtlijn_2012_elektrolytstoornissen.pdf.
- VMS Veiligheidsprogramma. Veilige zorg voor zieke kinderen. Utrecht: VMS; 2011. https://www.vmszorg.nl/wp-content/uploads/2017/11/web_2011.0114_praktijkgids_kinderen.pdf.
- Nederlandse Vereniging voor Arbeids- en Bedrijfsgeneeskunde. Richtlijn Influenzapandemie. Utrecht: Nederlandse Vereniging voor Arbeids- en Bedrijfsgeneeskunde; 2011.
- Long Alliantie Nederland. Richtlijn Palliatieve zorg bij COPD. Amersfoort: Long Alliantie Nederland; 2011.
- Nederlandse Vereniging van HIV Behandelaren. Richtlijn HIV. 2010-2024. https://richtlijnhiv.nvhb.nl/index.php/Inhoud.
- Nederlandse Vereniging voor Reumatologie, Nederlands Oogheelkundig Gezelschap, Nederlandse Internisten Vereniging, Nederlandse Vereniging van Artsen voor Longziekten en Tuberculose, Nederlandse vereniging van Maag-Darm-Leverartsen, Nederlandse Vereniging voor Dermatologie en Venereologie, et al. Richtlijn Kleine vaten vasculitis. Utrecht: Nederlandse Vereniging voor Reumatologie; 2010. https://richtlijnendatabase.nl/richtlijn/kleine_vaten_vasculitis/vasculitis_-_korte_beschrijving.html.
- VMS Veiligheidsprogramma. High risk medicatie: klaarmaken en toedienen van parenteralia. Utrecht: VMS; 2009. https://www.vmszorg.nl/wp-content/uploads/2017/03/2009.0108_praktijkgids_high_risk.pdf.
- VMS Veiligheidsprogramma. Verwisseling van en bij patiënten. Utrecht: VMS; 2009. https://www.vmszorg.nl/wp-content/uploads/2017/11/web_2009.0107_praktijkgids_verwisseling.pdf.
- Nederlandse Vereniging voor Dermatologie en Venereologie. Richtlijn basaalcelcarcinoom. Utrecht: Nederlandse Vereniging voor Dermatologie en Venereologie; 2008.
- Nationaal Borstkanker Overleg Nederland. Multidisciplinaire richtlijn borstkanker. Utrecht NABON; 2008.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Medicamenteuze therapie van COPD. Utrecht: CBO; 2007.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Foto(chemo)therapie en systemische therapie bij ernstige chronische plaque psoriasis. Utrecht: CBO; 2003.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Aspecifieke lage rugklachten. Utrecht: CBO; 2003.
- Nationaal Borstkanker Overleg Nederland. Richtlijn Behandeling van het mammacarcinoom. Utrecht: NABON; 2002.
- Nationaal Borstkanker Overleg Nederland. Richtlijn Mammacarcinoom: screening en diagnostiek. Utrecht: NABON; 2000.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Hoge bloeddruk. Utrecht: CBO; 2000.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Beroerte. Utrecht: CBO; 2000.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Antiretrovirale behandeling in Nederland. Utrecht: CBO; 2000.

- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Diepe veneuze trombose en longembolie. Utrecht: CBO; 1999.
- Kwaliteitsinstituut voor de Gezondheidszorg CBO. Richtlijn Behandeling en preventie van coronaire hartziekten door verlaging van de plasmacholesterolconcentratie. Utrecht: CBO; 1998.

Over de auteur / About the author

Over de auteur

Mariska Tuut werd geboren op 26 januari 1975 in Groningen. Na haar studie Biomedische Gezondheidswetenschappen, richting Epidemiologie, aan de Katholieke Universiteit Nijmegen werkte zij korte tijd als epidemiologisch onderzoeker aan de Westfälische Wilhelms-Universität Münster (Duitsland). Al snel werd zij gegrepen door de combinatie van wetenschap en beleid en werd aangenomen als procesbegeleider richtlijnontwikkeling bij het Centraal Begeleidingsorgaan voor de Intercollegiale Toetsing (het latere Kwaliteitsinstituut voor de Gezondheidszorg CBO) in Utrecht. Daar werkte zij mee aan de transitie van consensus-based richtlijnontwikkeling naar evidence-based richtlijnontwikkeling. Naast het begeleiden van richtlijnwerkgroepen was zij ook betrokken in de ontwikkeling van methodieken en training. Na enkele omzwervingen binnen de kwaliteitszorg maar buiten de richtlijnontwikkeling, startte Mariska in 2009 haar eigen bedrijf PROVA. Als onafhankelijk richtlijnmethodoloog is zij betrokken bij de ontwikkeling van richtlijnen voor diverse doelgroepen van zorgverleners en patiënten/cliënten/burgers. Ook geeft zij training en coaching aan beginnende en meer ervaren richtlijnontwikkelaars. Daarnaast is ze actief in het Dutch GRADE Network, Richtlijnennetwerk Nederland (RNN) en RNN-GENEVER met als doel de ontwikkeling en implementatie van richtlijnmethodieken te bevorderen. In 2018 startte zij met een buitenpromotietraject aan het Care and Public Health Research Institute (CAPHRI) van de Universiteit van Maastricht. Het onderzoek heeft zij steeds gecombineerd met haar werkzaamheden als richtlijnmethodoloog.

Mariska woont gelukkig samen met Erie Vriese in Varsseveld. Zij hebben drie jongvolwassen kinderen: Laura, Jara en Marijn.

About the author

Mariska Tuut was born on 26 January 1975 in Groningen, the Netherlands. After graduating in Biomedical Health Sciences with a major in Epidemiology from the Catholic University of Nijmegen, she worked briefly as an epidemiological researcher at the Westfälische Wilhelms-Universität Münster (Germany). She was quickly attracted to the combination of science and policy and was appointed process lead guideline development at the Dutch Institute for Healthcare Improvement in Utrecht. At this institution, she was involved in the transition from consensus-based guideline development to evidence-based guideline development. In addition to working on guideline panels, she was also involved in the development of methodologies and training. After some time spent in quality assurance but outside guideline development, Mariska started her own company PROVA in 2009. As an independent guideline methodologist, she has been involved in the development of numerous guidelines for various target groups of healthcare providers and patients /citizens. She also provides training and coaching to novice and more experienced guideline developers. Additionally, she is active in the Dutch GRADE Network, Guidelines Network Netherlands (RNN) and RNN-GENEVER, with the objective of advancing the development and implementation of guideline development methods. In 2018, she started an external PhD track at the Care and Public Health Research Institute (CAPHRI) on Maastricht University. She has consistently combined research with her work as guideline methodologist.

Mariska lives happily with Erie Vriese in Varsseveld. They have three adult children: Laura, Jara and Marijn.

Dankwoord

Dankwoord

'Nobody said it was easy'
The scientist – Coldplay

De totstandkoming van dit proefschrift was niet mogelijk en niet zo leuk geweest zonder de hulp van anderen. In dit ongetwijfeld meest gelezen hoofdstuk van mijn proefschrift wil ik deze mensen bedanken.

'You can't start a fire without a spark'
Dancing in the dark – Bruce Springsteen

Allereerst mijn 'promoteam' zonder wie dit traject onmogelijk was geweest: prof. dr. Trudy van der Weijden, prof. dr. Jako Burgers en dr. Miranda Langendam.

Trudy, tijdens GIN in Amsterdam spraken wij over het belang van goede richtlijnen, methodiek van richtlijnontwikkeling, bekwaamheid van richtlijnontwikkelaars en de mogelijke gevolgen van beperkingen daarin voor de gezondheidszorg, mensen en maatschappij. Jij deelde mijn gevoel van urgentie en gaf mij de mogelijkheid met dit proefschrift een steentje bij te dragen aan een oplossing. Héél veel dank hiervoor! Daarnaast was jij het die tijdens dit traject altijd de haalbaarheid en de stip op de horizon in het vizier hield, waardoor er nu een mooi boekje ligt waarop ik heel trots ben.

Jako, jij ging me voor met een proefschrift op richtlijnmethodologisch gebied, ook een meta-onderwerp, dat heel goed geïmplementeerd is in de richtlijnenwereld, en alleen daarom al is het een voorrecht dat je mijn promotor wilde zijn. Daarnaast was jouw klinische en wetenschappelijke inbreng van grote waarde en zorgde je er ook voor dat ik met beide voeten op de grond bleef, ook al gingen we soms op de filosofische toer. Veel dank voor je kritische inbreng en vele, vele, correcties!

Miranda, wat fijn dat jij er bij was! Niet alleen bracht je enorme deskundigheid en ervaring in met betrekking tot de methodologie van systematische reviews, GRADE en testevaluatie, maar vooral: je was er! Bij grote vragen, kleine vragen, als het even niet zo liep, of als het juist heel goed ging. Met vragen, opmerkingen, mailtjes, telefoontjes en appjes. En met koffie, wijn en ander lekkers. Door jou was dit traject een feestje!

'We are the champions, my friends'
We are the champions – Queen

My name is on the front cover of this booklet, but its creation is a team effort: A big, big, thank you to my co-authors, it was an honour and a pleasure working with you: dr. Hans de Beer, prof. dr. Patrick Bindels, prof. dr. Patrick Bossuyt, prof. dr. Jochen Cals, dr. Gowri Gopalakrisha, dr. Erik-Jonas van de Griendt, dr. Jesse Jansen, prof. dr. Mariska Leeflang, dr. Reem Mustafa, Hester Rippen, dr. Corinna Schaefer and prof. dr. Holger

Schünemann, keep up the good work! Ook de (anonieme) deelnemers van de studie die is beschreven in hoofdstuk 6 ben ik erg dankbaar voor hun belangeloze medewerking.

'Speaking words of wisdom'
Let it be – The Beatles

De voorzitter en leden van de beoordelingscommissie prof. dr. Marian Majoie, prof. dr. Silvia Evers, dr. Merit Tabbers en prof. dr. Philip van der Wees ben ik dankbaar voor hun kritische beoordeling van mijn proefschrift.

'We can be heroes, just for one day'
Heroes – David Bowie

Ik voel me zeer vereerd dat de hooggeleerde en zeergeleerde opponenten prof. dr. Marian Majoie, prof. dr. Philip van der Wees, prof. dr. Silvia Evers, prof. dr. Erwin Berkhout, dr. Dunja Dreesens en prof. dr. Jochen Cals met mij van gedachten willen wisselen over mijn proefschrift. Ik kijk erg uit naar 8 oktober!

'You paved the way, believe it'
Waka waka – Shakira

Ook al was ik 'external PhD candidate' en ben ik (mede 'dankzij' COVID-19 en telecommunicatiemiddelen als Zoom) maar enkele malen 'op de Uni' geweest, ik heb me er buitengewoon welkom gevoeld. Dank aan alle HAG'ers daarvoor. Twee medepromovendi hebben daar een extra belangrijke rol in gespeeld. Als eerste mijn 'buddy' Romy Richter met wie ik ongeveer gelijk op liep en met wie ik 'tips & tricks' over van alles kon uitwisselen. En ten tweede, heel bijzonder, mijn dochter Laura Vriese, 'mijn wicht in Maastricht', door wie dit avontuur een gouden randje kreeg.

'I'll be there for you (, 'cause you're there for me too)'
I'll be there for you – The Rembrandts

Het werk aan dit proefschrift gebeurde naast mijn werk als zelfstandig onafhankelijk richtlijnmethodoloog. Tijdens dit proefschrifttraject werkte ik mee aan diverse richtlijnen en aanverwante producten, gaf ik veel training en coaching en zat ik in diverse commissies. Besturen, directies, medewerkers, voorzitters en leden van richtlijnwerkgroepen/klankborden/subgroepen/commissies en cursisten van ACTA, AQUA, CLMC, Dutch GRADE Network, Erasmus MC, IQ Healthcare, KIMO, KNGF, NHG, NIV, NJI, NVHB, NVK, RAILZ, RIVM, RNN, SAN, SKILZ, V&VN en ZonMw: heel veel dank voor de samenwerking met jullie, voor alle inspiratie, voor de ervaring die we hebben opgedaan, voor de mooie producten die we hebben gemaakt, voor het plezier dat we daarbij hebben gehad, voor jullie vertrouwen en voor jullie belangstelling voor mijn

proefschrift. Ondanks, maar vooral dankzij jullie is het proefschrift op deze manier af gekomen. Eén persoon wil ik daarbij speciaal noemen: dr. Jannes van Everdingen, met wie ik de laatste jaren vooral samenwerkte in de door hem voorgezeten ZonMw-commissie Beter Thuis, maar die ik vooral zie als 'de vader van de Nederlandse richtlijnontwikkeling'. In 1998 zag hij een functie voor mij als junior procesbegeleider bij het CBO wel zitten, waarmee hij mijn enthousiasme voor richtlijnontwikkeling aanwakkerde.

'Side by side, we'll make things better'
Never alone – 2 Brothers on the 4th floor

GENEVER heeft een bijzonder plekje in mijn richtlijnenhart. Ooit begonnen als EBRO-platform bij het CBO en NHG, uitgegroeid tot een zeer gewaardeerd netwerk van betrokken richtlijnontwikkelaars, nu klaar om opgenomen te worden in RichtlijnenNetwerk Nederland (RNN). De halfjaarlijkse GENEVER-bijeenkomsten zitten altijd vól enthousiasme van sprekers, deelnemers en lokale hosts. Ook de voorbereiding door ons organisatiecomité verloopt altijd heel energiek en vol goede ideeën. Ik ben dan ook heel trots dat de verdediging van mijn proefschrift in de ochtend voorafgegaan wordt door een GENEVER-bijeenkomst, tevens de eerste bijeenkomst van GENEVER onder de vlag van RNN. Veel dank aan de sprekers, deelnemers en CAPHRI als lokale host op 8 oktober. En vooral wil ik op deze plek mijn mede-organisatiecomitéleden Mitchell van Doormaal, Ilse Verstijnen en Jolanda Wittenberg bedanken voor de organisatie van een ongetwijfeld weer geslaagde bijeenkomst, maar bovenal voor de geweldige samenwerking in de afgelopen jaren.

'I get by with a little help from my friends'
With a little help from my friends – Joe Cocker

Geen inspanning zonder ontspanning, geen ontspanning zonder inspanning: daarom is het fijn dat er naast mijn werk en proefschrift ook nog andere inspannende en ontspannende zaken zijn! Dank daarom aan: Reunited, mijn fiets- (en triathlon!)maatje (what happens on de fiets, stays on de fiets), mijn lokaal belangrijke vriendin, de culturele bezighedengroep, mijn vreugde- & geluk club, mijn medewaterpoloërs van DOS Varsseveld, de dames van Houdt Moed (in het bijzonder mijn medebestuursleden), de wijnclub, mijn lieve buren, familie en overige vrienden! Ondanks dat niet iedereen van jullie precies wist waar ik nou exact mee bezig was, heb ik genoten van jullie belangstelling en vragen of mijn opleiding/studie/project/'dat in Maastricht'/onderzoek/boekje al een beetje opschoot/naar wens verliep/nou nog niet klaar was; het is af.

'Winter, spring, summer or fall, all you have to do is call. and I'll be there'
You've got a friend – Carole King

Eén vriendin noem ik hierbij wél bij naam: Mika, wat zijn we verschillend, wat zijn we hetzelfde en wat delen we al heel lang heel veel: vooral veel lief, en ook leed: mijn (iets te letterlijke) boezemvriendin. Wat een geluk dat ik je ken.

Ik begon dit dankwoord met het noemen van mijn promoteam zonder wie de start van dit traject niet mogelijk was geweest. Ik eindig met de mensen zonder wie het niet mogelijk was geweest om het door te zetten en af te ronden:

'Ja, 't giet zoas 't giet, daor ku'j van op an, heb der mar fiducie in'
't Giet zoas 't giet – Skik

Papa en mama ('Oend en Peun'), en later Marjan en Luuk: van jullie leerde ik verantwoordelijkheid dragen, voor jezelf, anderen en de wereld. En ik leerde ergens voor te gaan en door te zetten, ook als dat niet zo makkelijk is. Jullie zijn een voorbeeld en wat ben ik blij en dankbaar dat jullie bij de verdediging van mijn proefschrift zijn.

'You can check out anytime you like, but you can never leave'
Hotel California – The Eagles

Bas, Merel en Rens (in volgorde van binnenkomst), wat fijn dat jullie ook bij ons gezin horen! Ik hoop van harte dat jullie en onze kinderen elkaar, en daarmee ook ons, nog heel lang gelukkig blijven maken!

'You should know, everywhere I go, always on my mind, in my heart, in my soul'
You're the inspiration – Chicago

Lieve, lieve, lieve Laura, Jara en Marijn! Wat ben ik een trotse moeder! Jullie zijn geweldige, verantwoordelijke, doorzettende, empathische en humoristische mensen geworden, wat fantastisch! En wat ben ik blij en dankbaar dat jullie – met z'n drieën - als paranimfen aan mijn zijde staan bij de verdediging van mijn proefschrift. Jullie maken de wereld mooier, ik heb er alle vertrouwen in!

'Gloria a te ogni volta'
E'un peccato morir – Zucchero

En tot slot: lieve, lieve, lieve Erie! Ik maak me belachelijk als ik beschrijf hoeveel je voor me betekent! Bedankt voor wat we delen, wat je voor mij en ons doet en vooral voor wie je bent ♥