BMJ Open

Barriers to making recommendations about medical tests: a qualitative study of European guideline developers

Gowri Gopalakrishna,[1] Mariska M G Leeflang,[1] Clare Davenport,[2] Andrea Juliana Sanabria,[3] Pablo Alonso-Coello,[3] Kirsten McCaffery,[4] Patrick Bossuyt,[1] Miranda W Langendam[1]

To cite: Gopalakrishna G, Leeflang MMG, Davenport C, *et al*. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. *BMJ Open* 2016;**6**:e010549. doi:10.1136/bmjopen-2015-010549

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-010549).

Received 24 November 2015
Revised 1 May 2016
Accepted 28 July 2016

For numbered affiliations see end of article.

Correspondence to
Gowri Gopalakrishna;
g.gopalakrishna@amc.uva.nl

ABSTRACT

Objectives: Development of medical test guidelines differs from intervention guideline development. These differences can pose unique challenges in building evidence-based recommendations to guide clinical practice. The aim of our study was to better understand these challenges, explore reasons behind them and identify possible solutions.

Setting and participants: In this qualitative study, we conducted in-depth interviews between February 2012 and April 2013 of a convenience sample of 17 European guideline developers experienced in medical test guideline development.

Outcomes measured: We used framework analysis with deductive and inductive approaches to generate the themes from the interviews. We kept interpretation grounded in the data.

Results: Guideline developers acknowledged that inclusion of patient important outcomes in their guideline development was necessary but lacking. This and other challenges raised fell into 3 broad and overlapping domains: methodological issues, resource limitations and a lack of awareness on the need for evidence that links testing to patient outcomes. Education was mentioned as a key solution to increase awareness and address the resources limitations mentioned.

Conclusions: Challenges guideline developers face were interlinked across the domains of methodological issues, resource limitations and a lack of awareness. Solutions that addressed these challenges in parallel are needed. Raising awareness, education and training of relevant stakeholders such as medical doctors, funders and regulators to look beyond test accuracy is key to having a long-term resolution to the issues faced in medical test guideline development.

Strengths and limitations of this study

- Our study findings are in line with other research that call for better regulatory frameworks for medical test approval.
- Our findings support other research that have shown health professionals find it difficult to understand test accuracy measures and relate that to downstream patient consequences.
- Our sample size demonstrates maximum variation in terms of guideline development expertise, topics covered and size of guideline organisations from which interviewees came from.
- We provide a number of practical solutions to overcome some of the challenges raised.
- Our interviewees were mainly methodologists and hence may not adequately represent the views of different types of panelists on a guideline panel.

INTRODUCTION

The process of guideline development combines technical and quantitative methods of evidence appraisal and synthesis with group processes within the guideline panel when moving from evidence synthesis to making recommendations.[1 2] Development of medical test guidelines should not be restricted to a synthesis of evidence of a test's accuracy but should also address the consequences of test use for the patient receiving the test. Known as patient important outcomes, these can range from clinical outcomes, such as mortality and morbidity, to quality of life outcomes, such as a reduction in anxiety as a result of testing.[3–5] The Agency for Healthcare Research and Quality (AHRQ)[6] defines these into five categories that may be the result of the test, testing process or both: (1) outcomes that result from clinical management based on the test results; (2) the direct health effects of testing; (3) the patients' emotional, social, cognitive and behavioural responses to testing; (4) the legal and ethical effects of testing and (5) the costs of testing.

Awareness on the need to base recommendations on patient important outcomes is growing. For example, the National Institute

for Health and Clinical Excellence's (NICE) adoption of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, an evidence-based grading system developed to ensure guideline recommendations are patient outcome centred, is an indication of the growing appreciation of the consequences of testing.[7]

However, evidence available to guideline developers remains predominantly focused on analytical or clinical test performance, such as that evaluated in test accuracy studies which rarely report on the downstream effects a test may have for patients.[8] In addition, test accuracy evidence is often of poor methodological quality and of limited applicability to the intended setting in which the test is to be used. This adds further complexity to the synthesis and interpretation of a test's performance.[9–11] Tests are never administered in isolation but as part of a testing strategy. Guideline panels therefore need to take into consideration the place of the test being appraised in a testing strategy in order to make judgements on the downstream patient consequences for further testing and treatment.[12 13] These multiple methodological complexities put together can make the process of making evidence-based recommendations for medical testing challenging.[5]

We adopted a qualitative approach involving in-depth interviews with an international group of guideline developers to explore these issues. Our aim was to identify ways in which the guideline development process for medical tests can be improved. Specifically, we wanted to gain a deeper understanding of these and other challenges guideline developers may face when making guidelines around medical tests, to explore their beliefs as to why these challenges exist and to identify possible solutions and areas for improvement that could help overcome these difficulties.

METHODS

Interview topic guide

Informed by previous research in the area of medical test guideline development, we prepared an interview topic guide outlining the areas in medical test guideline development that we intended to better understand (see online supplementary appendix 1). The topic guide was based on major steps in guideline development. This was piloted and modified based on results from the piloting. Semistructured, in-depth interviews were chosen for this study so as to allow interviewees to bring up and discuss, in their own words, relevant points beyond those anticipated by the researchers and to allow us to seek clarification about viewpoints expressed throughout the process.

Sampling and recruitment

Purposeful, theoretical sampling was used to recruit a convenience sample of international guideline developers, known within our network, which consisted mostly of methodologists with experience in medical test guideline development. This group we felt would provide us with a good understanding of challenges faced by the panel given that their role is in assisting all the members of the panel from start to end of a guideline. We sought maximum variation in our sample in terms of guideline topic (ranging from mental health, to laboratory medicine to radiology) and level of guideline developer experience (ranging from small local guideline groups within an institution to large national organisations such as NICE). The participants were approached by email and invited for an interview.

Data collection

We conducted face-to-face interviews at the interviewee's workplace whenever possible. All interviews were conducted in English except when interviewees felt more comfortable in their native language. These were then translated into English. When limited by distance, interviews were conducted over the telephone. GG, CD and AJS each individually conducted the interviews with GG conducting majority of the interviews. Interviewers were sufficiently knowledgeable in the area of testing and guideline development as part of their daily research. Each interviewer received the topic guide with interviewer instructions for use during their interviews. Each interview consisted of interviewer and interviewee only.

Participants were not personally known to the interviewers prior to this study and were provided a brief introduction of the background of interviewer and purpose of the study at first email contact. Interviews were recorded using a digital recorder and transcribed verbatim. Transcripts were checked for accuracy (GG listened to all audio recordings and compared these to the transcribed text). Any inaccuracies (eg, place, names and spelling errors) in the transcribed text were edited appropriately. Transcripts were not returned to participants for comment and/or correction.

Data analysis

Analysis of data generated from interviews was performed concurrently with data collection. Deductive and inductive approaches were used, but interpretation was kept grounded in the data (ie, grounded theory). Preliminary analysis of the first 10 interviews was conducted to evaluate if data saturation was reached. The interview topic guide was subsequently modified after discussion among the authors to focus on those aspects where data saturation was not yet reached. Data collection ceased when no new themes emerged from the interviews.

We used the framework analysis method[14 15] and followed the five analytical stages as recommended: (1) familiarisation with the data; (2) developing a thematic framework through the identification of main topics and subtopics; (3) indexing (coding the data) into themes; (4) charting by arranging summaries of the data on a case-by-case basis (which in our study was supported

through the use of an Excel spreadsheet) and (5) mapping and interpreting the data by examining the data for patterns. We used Microsoft Excel to develop the coding framework. In deriving the coding framework, we started out with a deductive list of codes derived from the interview topic guide and our own understanding of guideline development in medical tests. GG, MWL and MMGL coded a random set of interviews in triplicate to validate the initial coding framework and to check for consistency in interpretation of the data. Once the coding framework was finalised, GG coded the remaining interviews. Charting of a random sample of coded interviews into summaries was also initially performed in triplicate by GG, MWL and MMGL to check for consistency in interpretation. Thereafter, GG charted the remaining interviews.

At every stage of the coding and analysis, we kept the coding framework deductive and inductive allowing for emergent, data-driven codes and subcodes to be included into the framework. Codes and subcodes were supported by quotations derived from the interviews. GG, MWL, MMGL and PB met at regular intervals throughout the analysis phase to discuss emergent issues and themes until consensus was reached on how data should be interpreted. Brief summaries and representative quotes for each category were abstracted from the transcripts for reporting purposes to illustrate typical responses and/or the diversity of views expressed. An audit trail was maintained throughout the study to document all reasons behind decisions taken from start to finish of this study. Participants did not provide feedback on the findings. We used the consolidated criteria for reporting qualitative studies (COREQ): a 32-item checklist[16] to ensure our study meets the recommended standards of qualitative data reporting.

RESULTS

A total of 17 interviews were conducted between Feb 2012 and April 2013. An additional two participants contacted declined participation on the grounds that they felt they had not sufficient medical test guideline development experience. The average length of an interview was 61 min (range 36–96 min). There were in total 359 pages of transcribed text. The background information of interviewees is described in table 1. Table 2 provides an overview of the key themes, messages and quotations, and online supplementary table S1 gives the full details.

Challenges reported by the interviewees could be categorised into three main domains: methodological issues, resource limitations and a lack of awareness. Interconnectedness of two or more of these domains was a central theme for all challenges identified throughout the study (figure 1). For example, when it came to key question formulation, interviewees felt guideline panels lacked the awareness of the importance of including patient outcomes explicitly in the guideline question. The use of a test-treatment pathway (to illustrate the position of the test in a testing strategy, and management decisions consequent on test results) was suggested by interviewees as a tool that could help the guideline panel understand the value of a test's downstream consequences and thereby help the development of patient outcome centred guideline questions. However, this could mean more time and training is needed of the panel in order to be able to do this. Here, we see the domains of awareness, resource limitations and methodological issues that are interconnected in key question formulation stage. This interconnectedness was a central theme throughout the study (figure 1, table 2 and see online supplementary table S1).

Table 1 Overview of characteristics of the 17 interviewees

Category (number)	
Type of guideline group (n)	Institutional (3) National (11) International (3)
Countries of interviewees (n)	UK (6), Germany (1), Belgium (2), the Netherlands (3), USA (1), Spain (2), Australia (1), Finland (1)
Size of guideline development group (range)	10–20 panel members
Areas interviewees have developed medical test guidelines for	Paediatrics, mental health, womens' health, point-of-care tests, oncology, acute pain, laboratory medicine, celiac disease, diabetes, tuberculosis
Role of interviewee in guideline development group (n)	Methodologist (17)
Interviewed face to face or via telephone (n)	Face to face (11) Telephone (6)

Guideline developers are limited by challenges in methodological issues, resources limitations and a lack of awareness at each of the following guideline development stages:

Scoping the topic

> It takes a fair chunk of an analyst's time over several months to go through scoping, yes. It's absolutely critical that the problem is well defined…to understand exactly what all the ins and outs of the problem are and it's not something you just throw together. It's one of the things that makes diagnostics different. It requires a vastly, a more complicated problem definition phase that you would typically have for treatment. (ID 1)

The scoping phase can involve a number of preliminary literature searches, and a multidisciplinary team ranging from experts in the field to public consultations to

Table 2 Summary of key themes and suggested solutions/future areas for development

Impacted area	Key message	Representative quote (ID)	Suggested solutions/future areas for development (when applicable)
1. Guideline development stages			
Scoping	Scoping in diagnostic guideline is more extensive and resource intensive compared to intervention guidelines.	"It takes a fair chunk of an analyst's time over several months to go through scoping, yes. It's absolutely critical that the problem is well defined…to understand exactly what all the ins and outs of the problem are and it's not something you just throw together. It's one of the things that makes diagnostics different. It requires a vastly, a more complicated problem definition phase that you would typically have for treatment." (ID 1)	
Key question formulation and test—treatment pathway	PICO format for question formulation is not very useful for diagnostics. The panel needs to be educated and trained on how to develop focused questions that include patient outcomes.	"PICO is not very relevant for diagnostic questions…Educating the panel on test's downstream consequences helps define exact questions to be answered." (ID 5) "The panel still need help with question formulation. There is a lack of appreciation of what's required in a question…because a lot of them don't know that, you know. You can call it education of the clinicians." (ID 13)	A test-treatment pathway can help panelists develop focused questions that are patient outcome centred, but the awareness of the panel on its importance needs to be raised and it requires resource commitment in terms of time, money and training.
Impacted area	**Key message**	**Representative quote (ID)**	**Proposed solutions/future areas for development (when applicable)**
Searching and synthesising the evidence	Search filters are not well developed for test accuracy studies; meta-analysis is often complex due to complexity of the methods and the heterogeneous nature of test accuracy studies.	"Yes, we have a lot of them that show very heterogeneous data. That indicates there are a lot of things still to be done and that we should be very cautious of single studies and drawing conclusions." (ID 10) "We do not do meta-analysis because we are not as familiar on how to do this compared to treatment." (ID 3)	We need good search filters for test accuracy studies, training and more explicit guidance on meta-analysis methods. There is a need for better quality primary studies on test accuracy for meaningful data syntheses to occur.
Types of outcomes and evidence	Resource is a major consideration as to whether the panel includes outcomes other than test accuracy. This is compounded by the lack of availability of such data.	"Define the budget, get another team to bring the resource … I can't just extract diagnostic test accuracy studies. It's not a complete enough picture." (ID 11) "Doing qualitative research as part of a guideline would be really useful, but is not possible because of time constraints." (ID 15) "For questions which the panel feel there will be very little or no evidence, other methods should be explored such as Delphi or focus groups for gathering the information." (ID 3)	Inclusion of qualitative data and/or methods (eg, Delphi method and focus groups) should be explored as alternative ways to include patient outcome-related evidence in a structured way in the guideline process.

Continued

Table 2 Continued

Impacted area	Key message	Representative quote (ID)	Suggested solutions/future areas for development (when applicable)
Making recommendations	Expert opinion is important, but in the face of the lack of good quality evidence this can make the process unstructured, not transparent and political. Usefulness of modelling to overcome this lack of evidence has conflicting views.	"There's a discussion about the benefits and harms, about resources and about patient values and preferences. Do we know those? No. Again, people give you their opinions about it, but that's the best we can do at this point." (ID 11) "It's political and then lack of evidence that's not there. You have to build on the expert opinion and that can be quite difficult." (ID 10) "Modelling is the only other answer I know of. Is it probably more accurate most of the time than people just making individual guesses in clinical practice? I would say yes." (ID 1) "Models need assumptions and the assumptions cannot be proved, so it's very uncertain…We don't want to accept this uncertainty. We think it's better to give some pressure on the community to perform such studies." (ID 14)	Delphi processes, focus groups or the use of modelling could make the process more systematic and transparent, but there are contrasting views on this.
1. Awareness, education and training			
Within the guideline panel	Educating the guideline panel about test accuracy statistics and how test accuracy can impact a range of patient outcomes prior to the start of guideline development is crucial.	"The panel finds it very hard to make a choice as to when high sensitivity is important and when is high specificity important. They do not understand the consequences of high sensitivity and low specificity … hence cannot guide the methodologist either on what are important characteristics for the tests." (ID 3)	Guideline panels should consider investing, prior to starting the guideline development, training of the panelists on test accuracy and downstream test consequences. This can be in the form of developing a test-treatment pathway, for example.
Outside the guideline panel	General medical education of doctors in test accuracy was seen as inadequate. Intervention research was perceived as receiving greater focus and funding in the form of RCTs. Regulatory authorities and the medical testing industry need to recognise the importance for end-to-end studies that report on downstream testing consequences.	"Most attention goes to intervention studies in journals and in guidelines normally…in the education of medical professionals there is less focus on diagnostic accuracy. They're not used to it." (ID 3) "RCTs that evaluate full strategies including tests and treatment—we need new funding mechanisms before we get them." (ID 4) "If we change the regulatory process which should be similar to drugs, then I feel we will in a few years have much better studies." (ID 14)	Having a regulatory framework that recognises the importance of tests' downstream consequences can help bring the needed attention to medical test evaluation at several levels that were identified as lacking.

RCTs, randomised controlled trials.

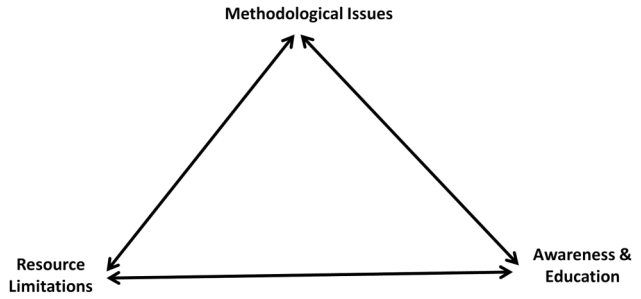

Figure 1 Challenges guideline developers face are interconnected among the domains of methodological issues, resource limitations and the need for awareness and education.

ensure the 'problem' is adequately defined while still remaining doable. In some instances, the scoping phase involves developing a test-treatment pathway as well. The scoping phase can therefore be resource demanding in terms of time and manpower as well as the methodologically demanding in terms of being able to adequately and appropriately define the problem.

Key question formulation

> PICO not very relevant for diagnostic questions but (it's) still used to try to be consistent with intervention format. (ID 5)

> Developing comparative test accuracy questions is complex, its sometimes not just a straightforward diagnostic test accuracy question but involves more the diagnostic pathway…so that's quite complex. (ID 2)

Interviewees mentioned using the Patient-Intervention-Comparator-Outcome (PICO)[17] format for developing the questions to be answered by the guideline yet; they also mention that the PICO is not particularly helpful when it comes to diagnostics questions, especially to particular types of testing questions such as comparative test accuracy questions.

The identification of a test-treatment pathway was mentioned as a useful tool to explicitly link test use and patient outcomes to facilitate question formulation by the guideline panel. Typically, a test-treatment pathway identifies the pathway a patient may take as a result of receiving the test of interest, identifying all other tests, medical decisions and treatment consequences as a result of the test application. Interviewees who were more likely to explicitly develop such a pathway tended to be health economists involved in modelling as part of the guideline development process. There was agreement among interviewees that incorporation of such a pathway into the guideline development process should have a more central role.

> Consideration of the clinical pathway and how tests fit in that pathway, that should have a more central role in the whole process. (ID 9)

> Defining clinical pathway helps to define diagnostic questions to be addressed. It helps clinicians make the link between test accuracy and clinical outcomes. That's usually when it kind of ticks off in their brain and helps them see the bigger picture. (ID 5)

It was however acknowledged that the creation of such pathways is a challenge to the guideline process. Although education and training of guideline panel members was mentioned as a potential mitigating factor, the complexity of the topic area, the presence of variation in testing and treatment process and lack of evidence often make it hard to reach consensus about a common pathway.

> Looking at a test in its context is useful, however some guidelines such as cancer are really broad and deal with the entire pathway from diagnosis to staging to treatment and follow up. For that reason, including a pathway is not doable due to resource constraints. (ID 7)

> Variation in practice can make it difficult to get consensus. Variation of practice sort of across the country because it's sometimes amazing how much variation there is and how people don't necessarily agree on each different pathway that is proposed. (ID 13)

Interviewees cited that education has an important contribution in helping the guideline panel formulate key questions that are informative, focused and which incorporate patient outcome(s). The limited time and budget within which guidelines need to be completed were cited as a barrier to the provision of education.

> The panel still needs help with question formulation. There is a lack of appreciation of what's required in a question. They found it quite hard to give us the details of what it actually means to do a PICO because a lot of them don't know that, you know. You can call it education of the clinicians. (ID 13)

> There is a need to be more specific in the guideline question as time and money is often short. (ID 3)

> PICO is based on the fact that they just want a good instrument and attention is not paid to other outcomes…my experience is that they are also a bit disappointed that there is too little time left to answer the question what it (the test) means for me as a clinician and what it (the test) means for the patient? (ID 3)

The search process

The search process for test accuracy evidence was expressed as being particularly resource intensive mostly due to methodological limitations; the lack of search filters to identify studies concerned with test accuracy. This makes the search process far more labour and time intensive compared to that for interventions.

Types of evidence and outcomes included in the guideline

All 17 interviewees shared the view that inclusion of patient important outcomes needs to be more explicit

than it currently is. Despite this, test accuracy, defined in this study as agreement in classification between the test results and the outcome of the clinical reference standard, was stated to be the dominant outcome of choice in their guideline development.

The main reason given for this was the lack of direct evidence that links testing to patient outcomes. When probed as to why direct evidence on patient outcomes tended to be scarce, interviewees felt this was a result of methodological, resource and awareness/educational issues among the panel. Interviewees felt there was a lack of awareness for the need for such evidence among funders and regulatory bodies in charge of approving tests that gain entry into the market. This in turn translates into difficulties in obtaining funding for studies that link testing to patient outcomes (see online supplementary tables S2–S4).

The influence of industry on the regulatory environment was also mentioned as having a role in hindering the conduct of studies linking testing to patient outcomes, which was felt to be in contrast to the funding and approval of drugs. Some interviewees felt there should be no distinction between the regulatory requirements for drugs and medical tests. The methodological challenges of conducting studies that link testing to patient outcomes (eg, large sample sizes and long study time lines) and 'technological obsolesce' (ie, tests that can become obsolete during study due to rapid technological development) were factors cited as inhibiting their conduct, particularly for smaller medical test companies.

Regarding the inclusion of other outcomes such as costs, interviewees cited lack of available resources as the main reason why this tended to not be included.

Synthesising and appraising the evidence

When synthesising and appraising test accuracy evidence, methodological, resource and educational challenges were raised. Methodologically, the large amount of study heterogeneity between individual studies limits the ability for secondary synthesis of the data. While it was acknowledged that the quality of test accuracy studies has improved since the arrival of guidelines such as Standards for Reporting of Diagnostic Accuracy (STARD),[10 18] overall studies are still more often than not of poor quality. The complexity of methods for performing diagnostic test accuracy systematic reviews and meta-analysis means guideline developers need education and training in these methods, which was a view expressed by interviewees.

> There's not much secondary synthesis you can do if you don't have good information from individual studies. (ID 2)

> Yes, we have a lot of them that show very heterogeneous data. That indicates there are a lot of things still to be done and that we should be very cautious of single studies and drawing conclusions. (ID 10)

> We don't have enough outside resources (training resources) to have a fully-fledged systematic review done properly. (ID 9)

> We do not do meta-analysis because we are not as familiar on how to do this compared to treatment. (ID 3)

From evidence to making recommendations

Interviewees mentioned that expert opinion and consensus discussion are the most frequent ways by which recommendations are derived. These are considered essential aspects of the recommendation making stage, particularly when evidence on patient outcomes is lacking and when test accuracy data are of poor quality. Yet, these factors also can make the process of making recommendations unstructured, lacking transparency and at risk of being biased by political and personal opinions.

> We're so often limited by the quality of the evidence available, and so we are taking the clinical expertise of a group of highly expert individuals as the next best thing. (ID 6)

> Now all systematic work is focused on the evidence, on the hard evidence and the rest is… not very systematic. We don't have the tools or a format for that process. (ID 3)

> It's political and then lack of evidence that's not there. You have to build on the expert opinion and that can be quite difficult. (ID 10)

Views about the role of modelling in the development of recommendations were varied. Some interviewees viewed modelling as an alternative and not the only way to making recommendations. One interviewee felt modelling was the only evidence-based way recommendations could be made in the face of lack of studies that link the impact of a test to downstream patient outcomes. In contrast, another interviewee felt having direct evidence on the effect of testing on patient outcome is the only way forward and modelling was too misleading. Whatever the view held, it was clear modelling brought with it resource considerations in terms of time, expertise and educational needs.

AWARENESS, EDUCATION AND TRAINING ARE CENTRAL TO ADDRESSING CHALLENGES

The need for increased awareness, education and training within the guideline panel and at a societal level emerged as a central theme across all 17 interviews with differing consequences (table 2, see online supplementary tables S1 and S4).

Within the guideline panel

There was consensus among interviewees that more time needed to be spent educating the guideline panel about the concept of test accuracy, test accuracy statistics and how test accuracy can impact a range of patient

outcomes prior to the start of guideline development. Methodologists on the guideline panel often struggle to get guidance from other panel members about key questions in the guideline development process.

> Normally the guideline development group are quite familiar with forest plots, per se. It doesn't matter if it's intervention or diagnostics. But anything slightly advanced, we would have to engage into it a bit more. We'd have to run a session of training. (ID 2)

> Educating the panel that actually they have to help me work out what it is that's the most important here and on that basis I can sort of guide them through the results better. (ID 13)

> The panel finds it very hard to make a choice as to when high sensitivity is important and when is high specificity important. They do not understand the consequences of high sensitivity and low specificity or vice versa and hence cannot guide the methodologist either on what are important characteristics for the tests. (ID 3)

It was noted that the extent of training needed depended on the medical specialty of the panelists. Some members need more assistance in understanding test accuracy and its implications than other medical specialists.

> Radiologists have quite a good basic knowledge about diagnostics but discuss this with, for example, a gastroenterologist or an internal specialist, it's more difficult. I think it's their background and the fact that they read a lot of evidence on the tests that they are using so they are quite familiar with the description of the evidence I guess. In contrast to the gastroenterologist, who is reading about how to treat a person and how a certain therapy should be used, yes or no. It's another kind of statistics and focus. (ID 7)

External to the guideline panel

Medical education of doctors

There was agreement among interviewees that the lack of understanding of medical tests is a reflection of the inadequacy of medical education with respect to interpretation of test accuracy statistics and an understanding of the implications of test results from a patient outcomes perspective.

> People do not understand, really understand, diagnostic accuracy. We have to go through extensive training. We do significant training of our panel before they ever get anywhere near anything about this. It's not the way people have practiced medicine historically. At the moment it's something that people are taught in medical school and promptly forget, because they don't have a good way to try and actually use it in practice. (ID 1)

> They (clinicians) don't understand the fundamental principles of how to interpret lab tests. They really don't capture the essence of diagnostics because it's not properly taught in medical schools. (ID 9)

> Clinicians understanding of the downstream consequences is poor. However, once their awareness is raised on the impact a test can have downstream of its application, their view changes. (ID 5)

> If we ask for instance if a certain test is very sensitive for diagnosing myocardial damage, how would you use this test? I think most clinicians would answer that they could use it to make the diagnosis of myocardial infarction and the truth is just the opposite. You can use it for excluding myocardial infarction but you can't use it for making the diagnosis, unless it is also very specific. But this is something that I think the majority of clinicians don't know. Majority of clinicians have difficulties in understanding the meaning of sensitivity and specificity. Training would be helpful because many clinicians did not learn this in medical school. (ID 4)

Inadequacies in medical education about tests were felt to add to the complexity of developing guidelines for medical tests compared to guideline development for interventions. In addition to inadequacies in medical education, the disproportionately greater focus on intervention research, particularly in the form of randomised controlled trials (RCTs), was perceived to further compound the lack of familiarity in understanding medical test evaluation and its impact on patient outcomes.

> Statistical expressions of the diagnostic performance of assays are not terribly appealing to clinicians and it's hard for even guideline developer clinicians to capture those details. That's why they need a bit more education in this area, to translate it into a language that is easily understood by them. Clinicians are a bit more used to big, mega, pharmaceutical trials and how to interpret that information. (ID 9)

> I think with RCTs (randomized controlled trials), they tend to be very familiar with them. Maybe because it's just sort of been what they've been exposed to the most I think. There's this perception that RCT is the best design, so sometimes they might even ask for RCTs for questions where it's not actually appropriate to do an RCT. (ID 13)

> Most attention goes to intervention studies in journals and in guidelines normally. Generally in the education of medical professionals there is less focus on diagnostic accuracy. They're not used to it. (ID 3)

Raising awareness among stakeholders beyond the guideline panel

Funders of primary research on tests, such as the medical test industry and regulatory organisations approving these tests, were identified as essential players beyond the guideline panel whose awareness needs to be raised on the importance of knowing a test's impact on patient outcomes. Raising their awareness would help facilitate getting access to funding of end-to-end studies that report on patient outcomes.

More primary studies assessing a diagnostic test's impact on patient outcomes would be helpful in this regard although these types of studies may not always be feasible for reasons of lack of awareness on the need to assess a test's impact on patient outcomes. (ID 7)

Randomised controlled trials that evaluate full strategies including tests and treatment—we need new funding mechanisms before we get them. (ID 4)

If we change the regulatory process which should be similar to drugs, then I feel that we will in a very few years we'll have much better studies. (ID 14)

DISCUSSION

Summary of results

In this study, guideline developers acknowledge that they need to explicitly incorporate patient outcomes in medical test guideline development, but they face challenges when trying to do so. The challenges they face could be classified into methodological issues, resource limitations, awareness and educational needs.

Limitations of our study

Our interviewees were mostly used to serving as clinician methodologists on a guideline panel. Hence, their views might be biased from a dominantly methodological point of view. Patient or other health professional guideline panelists might have different views on the challenges experienced in guideline development for medical tests that the findings of our study may not include. Despite this limitation, the views expressed in our study are in line with other research that call for better regulatory frameworks for medical test approval and corroborate our findings that health professionals in general find it difficult to understand test accuracy measures and relate that to downstream patient consequences.[19–22] Our sample of interviewees play an intricate role, as methodologists, in assisting all panelists with issues they may encounter in the guideline process, hence they have a good understanding of the challenges faced by different types of panelists.

In this study as in other qualitative studies, we did not strive for statistical representation but rather maximum variation in our sampling so as to adequately illustrate the range of experiences and perceptions relevant to our topic. This, we accomplished by striving for a selection of guideline developers representing a range of different clinical areas and levels of medical test guideline development expertise. This does not negate the generalisability of our findings in the broader context as explained by Morse on the generalisability of qualitative research.[23]

Our study identified a number of methodological solutions that could, in the short term, better facilitate the development of patient outcome centred guidelines. These include the inclusion of test-treatment pathways as an explicit step in guideline development, including different methods such as qualitative approaches to search for and/or collect patient outcome data, using modelling to bridge the gap between test accuracy and patient outcomes. In addition, employing a structured process and format when moving from test accuracy evidence to making recommendations such as those in GRADE for Diagnostics[4 5] would help the decision-making processes more transparent and less prone to bias.

In recent years, organisations such as the AHRQ, NICE, the Cochrane Diagnostic Test Accuracy Working Group and the GRADE Working Group have made considerable progress in introducing different strategies that focus on the inclusion of patient outcomes in medical tests guideline development.[4–6 24] The continued adoption of these approaches by national[7] and international guideline organisations is a step in the right direction. However, despite these developments, the inclusion of test treat pathways is not commonplace in medical test guideline development.[25] While NICE advocates the construction of a test-treatment pathway as part of guideline development and AHRQ makes explicit mention of the need to develop an analytical framework in its medical test manual,[6 26] there is no explicit methodology on how such a pathway can be created as part of these organisations' processes. This was corroborated by our own findings in this study which demonstrated much variation in how and when these pathways are created and used and that they are not a standard step with an explicit approach.

None of our interviewees from smaller, local guideline development groups explicitly developed such pathways as part of their guideline process. This suggests that the resource implications in terms of time, money and education that is needed to implement these solutions cannot be ignored and may be a barrier.

Interviewees felt very strongly that there still exist a number of differences between the guideline process for drugs versus medical tests (see online supplementary table S5). Central to this was the view that clinicians were more familiar and had more confidence in intervention research. Several factors were seen to contribute to this position: the presence of clear and stringent regulatory requirements, the investment and commitment of industry resulting in patient outcome-based research and relatively greater exposure of intervention research in the medical literature and in medical education of doctors.

A recognition of the importance for a medical test to demonstrate patient benefit at the regulatory level is a concrete way to help overcome the resource limitations guideline developers face at various levels: from more resources for incorporating different methodological approaches, and training of guideline development panels, to exerting pressure on the medical test industry to include patient outcomes in their test development processes.

While initiatives such as STARD and GRADE for Diagnostics[4 18] are clear indicators of the growing awareness in the medical test research community on the need for improvement in the field of medical testing and evidence appraisal, there is certainly still much room for improvement. The call for a change in the way medical tests are taught in undergraduate and postgraduate education is probably a more immediately achievable goal in the near future. Clinicians and healthcare providers need to be trained not just in understanding conventional test accuracy statistics but to also be able to translate these statistics into a meaningful clinical perspective of what it means to a patient receiving the test. Research has demonstrated that clinicians find it challenging to understand and interpret medical test statistics[19–21] let alone translate this into downstream patient impact. Nevertheless, the importance of tests' impact on patient outcomes has been well recognised —the most visible being screening tests.[27] Introducing this change in medical education will lead to future guideline panels being able to bring this understanding and perspective from the outset of the guideline development process. It can also help improve the conduct and quality of studies of test performance, which are still an essential though not exclusive piece of evidence needed in the guideline development process. Good quality test accuracy studies can form an important foundation on which medical test recommendations may be based on.

In order to have a longer term impact on the challenges identified in our study, a paradigm shift is required, not just within the medical test research community but at a societal level. We should start asking different questions —as a patient receiving a test, a clinician about to purchase or administer a test to regulators approving tests or as funders, including the medical test industry creating new tests. A shift in the way we view the value of a test is required: to move away from solely considering how accurate a test may be in diagnosing a condition to including the value it may bring to the patient receiving the test.

Recommending, ordering or reimbursing medical tests should not only be guided by the information that is generated by these tests, but by the effects they have on patient-relevant outcomes and costs. Developing guidelines that reflect this basic principle is not yet straightforward but is absolutely necessary if we want to safeguard and improve patients' health.

Author affiliations
[1]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
[2]Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK
[3]Iberoamerican Cochrane Centre, Biomedical Research Institute Sant Pau-CIBER of Epidemiology and Public Health (CIBERESP-IIB Sant Pau), Barcelona, Spain
[4]School of Public Health, Sydney Medical School, Public Health Section, Centre for Medical Psychology and Evidence based Decision Making (CeMPED), Sydney, New South Wales, Australia

Contributors PB initiated the original research idea. GG developed the interview topic guide, and data analysis plan together with contributions from MWL, MMGL, PB, CD and KM. GG, CD and AJS conducted the interviews. GG, MWL and MMGL participated in all stages of the framework analysis with contributions from PB, CD and KM. GG drafted the manuscript. MWL, MMGL and PB participated in the definition of the data tables and figures. All authors contributed to and approved all versions of the manuscript.

Funding This study has been funded by the DECIDE project under the European Union Seventh Framework Programme (FP7/2007-2013) Grant number 258583.

Competing interests All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare that GG, MML, PB, and MWL had financial support for the submitted work from DECIDE Project which is funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 258583. All authors declare no financial relationships with any other organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethics approval According to Dutch Law, there is no legal requirement for this study to receive medical ethics committee approval.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data set containing the codes and subcodes derived from the interviews and anonymised transcripts can be made available to interested researchers.

REFERENCES

1. Shekelle PG, Woolf SH, Eccles M, *et al*. Clinical guidelines: developing guidelines. *BMJ* 1999;318:593–6.
2. Qaseem A, Forland F, Macbeth F, *et al*. Guidelines International Network: toward international standards for clinical practice guidelines. *Ann Intern Med* 2012;156:525–31.
3. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. *Med Decis Making* 2009;29:E30–8.
4. Brozek JL, Akl EA, Jaeschke R, *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;64:1109–16.
5. Schünemann HJ, Oxman AD, Brozek J, *et al*. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106–10.
6. Smetana GW, Umscheid CA, Chang S, *et al*. Methods guide for authors of systematic reviews of medical tests: a collaboration between the Agency for Healthcare Research and Quality (AHRQ) and the Journal of General Internal Medicine. *J Gen Intern Med* 2012;27(Suppl 1):S1–3.
7. Thornton J, Alderson P, Tan T, *et al*. Introducing GRADE across the NICE clinical guideline program. *J Clin Epidemiol* 2013;66:124–31.
8. Bossuyt PM. Room for improvement in national academy of clinical biochemistry laboratory medicine practice guidelines. *Clin Chem* 2012;58:1392–4.
9. Whiting P, Rutjes AW, Reitsma JB, *et al*. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189–202.
10. Leeflang MM, Deeks JJ, Gatsonis C, *et al*. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;149:889–97.
11. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol* 2005;5:20.
12. Bossuyt PM, Irwig L, Craig J, *et al*. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089–92.

13. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, *et al*. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012;344:e686.
14. Gale NK, Heath G, Cameron E, *et al*. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Med Res Methodol* 2013;13:117.
15. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Analyzing Qualitative Data. Bryman A, Burgess RG, eds. *Taylor & Francis Books Ltd*, 1994, pp 173–94.
16. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19: 349–57.
17. Medicine CEBM. Asking Focussed Questions University of Oxford [updated 2014]. http://www.cebm.net/index.aspx?o=1036
18. Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–4.
19. Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. *Syst Rev* 2013;2:32.
20. Whiting PF, Davenport C, Jameson C, *et al*. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open* 2015;5:e008155.
21. Ben-Shlomo Y, Collin SM, Quekett J, *et al*. Presentation of diagnostic information to doctors may change their interpretation and clinical management: a web-based randomised controlled trial. *PLoS ONE* 2015;10:e0128637.
22. Horvath AR, Lord SJ, StJohn A, *et al*. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta* 2014;427:49–57.
23. Morse JM. Qualitative generalizability. *Qual Health Res* 1999;9:5–6.
24. Deeks JJ, Wisniewski S, Davenport C. Chapter 4: Guide to the contents of a Cochrane Diagnostic Test Accuracy Protocol. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration, 2013. http://srdta.cochrane.org/
25. Gopalakrishna G, Langendam MW, Scholten RJ, *et al*. Guidelines for guideline developers: a systematic review of grading systems for medical tests. *Implement Sci* 2013;8:78.
26. NICE. Centre for Health Technology Evaluation. Diagnostics Assessment Programme (DAP). Programme manual. 2011. http://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-diagnostics-guidance/Diagnostics-assessment-programme-manual.pdf
27. Andermann A, Blancquaert I, Beauchamp S, *et al*. Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. *Bull World Health Organ* 2008;86:317–9. http://www.who.int/bulletin/volumes/86/4/07-050112/en/ (accessed Jul 2015).