Evaluation of prediction models and diagnostic tests

Beyond the standards

Kevin Jenniskens

Evaluation of prediction models and diagnostic tests
Beyond the standards

Kevin Jenniskens

Bekijk dit proefschrift in enkele minuten

Evaluation of prediction models and diagnostic tests: beyond the standards

PhD thesis with Dutch summary

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, The Netherlands

ISBN: 978-90-393-7211-1

Author: Kevin Jenniskens

Cover: David de Groot | https://www.persoonlijkproefschrift.nl/

Layout: Kevin Jenniskens

Printed by: Ipskamp Printing | https://www.ipskampprinting.nl/

Evaluation of prediction models and diagnostic tests: Beyond the standards

Evaluatie van predictie modellen en diagnostische testen: Verder dan de standaard

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

donderdag 21 november
des middags te 4:15 uur

door

Kevin Jenniskens

geboren op 31 juli 1988

te Venlo

Promotor:
Prof. dr. K.G.M. Moons

Copromotoren:
Dr. C.A. Naaktgeboren
Dr. L. Hooft

Remember: it's all luck

You are lucky to be here
You were incalculably lucky to be born,
And incredibly lucky to be brought up by a nice family that helped you get educated and encouraged you to go get a PhD.

Or if you were born into a horrible family, that's unlucky and you have my sympathy…

But you were still lucky:
Lucky that you happened to be made of the sort of DNA that made the sort of brain which – when placed in a horrible childhood environment – would make decisions that meant you ended up, eventually, obtaining a PhD.

Well done you, for dragging yourself up by the shoelaces, but you were lucky. You didn't create the bit of you that dragged you up.

They're not even your shoelaces

Tim Minchin, Australian comedian

Contents

"We're not here to change the world we're here to laugh at others. Maybe have a nap along the way"

Bowling for Soup, punk rock band

Chapter 1

General introduction

In a perfect world…

… individuals are only labelled as having a certain disease or health condition when it is truly present. Giving a diagnostic label to an individual can have extreme impact, both directly and indirectly, on the individual, the family and of course on subsequent medical decision making and care. Research on the value of diagnostic tests involves quantifying the extent to which a novel (index) test adds diagnostic information to current care. This is done by assessing the test's predictive performance or accuracy in combination with other test results (often referred to as a diagnostic model). This diagnostic performance is assessed in a diagnostic accuracy study by comparing the results of the index test or diagnostic model to the result of the so-called reference test (reference standard). The reference test is considered to provide a completely error-free diagnostic classification of all study participants into those with and those without the target condition of interest. A fully error free reference standard would allow for the correct estimation of the diagnostic accuracy of the index test(s) under study.

… the reference standard only labels individuals as having the target condition when they would benefit from receiving that diagnostic label. Benefit can be diverse, ranging from improved health, extended life expectancy, reduced healthcare costs, or even just reassurance. The reference standard defines what disease is, and who is and isn't affected. Some diseases might be symptomatic, whereas others might be asymptomatic. Some with the disease might be severely affected, whereas others less so. In a perfect world, the reference standard will only detect and label individuals as having a disease if that disease would ultimately indeed lead to morbidity or mortality.

… implementation of a novel, sufficiently accurate, diagnostic test or model in clinical practice, will improve effectiveness and/or efficiency of the current diagnostic pathway, and subsequently lead to better choices in patient management and care. The results of such tests or models are always correctly used to guide treatment decisions, maximizing net benefit in the

long run, improving quality of life, prolonging life-expectancy, or reducing healthcare costs as a result of optimal patient management.

In reality…

… the reference standard is often not 'gold', and does not provide perfect, error-free classification of the study participants with and without the target condition. However, when estimating diagnostic accuracy of an index test or model in a scientific study, it is assumed that it does. Several approaches have been proposed and evaluated to overcome imperfect reference standards when evaluating diagnostic tests or models, such as the use of composite reference standards or expert panels that use multiple component test results to come to a final diagnostic classification for each study participant. In both approaches a dichotomous classification of the target condition (present or absent) is forced upon all study patients, such that the diagnostic accuracy of the index test can be calculated in the traditional way. The drawback of forcing a reference standard to make such a black-and-white diagnostic classification, when it is known that the reference standard is imperfect, is that it ignores any remaining uncertainty, for example, due to that fact that component test results of a composite reference standard are conflicting, or experts in a panel disagree. This in turn leads to biased accuracy estimates of the diagnostic index test or model under study.

… not all patients with abnormalities that are labelled as disease present will have benefit from receiving a diagnosis. Disease presence versus absence is typically defined by deciding where to draw the line between what is "normal" and "abnormal". But definitions of disease, and thus of normality versus abnormality may change over time, for example due to an increased ability to detect smaller abnormalities, due improved understanding of associations between risk factors and long-term health outcomes, or due to lobbying of invested parties. These factors can thus influence the concept of disease ('abnormality') itself, which in turn may result in more individuals being labelled with the disease. The question arises then whether these individuals that are now suddenly classified as diseased, would indeed experience symptoms and health deterioration if left untreated. Patients

diagnosed with a condition which would not lead to net benefit in terms of health outcomes, are classified as "overdiagnosed". This and many other related terms, such as overscreening, overtesting, overtreatment, can be grouped under the broad umbrella term of "Too much medicine". Unfortunately, all these terms are still loosely defined and have different meaning across clinical fields leading to a myriad of definitions and much linguistic confusion. Conceptual and methodological frameworks have been proposed to give guidance on estimating and reporting the amount of overdiagnosis, and to describe the intricate workings and interrelatedness of terminology surrounding "Too much medicine". Although these are valuable frameworks, they are often limited to a specific clinical field (e.g. screening in oncology), and tend not to provide guidance on how to reduce overdiagnosis and its consequences.

… good accuracy or performance of a diagnostic test or model does not necessarily guarantee that diagnostic tests or models are (correctly) used to inform patient management decision making, leading to a positive impact on downstream patient health outcomes and/or healthcare costs. Issues such as lack of compliance by healthcare providers or patients to follow management recommendations based on certain test or model results, limited effectiveness of treatment, or unexpected complications after using tests or treatments, may result in limited impact on short- and long-term patient health outcomes. So-called randomized 'test/model-treatment trials' are the most rigorous research approach to assess the impact of diagnostic tests or models. However these are often complex in setting up, require significant time investments, and tend to be costly. Moreover, the results of such test/model-treatment trials have frequently been disappointing, as the observed impact of the index test or model was lower or even absent compared to what was expected. Alternative approaches for assessing impact of a new diagnostic test or model, and providing guidance for consistent reporting, could prove valuable to prevent (diagnostic) research waste.

Aim and outline of this thesis

The aim of this thesis is to expose methodological issues surrounding evaluation of diagnostic tests and models, and their associated impact on patient health outcomes and healthcare related costs, as well as to propose alternative approaches to reduce bias, research waste and improve methodological quality.

Chapter 2 highlights an overlooked problem that may arise when using an expert panel as a reference standard in diagnostic studies. An expert panel typically provide a dichotomous target condition classification for each study participant, thereby ignoring any uncertainty that may (still) exist, which may result in bias in the accuracy estimates of the index test. Through a step by step illustrative example we first demonstrate how forcing a black-and-white classification by the expert panel can introduce bias. Next a series of simulations is performed, in which the number and accuracy of component tests that is available to the expert panel, as well as the target condition prevalence are varied, to assess the amount of bias when estimating the diagnostic accuracy of an index test.

In **chapter 3** an alternative method to dichotomous target condition classification is proposed, using probabilistic estimates of target condition presence from the expert panel. These estimates were elicited from expert panel members in the SPACE study, a study aimed at evaluating the accuracy of clinical prediction rules for detecting sepsis in the emergency room. We use these empirical study data to demonstrate how probabilistic estimates of target condition presence elicited from an expert panel can be used to derive diagnostic accuracy measures of the index test under study. The results from the probabilistic approaches are compared to those of the traditional approach, in which the panel is forced to provide a dichotomous (present or absent) target condition classification for each patient.

Chapter 4 provides a scoping review on the topic of overdiagnosis and related concepts in the field of "Too much medicine". It addresses the clinical fields in which the problems are described and in what context they

are discussed. The paper concludes with the topics for which there is not yet consensus and where more research may be warranted.

Subsequently, **chapter 5** addresses one of the topics identified in the scoping review, namely the need for a uniform framework describing "Too much medicine" related concepts across different clinical fields. Mechanisms leading to too much medicine are described at various stages of the clinical pathway, using examples from clinical practice. Ultimately the reader is provided with strategies to reduce too much medicine linked to the mechanism applicable to their specific clinical problem.

Chapter 6 evaluates the use and feasibility of performing a decision analytic modelling approach before conducting a test/model-treatment trial to assess the impact of a diagnostic prediction model on patient health outcomes and healthcare costs. We illustrate this approach using the HEART score for predicting cardiovascular events in the emergency room as a case study example. The feasibility of such decision analytic modelling approach to assess the impact of diagnostic tests and models is discussed, and guidance is given on the various steps within such an approach.

Chapter 7 illustrates the value of early health economic modelling at the stage where one is (considering to) develop a novel diagnostic test. This approach will be illustrated by looking at the potential cost-effectiveness of a novel biomarker for diagnosis of primary aldosteronism in hypertensive patients, by using a so-called headroom approach.

In **chapter 8** a systematic review is performed, looking at the methods used to determine the pretest probabilities (prevalence of disease) to facilitate the interpretation of diagnostic accuracy parameters in the summary of findings tables of Cochrane reviews on diagnostic tests. The pretest probabilities chosen in these tables directly affect the absolute number of (true and false) positive and (true and false) negative individuals in a hypothetical cohort. These numbers play a key role when judging the clinical usefulness of a diagnostic test, and any shortcomings in selecting the pretest probabilities may potentially misinform readers of systematic reviews of diagnostic tests.

We provide an overview of the various methods used for selecting pretest probabilities and discuss their validity.

Chapter 9 discusses the main findings from this thesis, and highlights main areas for future research.

"A famous bon mot asserts that opinions are like arse-holes: everyone has one

There is great wisdom in this… but I would add that opinions differ significantly from arse-holes, in that yours should be constantly and thoroughly examined"

Tim Minchin, Australian comedian

Chapter 2

Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: A simulation study

K. Jenniskens

C.A. Naaktgeboren

J.B. Reitsma

L. Hooft

K.G.M. Moons

M. van Smeden

J Clin Epidemiol. 2019;111:1-10

Abstract

Objective: To study the impact of ignoring uncertainty by forcing dichotomous classification (presence or absence) of the target disease on estimates of diagnostic accuracy of an index test.

Study Design and Setting: We evaluated the bias in estimated index test accuracy when forcing an expert panel to make a dichotomous target disease classification for each individual. Data for various scenarios with expert panels were simulated by varying the number and accuracy of "component reference tests" available to the expert panel, index test sensitivity and specificity, and target disease prevalence.

Results: Index test accuracy estimates are likely to be biased when there is uncertainty surrounding the presence or absence of the target disease. Direction and amount of bias depend on the number and accuracy of component reference tests, target disease prevalence and the true values of index test sensitivity and specificity.

Conclusion: In this simulation, forcing expert panels to make a dichotomous decision on target disease classification in the presence of uncertainty, leads to biased estimates of index test accuracy. Empirical studies are needed to demonstrate whether this bias can be reduced by assigning a probability of target disease presence for each individual, or using advanced statistical methods to account for uncertainty in target disease classification.

Introduction

In diagnostic test accuracy studies, the discriminatory ability of the test of interest (index test) is evaluated by comparing its results to those of the reference standard in a group of individuals suspected of the target disease. While analysing this comparison, it is often assumed that the reference standard can perfectly distinguish two groups of individuals: those with and without the target disease. (1, 2) For many diseases, however, the best available reference standard isn't perfect. (3, 4) In the absence of a single perfect test that can be held as the reference standard, alternative approaches have been proposed, including composite reference standards (applying multiple tests and combining their results using a fixed rule), latent class models (using a statistical method to link multiple tests results to a latent class) and expert panels. (5)

Using an expert panel is a common approach to assign a final diagnosis in fields where an accepted reference standard is lacking. (6) In such a panel, multiple experts combine information from multiple tests, patient characteristics, and clinical expertise to make a final decision on whether the target disease is present or absent for each individual. Typically, in an expert panel all individuals are ultimately classified as either having or not having the target disease based on a decision making procedure, such as majority vote or by consensus. (6, 7) With this dichotomization (presence or absence) of the target disease, measures of index test accuracy can be calculated in the traditional way. (8, 9)

Compared to a single test error-prone reference standard, the panel diagnosis may improve reference standard accuracy and subsequently reduce reference standard bias (5, 10). However, panel diagnoses almost by definition lead to imperfect target disease classification, as evidenced by studies of panel intra- and interobserver variability. (11-13) Different experts within a panel can disagree on the presence/absence of the target disease, in particular in patients presenting with atypical signs and symptoms. Forcing a dichotomous decision in every individual thus ignores this uncertainty about the target disease status. Simply ignoring this uncertainty may lead to biased accuracy estimates of the index test under study. This has already

been demonstrated for composite reference standards using explicit decision rules (e.g. at least two out of four tests should be positive to assign a target disease present classification to an individual) (14, 15), but has not been described in the context of expert panels, which is the goal of this paper.

In this study we aim to assess the impact of dichotomization of the target disease classification on accuracy estimates of the index test. An expert panel with multiple imperfect tests at its disposal will be used as a reference test. We first present an example to illustrate how ignoring uncertainty in the target disease classification can lead to biased accuracy estimates of the index test. Readers familiar with this type of bias can skip this section ("The source of bias: an illustrative example") and directly go to the description of the methods and results of our simulation study, illustrating the bias due to dichotomization of target disease status across a range of scenarios. Implications of the results for diagnostic research will be discussed and alternative strategies for reducing bias in index test accuracy estimates will be proposed.

The source of bias: an illustrative example

Consider the following hypothetical example of 1,000 individuals with a target disease prevalence of 40% to which an index test with sensitivity and specificity 80% is applied. Assuming we have a perfect reference standard, we can construct Table 1, which we will refer to as the true contingency table.

Table 1. True contingency table for a hypothetical index test when compared to a gold standard.

	Disease present (according to gold standard)	Disease absent (according to gold standard)	
Index test +	320	120	**440**
Index test -	80	480	**560**
	400	**600**	**1000**

Now suppose that there is no perfect reference standard, and that the disease classification is made by a panel of experts. These expert panels are frequently applied in various clinical domains such as psychiatric disorders and cardiovascular diseases when a single reference standard is lacking. (6) For example, when assessing screening tools for diagnosis of autism spectrum disorder, expert panels are used as a reference standard, which requires (subjective) interpretation of different components from the Diagnostic and Statistical Manual of Mental Disorders (DSM). (16, 17) Note that in this paper we will not consider a continuous spectrum of target disease severity, but rather focus on expert panel's uncertainty regarding the presence or absence of a single well-defined target condition.

Expert panels combine the results of several imperfect tests to make the final classification whether the target disease is present or absent. Each separate test available to the expert panel will hereinafter be referred to as a "component test". We use the term component test in a broad sense, as any piece of information (e.g. patient characteristic, biomarker, imaging) that might help in making the disease classification. In this example we use two dichotomous component tests, the first having a true sensitivity and

specificity of 80%, and the second a sensitivity and specificity of 90%. For simplicity, we assume that the errors of these component tests are uncorrelated, in other words the test results are conditional independent given the true target disease status. (18)

We simulate the implicit decisions by an expert panel on target disease classification by making them explicit, solely based on the assigned probability of target disease presence given the component test results for any given individual. Individuals are then classified to target disease present (i.e. probability of disease presence of 50% or higher) or target disease absent (i.e. probability of disease presence below 50%). We assume that the panel is well calibrated (they assign correct target disease presence probabilities) and consistent (they apply the same threshold value of 50% across all individuals when dichotomizing). Ultimately the panel is forced to classify each individual as either being disease present or disease absent. This final classification is used to calculate sensitivity and specificity of the index test. Because this is a simulation study, we know the true values of sensitivity and specificity of the index test, and therefore the corresponding bias can be calculated.

In this example, there are four possible component reference test patterns (++, +-, -+, --). The probability of observing a specific test pattern is given by the sensitivity (Se) and specificity (Sp) of the component tests, and the target disease prevalence (prev). When a ++ pattern is observed, the first and second component test are positive. This can occur in two ways: an individual has the disease and these are two true positive component test results (with probability: $prev*Se_{comp1}*Se_{comp2}$) or an individual does not have the disease and these are two false positive component test results (with probability: $(1-prev)*(1-Sp_{comp1})*(1-Sp_{comp2})$). The total probability of observing pattern ++ is the sum of these two probabilities. This can be generalized for each possible component test pattern, obtaining the following formulas for the probability of each pattern:

Formulas for the probability for observing each possible component test pattern

Pattern (k)		*Probability for diseased cases*		*Probability for non-diseased cases*
1	+ +	$prev*Se_{comp1}*Se_{comp2}$	+	$(1-prev)*(1-Sp_{comp1})*(1-Sp_{comp2})$
2	- +	$prev*(1-Se_{comp1})*Se_{comp2}$	+	$(1-prev)*Sp_{comp1}*(1-Sp_{comp2})$
3	+-	$prev*Se_{comp1}*(1-Se_{comp2})$	+	$(1-prev)*(1-Sp_{comp1})*Sp_{comp2}$
4	- -	$prev*(1-Se_{comp1})*(1-Se_{comp2})$	+	$(1-prev)*Sp_{comp1}*Sp_{comp2}$

The probability of target disease presence given a component test pattern can be derived by using Bayes Theorem. (19, 20) For pattern ++, this is post-test probability is given by the probability of observing that pattern among diseased (probability: $prev*Se_{comp1}*Se_{comp2}$) divided by the total probability of getting that pattern (probability: $prev*Se_{comp1}*Se_{comp2}$ + $(1-prev)*(1-Sp_{comp1})*(1-Sp_{comp2})$). Applying this line of reasoning to all component test patterns leads to the following formulas:

Formulas for the probability of disease presence within a component test pattern

Pattern (k)		*Probability for diseased cases*		*Probability for non-diseased cases*
1	+ +	$prev*Se_{comp1}*Se_{comp2}$	/	$(prev*Se_{comp1}*Sec_{omp2}+ (1-prev)*(1-Sp_{comp1})*(1-Sp_{comp2}))$
2	- +	$prev*(1-Se_{comp1})*Se_{comp2}$	/	$(prev*(1-Se_{comp1})*Se_{comp2}+ (1-prev)*Sp_{comp1}*(1-Sp_{comp2}))$
3	+-	$prev*Se_{comp1}*(1-Se_{comp2})$	/	$(prev*Se_{comp1}*(1-Se_{comp2})+ (1-prev)*(1-Sp_{comp1})*Sp_{comp2})$
4	- -	$prev*(1-Se_{comp1})*(1-Se_{comp2})$	/	$(prev*(1-Se_{comp1})*(1-Se_{comp2})+ (1-prev)*Sp_{comp1}*Sp_{comp2})$

For our illustrative example, in Table 2 we can see that the probability of observing a component test pattern with two positive test results is 30%, and within that component test pattern there is 96% (not 100%) probability of truly having the target disease. Hence in a sample of 1,000 individuals, the expert panel will assign the target disease to all 300 individuals having the ++ pattern, of which only 288 would truly have the target disease.

Table 2. Distribution pattern of two component tests, mapped on a theoretical sample. The table shows how a sample with a known target disease prevalence, is classified by the expert panel. Dichotomous class (DC) is assigned using a threshold for target disease presence probability of 50%.

Test pattern	Prob. test pattern	Prob. target disease presence	Dicho-tomous class (DC)	Total[1] n=1,000	Truly disease present (D1)	Truly disease absent (D0)
+ +	0.30	0.960	1	300	288	12
- +	0.12	0.600	1	120	72	48
+ -	0.14	0.229	0	140	32	108
- -	0.44	0.018	0	440	8	432

[1]Although individuals are classified as diseased (DC1) or non-diseased (DC0), note that not all of them are.

In practice, the expert panel makes a dichotomous decision about the presence of the target disease for each individual based on the results of the two component reference tests. To reach this decision, the expert panel applies a threshold (either implicitly or explicitly) on the probability of target disease being present. An intuitive threshold for dichotomizing disease status would be 50%, such that each individual is classified to the most likely disease status (present or absent). In our example this would result in all individuals with component test pattern ++ and -+ being classified as disease present, as their probabilities of having the disease are higher than 50% (96% and 60% respectively). The remaining component test patterns have probabilities below the 50% threshold, consequently individuals with these patterns will be classified as disease absent. We will refer to component test patterns above the threshold as dichotomous classification 1 (DC1) and under the threshold as dichotomous classification 0 (DC0).

In our illustrative example, the expected distribution of the 1,000 individuals can be calculated using the probability of observing a component test pattern and the probability of target disease presence. (Table 2) In DC1 there are two test patterns (++ and -+) in which 420 (300+120) individuals are classified as target disease present, and of which 360 (288+72) are truly diseased. In the test patterns in DC0 (+- and --) zero individuals are

classified as diseased, whereas in reality 40 (32+8) are truly diseased. From this we can also derive that the prevalence according to the expert panel's classification has changed from 40% to 42%. Note that within DC1, only 85% truly has the disease present, and in DC0 93% is truly non-diseased, hence a 15% overestimation of the number of diseased individuals in DC1, and 7% overestimation of non-diseased individuals in DC0.

Now we can construct the contingency table that we would expect to obtain for our index test when compared to the expert panel's target disease classification. DC1 consists of component test patterns ++ and -+, of which we know that 360 individuals are diseased (D1) and that the remaining 60 are non-diseased (D0). For simplicity, we will denote these diseased individuals by N(D1, DC1) and non-diseased individuals by N(D0, DC1). Given a positive index test and its sensitivity (Se_{index}) and specificity (Sp_{index}), we can calculate the number of true positives Se_{index} * N(D1, DC1) and number of false positives (1 - Sp_{index}) * N(D0, DC1) in DC1. This can also be done given a negative index test, and for DC0, using the following formulas:

Index test + | DC1 = Se_{index} * N(D1, DC1) + (1 - Sp_{index}) * N(D0, DC1)

Index test - | DC1 = Sp_{index} * N(D0, DC1) + (1 - Se_{index}) * N(D1, DC1)

Index test + | DC0 = Se_{index} * N(D1, DC0) + (1 - Sp_{index}) * N(D0, DC0)

Index test - | DC0 = Sp_{index} * N(D0, DC0) + (1 - Se_{index}) * N(D1, DC0)

The resulting contingency table, which we refer to as the observed contingency table, can be found in Table 3. This table shows the shift between true disease status by a perfect (gold) reference standard and observed disease status after disease classification by the expert panel (marked by the grey arrows). The misclassification of true diseased and non-diseased individuals resulting in the observed classification is indicated by the red arrows. Consider the cell with a positive index test and disease present according to classification. In the true contingency table, this consisted of 320 individuals, while in the observed contingency table, this number has dropped to 300.

Table 3. An illustrative example showing the differences in contingency tables and accuracy when comparing an index test to a gold standard (true) versus an index test to an expert panel using two imperfect tests (observed). The shift between true and observed disease classification is given by the grey arrows. The shift between disease present and disease absent, resulting in the observed classification, is indicated by the red arrows. Note that out of the 420 individuals classified as disease present by the expert panel, only 360 actually have the disease, and out of the 580 individuals classified as disease absent by the expert panel, only 540 actually do not have the disease.

	Disease Classification method	Disease present according to classification		Disease absent according to classification	
Index test +	Gold standard	320	32 / 12	120	440
	Expert panel	300		140	
Index test -	Gold standard	80	8 / 48	480	560
	Expert panel	120		440	
	Gold standard	400	40 / 60	600	1000
	Expert panel	420		580	
		Sensitivity		**Specificity**	
Gold standard (true value)		320 / 400 = 80.0%		480 / 600 = 80.0%	
Expert panel (observed value)		300 / 420 = 71.4%		440 / 580 = 75.9%	
Bias in estimates		80.0% - 71.4% = 8.6%		80.0% - 75.9% = 4.1%	

Values for sensitivity and specificity of the index test are calculated from the observed table, yielding 71.4% and 75.9% respectively, which, compared to the true values of 80%, correspond to a bias of 8.6% and 4.1% respectively. We can also derive the total proportion of misclassifications by adding the misclassifications in disease classification by the expert panel in both directions and dividing by the total number of individuals. For our example this yields (40+60) / 1000 = 10% misclassification.

Methods of the simulation study

We investigated a series of hypothetical scenarios to study the impact of dichotomous classification of the target disease on the bias in sensitivity and specificity estimates of an index test. Each scenario consists of an expert panel with multiple component tests at their disposal, with varying accuracy (all less than 100%). The calculations as described in the preceding illustrative example were used. For further background on the methods used, we refer to the Supplementary file.

Ten different expert panel scenarios were assessed in our simulation study, described in Table 4. In the base scenario the expert panel was provided with four component tests, each with a sensitivity and specificity of 70%, and the target disease prevalence was 20%. In other scenarios one of the following factors was varied: the number of component tests (two, four, and eight), the diagnostic accuracy of these component tests (60%, 70%, 80%, and a combination of high and low accuracy component tests), and target disease prevalence (10%, 20%, and 40%). Threshold for dichotomizing and assigning target disease status was kept constant at 50% across all scenarios (i.e. classification to the most likely target disease status). In all scenarios we assumed conditional independence between results of component tests.

Table 4. Description of expert panel scenarios, which include changing the number of component tests (a), the accuracy of component tests (b), and the target disease prevalence (c). A probability of 50% was chosen as a threshold for dichotomization of the target disease. The base scenarios are marked with an asterisk.

Scenario	# of tests	Component reference test sensitivity	Component reference test specificity	Target disease prevalence
Low number of tests[a]	2	70%	70%	20%
Medium number of tests[a]*	4	70%	70%	20%
High number of tests[a]	8	70%	70%	20%
Low accuracy[b]	4	60%	60%	20%
Medium accuracy[b]*	4	70%	70%	20%
High accuracy[b]	4	80%	80%	20%
Mirrored accuracy[b]	4	60%-70%-80%-90%	90%-80%-70%-60%	20%
Low prevalence[c]	4	70%	70%	10%
Medium prevalence[c]*	4	70%	70%	20%
High prevalence[c]	4	70%	70%	40%

Performance of the expert panel

Diagnostic performance of the expert panel was assessed by calculating the area under the receiver operator characteristic (AUROC) and the proportion of misclassifications. (21) AUROC is a measure for overall discriminative performance of the expert panel, that can be derived using probability of target disease presence as a continuous cut-off threshold. The proportion of misclassifications was calculated as the proportion of incorrect target disease classifications using the aforementioned threshold of 50% for dichotomization.

Bias in sensitivity and specificity estimates of the index test

We calculated the resulting bias in sensitivity and specificity estimates of the index test after dichotomous target disease classification by the reference standard for each of the scenarios. A comprehensive range (0 – 100%) of true index test sensitivity and specificity values was analysed to assess the amount and direction of bias in each scenario. Only either index test sensitivity or specificity was varied at a time. When varying index test

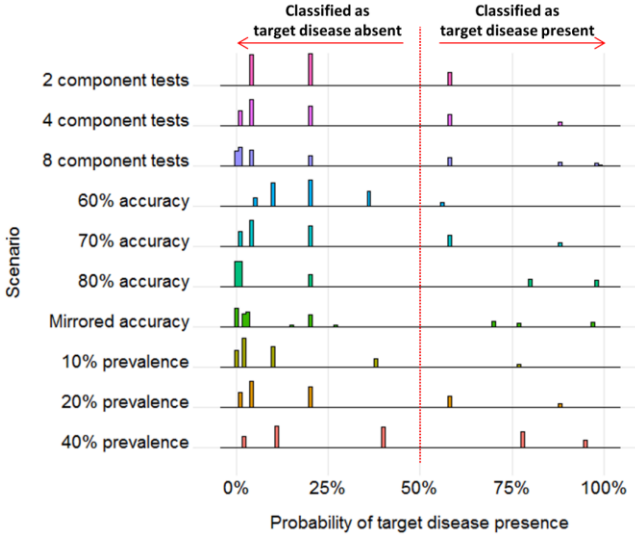
sensitivity, the specificity was kept constant at 80%. In a similar way, when specificity was varied, the sensitivity was fixed at 80%. Conditional independence between the index test and the component tests was assumed.

Results of the simulation study

Performance of the expert panel

The expected distribution of component reference test patterns and their corresponding probability of target disease presence are visualized in Figure 1. The bars visualize the expected relative frequencies of target disease probabilities corresponding to different component test patterns. The total number of these patterns possible for a given scenario is given by two to the power of the number of component tests. (i.e. $2^4 = 16$ patterns for the base scenario) One bar may contain more than one component test pattern when patterns have an equal probability of target disease presence. Target disease probability estimates towards the extremes (zero or one) are likely to yield the least incorrect classifications; almost all individuals in these patterns are likely to be either truly diseased or non-diseased, hence forced dichotomization of the expert panel will result in minimal incorrect classifications. However when there are patterns around the target disease dichotomization threshold (in this case 50%), and probability of observing these patterns is high, the likelihood of errors after dichotomization will increase.

Figure 1 Distribution of component reference test patterns and their associated probability of target disease presence, for each scenario. Proportion of misclassifications (at a threshold of 50% given by the red dotted line) and area under the receiver operator characteristic (AUROC) are given as measures of diagnostic performance. If multiple component test patterns have the same probability of disease presence, they are aggregated together in a bar. The base scenarios are marked with an asterisk.



AUROC	Proportion of misclassifications
0.784	0.174
0.874*	0.137*
0.950	0.086
0.710	0.195
0.874*	0.137*
0.967	0.058
0.940	0.086
0.874	0.083
0.874*	0.137*
0.874	0.190

As shown in the figure, when all component tests have identical accuracy many combinations of test patterns will have the same probability of the target disease being present. When comparing the scenarios with a low and high number of component tests, there was a higher likelihood of observing test patterns closer to the extremes for the latter, which resulted in higher discrimination (AUROC) and fewer disease misclassifications by the expert panel. Similar trends were observed when increasing the accuracy of the component tests. In the mirrored accuracy scenario, there was more spread in the probability of target disease presence for the various combinations of component test patterns, but overall provided similar discriminative performance (0.940 vs 0.967) and proportion of misclassifications (0.086 vs 0.058) compared to the high accuracy scenario. Changes to the target disease prevalence did not affect discriminative performance, however it did affect the expected number of misclassifications.

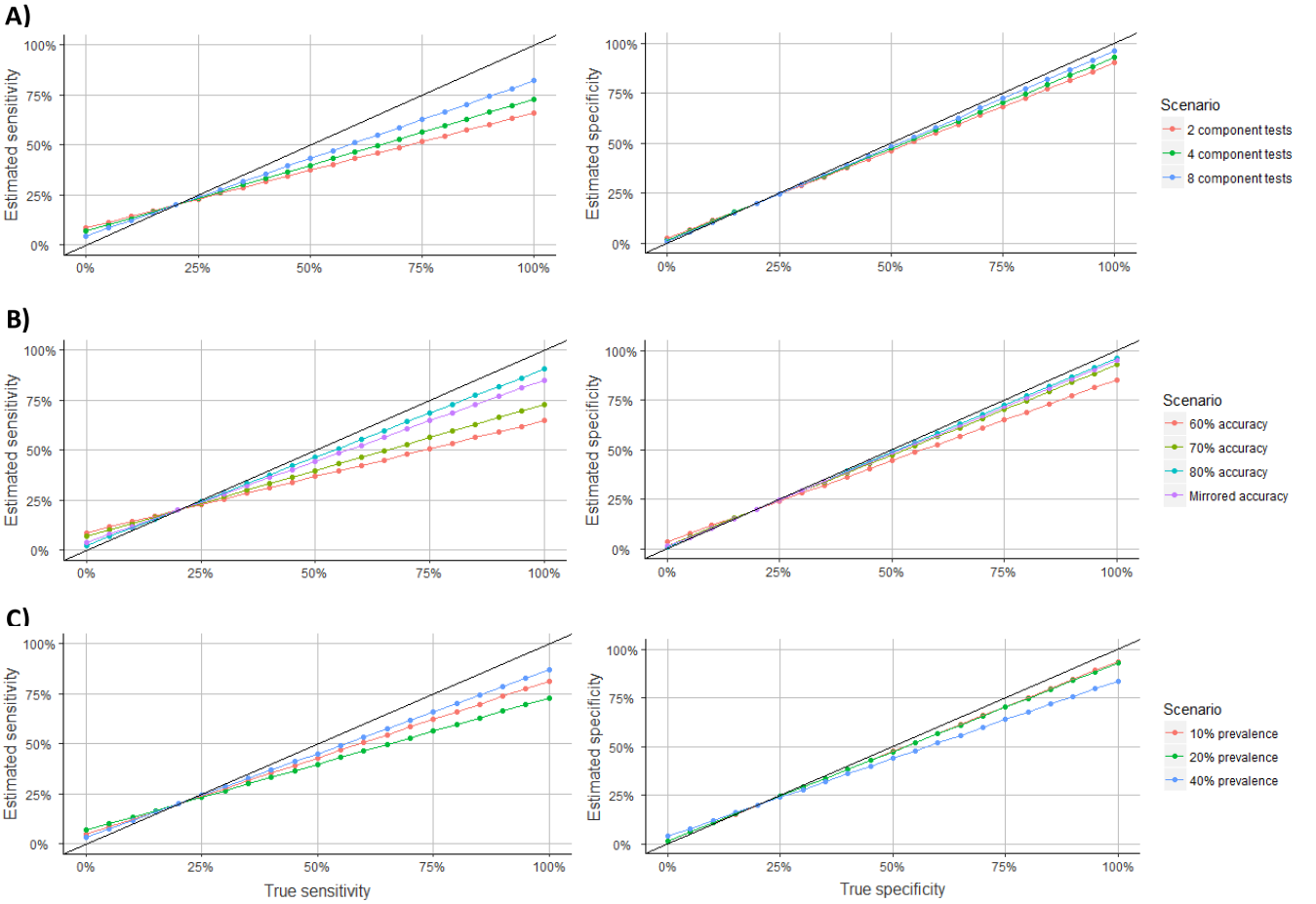
Bias in sensitivity and specificity estimates of the index test

Dichotomous target disease classifications by the expert panels in the aforementioned scenarios were used to estimate bias in sensitivity and specificity for a range of true values of an index test. (Figure 2) In all investigated scenarios there was deviation from the reference line, indicating that in virtually all cases there is bias in estimates of index test sensitivity and specificity. When considering the base scenario, combined with for example true values of 80% sensitivity and specificity of the index test, estimates for index test sensitivity and specificity by the expert panel were 60% and 75% respectively, leading to an absolute bias of 20% and 5%.

The amount of bias differed across scenarios. Figure 2A shows the shift for the low, medium and high number of component test scenarios. A larger number of component tests resulted in a lower bias for both index test sensitivity and specificity. In a similar fashion, increasing accuracy of component tests led to less bias in estimates of sensitivity and specificity. The mirrored and high accuracy scenarios showed similar bias in estimates.

While changes in target disease prevalence did not affect the AUROC of the reference standard (Figure 1), it did produce irregular results in terms of bias of sensitivity and specificity of the index test. In figure 2C, the bias in sensitivity was highest at medium target disease prevalence, lowest at high target disease prevalence, and intermediate at the lowest target disease prevalence. This can be explained by examining the distribution of test patterns shown in Figure 1. The difference in bias of sensitivity and specificity estimates of an index test between two scenarios was influenced by whether a component test results pattern shifted across the threshold used in the dichotomization process. For example, when looking at low and medium prevalence scenarios, there was a shift of the fourth test pattern across the 50% threshold. Individuals with that test pattern result were suddenly all classified as disease being absent in the low prevalence scenario, and all as disease being present in the medium prevalence scenario. As a consequence, there was a strong increase in bias of sensitivity estimates, while the effect on specificity bias was limited.

Figure 2. Range of true values of sensitivity and specificity of a hypothetical index test and their recalculated estimates. Scenarios with different number of component reference tests (A), accuracy (B), and prevalence (C) were taken for the reference standard. A reference line is given in solid black. Dichotomization was based on a target disease probability threshold of 50%. Index test specificity was fixed at 80% when calculating sensitivity, and vice versa.



Discussion

Forcing expert panels to dichotomize target disease classification leads to both target disease misclassification and biased accuracy estimates of the index test under study, even when individuals are consistently classified to their most likely target disease status. A series of scenarios were assessed in which an expert panel was given a set of component reference tests with varying characteristics combined with a range of true accuracy values for the index test. Virtually all scenarios lead to biased index test accuracy estimates. Increasing the number and/or accuracy of component reference tests reduced bias in the index test accuracy estimates. Varying target disease prevalence led to irregular shifts in bias of index test accuracy.

The scenarios that were investigated demonstrated a structural underestimation of index test sensitivity and/or specificity when (realistic) true values of at least 50% for both parameters were considered. However, it would be an error to assume that index test accuracy will always be underestimated when expert panels are used as a reference standard in diagnostic studies. In particular, the index test results might be correlated (conditionally dependent) for a given true disease status, which might lead to overestimation rather than underestimation of sensitivity and/or specificity of the index test. Also, in case of conditional dependence between component reference test results, adding more component tests may not always improve estimation of the accuracy of the index test. (14)

When looking at the distribution of the probability of target disease presence for different component test patterns (Figure 1), one might anticipate that a symmetrical distribution (i.e. equal distributions left and right of the threshold) will cancel out any target disease misclassifications made by a reference standard, which should then consequently reduce bias in accuracy estimates of the index test. When we simulated a scenario with such a symmetrical distribution, bias in estimated index test sensitivity and specificity were equal across the range of true values, however more research is required to investigate whether this minimalizes bias in sensitivity and specificity of the index test.

One tempting option in diagnostic studies would be to exclude individuals where there is significant uncertainty about the true disease status, as these have the highest probability of leading to erroneous target disease classification by the expert panel. However, this is ill-advised. Excluding cases in which there is uncertainty about the true disease status (i.e. close to the threshold) would mean the accuracy of the index test would only be generalizable to the assessment of the 'easy' cases with a high probability of either having or not having the target disease. This obviously does not represent the true target population of the index test, hence, such study patient exclusion will yield a distorted and too optimistic accuracy of the index test. Similar issues have been described for diagnostic case-control studies. (22-24)

Earlier studies have demonstrated similar effects on estimates of sensitivity and specificity of index tests when composite reference standards based on explicit decision rules were used. (14, 25) Expert panels as reference standards deal with similar issues as these composite reference standards, resulting in biased index test accuracy estimates. However, unlike composite reference standards with explicit decision rules, we studied the effect of target disease dichotomization based on the probability of target disease presence, which is commonly ignored when developing a composite reference standard. A recent paper expressed further concerns regarding such types of composite reference standards, and suggested alternatives such as latent class models to take into account uncertainty surrounding target disease classification. (15)

An alternative approach to minimize bias from dichotomous classification of target disease status, would be to allow for probabilistic target disease estimates on a continuous or ordinal scale, which have already been applied in a few diagnostic settings. (26, 27) Taking such probabilistic estimates of target disease presence is currently seldom being applied in studies exploiting expert panels as a reference standard, however have been described in the context of record linkage. (28, 29) Some authors have suggested obtaining ordinal target disease classes between the traditional disease present and absent options, such as "possible disease" and "intermediate classes". (30, 31) Others have suggested using methods such

as diagnostic probability functions based on expert diagnosis to obtain target disease probabilities. (27) Although it has been emphasized that eliciting expert judgments on disease status is a complex task. (32)

To fully appreciate the findings of this paper, there are some limitations that should be considered. First, in our simulations we have only considered dichotomous component reference tests, whereas in practice some test results may produce continuous outcomes. Unless these continuous tests can be used to perfectly separate individuals with and without the target disease, uncertainty in target disease classifications will remain present. Therefore, bias in index test accuracy estimates after dichotomization of target disease classification based on continuous diagnostic component reference tests, is also to be anticipated.

Secondly, we have not included conditional dependence between component reference tests nor between component tests and the index test. Conditional dependence is likely to be present in real-life situations, for instance, because tests are likely to make fewer errors in more severe cases compared to less severe cases. (33) We anticipate that similar problems as observed would occur for test results that are conditionally dependent. The exact influence of dependent test results may be a complicated interplay between the mechanism of the dependence between the tests, which may obviously vary between settings, the accuracy of the component tests and index test, and the prevalence of the disease. (14) While our results may be viewed as a simplification, the fact that the bias occurs even in the simplest situations should already be of great concern.

Finally, we have assumed that the expert panel is able to correctly estimate the target disease probability for all individuals, and that these individuals are consistently classified to the target disease status with the highest probability. In diagnostic research this may not always be realistic, especially when target disease probability estimates of individuals are centred around the threshold for dichotomous classification. Thus we may expect even more target disease classification errors when not all subjects are classified to the target disease status with the highest probability.

Our findings are not only applicable to expert panels serving as a reference standard in diagnostic studies, but also to other situations in which a dichotomous outcome classification is used, and uncertainty is not taken into account. Composite reference standards in diagnostic research (34, 35), adjudication committees used to classify endpoints in intervention or prognostic studies (36), and probabilistic medical record linkage (28, 37, 38), frequently force dichotomisation from their respective reference standards. As a result, similar biases may occur.

We conclude that dichotomizing target disease classification by a reference standard based on multiple imperfect component tests, such as a panel diagnosis, leads to biased accuracy estimates of the index test in a simulation study. The direction and magnitude of these biases were found to depend on the combination of the number of component reference tests, their accuracy, and the target disease prevalence. The bias found in this simulation study may not reflect the true bias in an empirical setting, as more complex interactions, such as conditional dependence and misclassification by expert panels (e.g. classifying an individual with a low probability of disease, as target disease present) may be at play. To potentially reduce these biases, alternatives to dichotomous classification of target disease by the reference standard should be sought after, such as obtaining target disease probability estimates per individual from the expert panel, or via a latent class analysis (39). Researchers involved in diagnostic studies that employ expert panels as a reference standard should be wary that solely asking for presence or absence of the target disease will limit the ability of unbiased estimation of index test accuracy. Performance of novel diagnostic tests needs to be established accurately in diagnostic research, and it should not have to suffer from the imperfectness of reference standard that it is being compared to.

References

1. Bachmann LM, Juni P, Reichenbach S, Ziswiler HR, Kessels AG, Vogelin E. Consequences of different diagnostic "gold standards" in test accuracy research: Carpal Tunnel Syndrome as an example. International journal of epidemiology. 2005;34(4):953-5. Epub 2005/05/25.

2. Reference Standard. (n.d.), *Mosby's Medical Dictionary, 8th edition.* (2009). Retrieved May 2 2018 from https://medical-dictionary.thefreedictionary.com/Reference+Standard.

3. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62(8):797-806. Epub 2009/05/19.

4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1):W1-12. Epub 2003/01/07.

5. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health technology assessment. 2007;11(50):iii, ix-51. Epub 2007/11/21.

6. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10(10):e1001531. Epub 2013/10/22.

7. Bertens LC, van Mourik Y, Rutten FH, Cramer MJ, Lammers JW, Hoes AW, et al. Staged decision making was an attractive alternative to a plenary approach in panel diagnosis as reference standard. J Clin Epidemiol. 2015;68(4):418-25. Epub 2014/12/03.

8. Bossuyt PM. Interpreting diagnostic test accuracy studies. Seminars in hematology. 2008;45(3):189-95. Epub 2008/06/28.

9. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ. 2002;324(7335):477-80. Epub 2002/02/23.

10. Trikalinos TA, Balion CM. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". Journal of general internal medicine. 2012;27 Suppl 1:S67-75. Epub 2012/06/08.

11. Miller DP, O'Shaughnessy KF, Wood SA, Castellino RA, editors. Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions. Medical Imaging 2004; 2004: SPIE.

12. Handels RL, Wolfs CA, Aalten P, Bossuyt PM, Joore MA, Leentjens AF, et al. Optimizing the use of expert panel reference diagnoses in diagnostic studies of multidimensional syndromes. BMC neurology. 2014;14:190. Epub 2014/10/05.

13. Thomeer M, Demedts M, Behr J, Buhl R, Costabel U, Flower CD, et al. Multidisciplinary interobserver agreement in the diagnosis of idiopathic pulmonary fibrosis. Eur Respir J. 2008;31(3):585-91. Epub 2007/12/07.

14. Schiller I, van Smeden M, Hadgu A, Libman M, Reitsma JB, Dendukuri N. Bias due to composite reference standards in diagnostic accuracy studies. Statistics in medicine. 2016;35(9):1454-70. Epub 2015/11/12.

15. Dendukuri N, Schiller I, de Groot J, Libman M, Moons K, Reitsma J, et al. Concerns about composite reference standards in diagnostic research. BMJ. 2018;360:j5779. Epub 2018/01/20.

16. Johnson S, Hollis C, Hennessy E, Kochhar P, Wolke D, Marlow N. Screening for autism in preterm children: diagnostic utility of the Social Communication Questionnaire. Archives of disease in childhood. 2011;96(1):73-7. Epub 2010/10/30.

17. Brugha TS, McManus S, Smith J, Scott FJ, Meltzer H, Purdon S, et al. Validating two survey methods for identifying cases of autism spectrum disorder among adults in the community. Psychological medicine. 2012;42(3):647-56. Epub 2011/07/30.

18. Fryback DG. Bayes' theorem and conditional nonindependence of data in medical diagnosis. Computers and biomedical research, an international journal. 1978;11(5):423-34. Epub 1978/10/05.

19. Simon D, Boring JR, III. Sensitivity, Specificity, and Predictive Value. In: rd, Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. Boston 1990.

20. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology. 2008;56(1):45-50. Epub 2007/12/26.

21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. Epub 1982/04/01.

22. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg. 2010;126(6):2234-42.

23. Kopec JA, Esdaile JM. Bias in case-control studies. A review. J Epidemiol Community Health. 1990;44(3):179-86.

24. Thomas SV, Suresh K, Suresh G. Design and data analysis case-controlled study in clinical research. Ann Indian Acad Neurol. 2013;16(4):483-7.

25. Walter SD, Macaskill P, Lord SJ, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Statistics in medicine. 2012;31(11-12):1129-38. Epub 2012/02/22.

26. van Houten CB, de Groot JA, Klein A, Srugo I, Chistyakov I, de Waal W, et al. A host-protein based assay to differentiate between bacterial and viral infections in preschool children (OPPORTUNITY): a double-blind, multicentre, validation study. Lancet Infect Dis. 2016.

27. Steurer J, Held U, Miettinen OS. Diagnostic probability function for acute coronary heart disease garnered from experts' tacit knowledge. J Clin Epidemiol. 2013;66(11):1289-95. Epub 2013/09/11.

28. Hof MH, Zwinderman AH. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. Statistics in medicine. 2012;31(30):4231-42. Epub 2012/07/19.

29. Hof MH, Zwinderman AH. A mixture model for the analysis of data derived from record linkage. Statistics in medicine. 2015;34(1):74-92. Epub 2014/10/03.

30. van Mourik Y, Bertens LC, Cramer MJ, Lammers JW, Reitsma JB, Moons KG, et al. Unrecognized heart failure and chronic obstructive pulmonary disease (COPD) in frail elderly detected through a near-home targeted screening strategy. Journal of the American Board of Family Medicine : JABFM. 2014;27(6):811-21. Epub 2014/11/09.

31. Poldervaart JM, Reitsma JB, Backus BE, Koffijberg H, Veldkamp RF, Ten Haaf ME, et al. Effect of Using the HEART Score in Patients With Chest Pain in the Emergency Department: A Stepped-Wedge, Cluster Randomized Trial. Ann Intern Med. 2017;166(10):689-97. Epub 2017/04/25.

32. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: eliciting experts' probabilities: John Wiley & Sons; 2006.

33. Brenner H. How independent are multiple 'independent' diagnostic classifications? Statistics in medicine. 1996;15(13):1377-86. Epub 1996/07/15.

34. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statistics in medicine. 1999;18(22):2987-3003. Epub 1999/11/02.

35. Naaktgeboren CA, Bertens LC, van Smeden M, de Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. BMJ. 2013;347:f5605.

36. Sepehrvand N, Zheng Y, Armstrong PW, Welsh R, Goodman SG, Tymchak W, et al. Alignment of site versus adjudication committee-based diagnosis with patient outcomes: Insights from the Providing Rapid Out of Hospital Acute Cardiovascular Treatment 3 trial. Clinical trials. 2016;13(2):140-8. Epub 2015/08/21.

37. Scheuren F, Winkler WE. Regression analysis of data files that are computer matched. Survey Methodology. 1993;19(1):39-58.

38. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. International journal of epidemiology. 2016;45(3):954-64. Epub 2015/12/22.

39. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard--a systematic review. American journal of epidemiology. 2014;179(4):423-31. Epub 2013/11/26.

Appendix

Strengthening the Reporting of Empirical Simulation Studies (STRESS)Discrete-event simulation guidelines STRESS-DES

Section/Subsection	Item	Recommendation
1. Objectives		
Purpose of the model	1.1	Explain the background and objectives for the model. The model aims to provide insight in the diagnostic performance of an expert panel, that have a number of component tests at their disposal, and provide insight in the bias in diagnostic accuracy estimates when an index test were to be assessed by such a panel.
Model Outputs	1.2	Define all quantitative performance measures that are reported, using equations where necessary. Specify how and when they are calculated during the model run along with how any measures of error such as confidence intervals are calculated. Proportion of misclassifications Area under the receiver-operator characteristic curve (AUROC) Percentage points bias in sensitivity Percentage points bias in specificity
Experimentation Aims	1.3	If the model has been used for experimentation, state the objectives that it was used to investigate. a.) Scenario based analysis – Provide a name and description for each scenario, providing a rationale for the choice of scenarios and ensure that item 2.3 (below) is completed. See table below b.) Design of experiments – Provide details of the overall design of the

experiments with reference to performance measures and their parameters (provide further details in *data* below).
First, for each scenario the distribution of test patterns is visually displayed against the likelihood of observing those patterns. For each scenario the proportion of misclassifications and AUROC were calculated. Second, for each scenario the estimated sensitivity and specificity of an index test is plotted against a range of true values of sensitivity and specificity of the index test. This shows the amount of bias for a given scenario.

c.) Simulation Optimisation – (if appropriate) Provide full details of what is to be optimised, the parameters that were included and the algorithm(s) that was be used. Where possible provide a citation of the algorithm(s).
Not applicable

Table 1

Scenario	# of tests	Component reference test sensitivity	Component reference test specificity	Target disease prevalence
Low number of tests[a]	2	70%	70%	20%
Medium number of tests[a]*****	4	70%	70%	20%
High number of tests[a]	8	70%	70%	20%
Low accuracy[b]	4	60%	60%	20%
Medium accuracy[b]*****	4	70%	70%	20%
High accuracy[b]	4	80%	80%	20%
Mirrored accuracy[b]	4	60%-70%-80%-90%	90%-80%-70%-60%	20%
Low prevalence[c]	4	70%	70%	10%
Medium prevalence[c]*****	4	70%	70%	20%
High prevalence[c]	4	70%	70%	40%

2. Logic		
Base model overview diagram	2.1	Describe the base model using appropriate diagrams and description. This could include one or more process flow, activity cycle or equivalent diagrams sufficient to describe the model to readers. Avoid complicated diagrams in the main text. The goal is to describe the breadth and depth of the model with respect to the system being studied. Illustration below depicts the process of determining test pattern options (Test 1 – 4), probability of observing that pattern (pP), probability of disease (pD1) and non-disease (pD0) in those patterns, the dichotomous class that pattern belongs to (Dich_class), and a worked out example for 100.000 individuals.

Table 2

	Test.1	Test.2	Test.3	Test.4	pP	pD1	pD0	Dich_class	Sample100K_Total	Sample100K_D1	Sample100K_D0
1	0	0	0	0	0.194	0.008	0.992	0	19370	162	19208
2	0	0	0	1	0.086	0.044	0.956	0	8610	378	8232
3	0	0	1	0	0.086	0.044	0.956	0	8610	378	8232
4	0	0	1	1	0.044	0.200	0.800	0	4410	882	3528
5	0	1	0	0	0.086	0.044	0.956	0	8610	378	8232
6	0	1	0	1	0.044	0.200	0.800	0	4410	882	3528
7	0	1	1	0	0.044	0.200	0.800	0	4410	882	3528
8	0	1	1	1	0.036	0.576	0.424	1	3570	2058	1512
9	1	0	0	0	0.086	0.044	0.956	0	8610	378	8232
10	1	0	0	1	0.044	0.200	0.800	0	4410	882	3528
11	1	0	1	0	0.044	0.200	0.800	0	4410	882	3528
12	1	0	1	1	0.036	0.576	0.424	1	3570	2058	1512
13	1	1	0	0	0.044	0.200	0.800	0	4410	882	3528
14	1	1	0	1	0.036	0.576	0.424	1	3570	2058	1512
15	1	1	1	0	0.036	0.576	0.424	1	3570	2058	1512
16	1	1	1	1	0.054	0.881	0.119	1	5450	4802	648

Base model logic	2.2	Give details of the base model logic. Give additional model logic details sufficient to communicate to the reader how the model works. Probability of observing a test pattern in combined with the probability of disease for that pattern. These can be used to determine the dichotomous class (0 or 1) based on a predefined threshold, in this case 0.5. For a given sample, the proportion of truly diseased and truly non-diseased can be calculated. These truly diseased and non-diseased samples (Sample100K_D1 and Sample100K_D0) can be used to calculate the number of positive and negative test results for an index test with a given sensitivity and specificity. (results not shown in table) Detailed instructions on how to obtain probability of pattern and/or disease and how these can be used to calculate the index test positives and negatives can be found in the manuscript under "Numerical example".
Scenario logic	2.3	Give details of the logical difference between the base case model and scenarios (if any). This could be incorporated as text or where differences are substantial could be incorporated in the same manner as 2.2. Base-case scenario is defined as 4 component tests, each with 70% sensitivity and specificity,

		and a prevalence of 0.2. The logic behind the scenarios is that each of these separate components was varied to a value lower and higher than the base-case values. For example: the low diagnostic accuracy scenario, meant using 60% instead of 70% sensitivity and specificity. For accuracy, a mirrored accuracy option was taken into account, since it was deemed to reflect a realistic situation in which expert panels are given both high sensitivity - low specificity tests, and low sensitivity - high specificity tests.	
Algorithms	2.4	Provide further detail on any algorithms in the model that (for example) mimic complex or manual processes in the real world (i.e. scheduling of arrivals/appointments/operations/maintenance, operation of a conveyor system, machine breakdowns, etc.). Sufficient detail should be included (or referred to in other published work) for the algorithms to be reproducible. Pseudo-code may be used to describe an algorithm. Any formulas used to calculate the aforementioned outcomes can be found in the R-script alongside the publication	
Components	2.5	2.5.1 Entities	Give details of all entities within the simulation including a description of their role in the model and a description of all their attributes. Entities are described here as the parameters used for the simulations. Probability of test pattern (pP): used in combination with probability of disease to determine the distribution of patterns across the Probability of disease (pD1): given a certain pattern, the probability of disease for an individual within that pattern. Dich_class: indicator to highlight whether individuals in a test pattern

			would be classified as diseased (D1) or non-diseased (D0) Sample100K: pP, pD1, and dich_class are applied to a hypothetical sample of 100.000 individuals. This ultimately allows for calculation of 2x2 tables and with that allows for calculation of misclassifications, as well as bias in sensitivity and specificity estimates of an index test.
		2.5.2 Activities	Describe the activities that entities engage in within the model. Provide details of entity routing into and out of the activity. Not applicable
		2.5.3 Resources	List all the resources included within the model and which activities make use of them. No resources, only simulated data
		2.5.4 Queues	Give details of the assumed queuing discipline used in the model (e.g. First in First Out, Last in First Out, prioritisation, etc.). Where one or more queues have a different discipline from the rest, provide a list of queues, indicating the queuing discipline used for each. If reneging, balking or jockeying occur, etc., provide details of the rules. Detail any delays or capacity constraints on the queues. Not applicable
		2.5.5 Entry/Exit Points	Give details of the model boundaries i.e. all arrival and exit points of entities. Detail the arrival mechanism (e.g. 'thinning' to mimic a non-homogenous Poisson process or balking) Not applicable

3. Data		
Data sources	3.1	List and detail all data sources. Sources may include: • Interviews with stakeholders, • Samples of routinely collected data, • Prospectively collected samples for the purpose of the simulation study, • Public domain data published in either academic or organisational literature. Provide, where possible, the link and DOI to the data or reference to published literature. All data source descriptions should include details of the sample size, sample date ranges and use within the study. All data have been simulated. Number of simulated patients, sensitivity and specificity of component tests, target disease prevalence, and range of sensitivity and specificity of the index test have been chosen based on scenarios, ranges or assumptions.
Pre-processing	3.2	Provide details of any data manipulation that has taken place before its use in the simulation, e.g. interpolation to account for missing data or the removal of outliers. No empirical data was used, hence no data manipulation has taken place
Input parameters	3.3	List all input variables in the model. Provide a description of their use and include parameter values. For stochastic inputs provide details of any continuous, discrete or empirical distributions used along with all associated parameters. Give details of all time dependent parameters and correlation. Clearly state:

		• Base case data • Data use in experimentation, where different from the base case. • Where optimisation or design of experiments has been used, state the range of values that parameters can take. Where theoretical distributions are used, state how these were selected and prioritised above other candidate distributions. For the variables in the model (on which the scenarios are based) we refer to table 1 (above). No distributions were used as we performed a deterministic assessment (no sample data was used). Conditional dependency between component tests, although optional, was not assessed (as pointed out in the manuscript) to avoid additional complexity.
Assumptions	3.4	Where data or knowledge of the real system is unavailable what assumptions are included in the model? This might include parameter values, distributions or routing logic within the model. No empirical data was used for these assessments. Scenarios, and with that the values for sensitivity, specificity, and prevalence, were chosen determined after deliberation amongst the researchers
4. Experimentation		
Initialisation	4.1	Report if the system modelled is terminating or non-terminating. State if a warm-up period has been used, its length and the analysis method used to select it. For terminating systems state the stopping condition. Not applicable State what if any initial model conditions have been included, e.g., pre-loaded queues and activities. Report whether initialisation of these variables is deterministic or stochastic. Not applicable

Run length	4.2	Detail the run length of the simulation model and time units. Model can be run within seconds
Estimation approach	4.3	State the method used to account for the stochasticity: For example, two common methods are multiple replications or batch means. Where multiple replications have been used, state the number of replications and for batch means, indicate the batch length and whether the batch means procedure is standard, spaced or overlapping. For both procedures provide a justification for the methods used and the number of replications/size of batches. Model is deterministic, as no observational data is used there is no uncertainty. Hence there was no need to account for stochasticity.
5. Implementation		

Software or programming language	5.1	State the operating system and version and build number.
		Windows 7 Enterprise SP1
		
		State the name, version and build number of commercial or open source DES software that the model is implemented in.
		Rstudio, version 1.1.383
		
		State the name and version of general-purpose programming languages used (e.g. Python 3.5).
		R's programming language
		
		Where frameworks and libraries have been used provide all details including version numbers.
		library(ggplot2)
		library(data.table)
		library(e1071) # for bincombinations
		library(pROC)
		library(ROCR)
		library(vcdExtra)
		library(grid)
		library(Hmisc)

		library(stringr) library(ggridges) library(plyr)
Random sampling	5.2	State the algorithm used to generate random samples in the software/programming language used e.g. Mersenne Twister. No random sampling, just discrete simulations are run If common random numbers are used, state how seeds (or random number streams) are distributed among sampling processes. No random sampling, just discrete simulations are run
Model execution	5.3	State the event processing mechanism used e.g. three phase, event, activity, process interaction. Not applicable *Note that in some commercial software the event processing mechanism may not be published. In these cases authors should adhere to item 5.1 software recommendations.* State all priority rules included if entities/activities compete for resources. Not applicable If the model is parallel, distributed and/or use grid or cloud computing, etc., state and preferably reference the technology used. For parallel and distributed simulations the time management algorithms used. If the HLA is used then state the version of the standard, which run-time infrastructure (and version), and any supporting documents (FOMs, etc.) Not applicable
System Specification	5.4	State the model run time and specification of hardware used. This is particularly important for large scale models that require substantial

		computing power. For parallel, distributed and/or use grid or cloud computing, etc. state the details of all systems used in the implementation (processors, network, etc.) Not applicable, model can be run within seconds
6. Code Access		
Computer Model Sharing Statement	6.1	Describe how someone could obtain the model described in the paper, the simulation software and any other associated software (or hardware) needed to reproduce the results. Provide, where possible, the link and DOIs to these. R-code is provided alongside the manuscript publication

"The more you know, the harder you will find it

To make up your mind, it doesn't really matter if you find

You can't see which grass is greener

Chances are it's neither, and either way it's easier

To see the difference, when you're sitting on the fence"

Tim Minchin, Australian comedian

Chapter 3

Estimating diagnostic accuracy in the presence of uncertainty surrounding the final target condition status provided by an expert panel: a sepsis case study

K. Jenniskens

C.A. Naaktgeboren

J.B. Reitsma

L. Hooft

K.G.M. Moons

M. van Smeden

J.J. Oosterheert

M.J.A. de Regt

J.W. van Uffen

(Work in progress)

Abstract

Introduction: Expert panels, used as reference standard in diagnostic accuracy studies, typically classify each patient as having or not having the target condition, even when they have remaining uncertainty about that classification. This has been shown to lead to biased diagnostic accuracy estimates of the index test. We aim to show how probabilistic estimates of presence of a target condition elicited from an expert panel can be used in diagnostic accuracy research.

Methods: The SPACE (SePsis in Acutely ill patients in the Emergency department) study, aimed at investigating the diagnostic value of clinical decision rules (SIRS, qSOFA, CBJ) for diagnosis of sepsis in the emergency room, was used as a case study. Both dichotomous (i.e. present or absent) and probabilistic estimates of sepsis status were obtained from the expert panels. Measures of diagnostic accuracy were calculated using three approaches: (1) (traditional) dichotomous sepsis classification; (2) an approach using probabilistic estimates for presence of sepsis as weights; (3) an approach using these probabilistic estimates in combination with the diagnostic odds ratio (DOR).

Results: A total of 306 patients were included in the analysis. A skewed distribution of probabilistic estimates for the presence of sepsis by the panel was observed (median=0.30). The panel expressed considerable uncertainty whether sepsis was present or not (probabilities between 0.2 and 0.8) in 57% of patients. Estimates of diagnostic accuracy varied considerably between the dichotomous and two probabilistic approaches, but also between the two probabilistic approaches. For example, sensitivity of SIRS was 91% for the dichotomous approach, 74% for the probabilistic weighting approach, and 99% for probabilistic DOR approach. Specificity was 46%, 47% and 60% for these approaches respectively.

Conclusions: Eliciting probabilistic estimates of target disease presence from expert panels, can provide valuable insight in the uncertainty that is normally ignored in dichotomous target disease classification. Different approaches exist on how to incorporate this uncertainty when estimating diagnostic accuracy measures and results can vary substantially depending

on the assumptions made. When substantial uncertainty about the final diagnosis is present in a considerable proportion of patients, it may be questioned whether the diagnostic accuracy framework is still useful.

Introduction

Detecting sepsis in patients presenting at the emergency department is of vital importance, as this could progress into septic shock which is associated with high mortality and morbidity. (1, 2) Diagnosing sepsis is challenging, due to a heterogeneous clinical presentation, and complex underlying pathophysiology. Hence numerous diagnostic accuracy studies have been performed, looking at index tests or clinical decision rules to improve clinical decision-making for patients suspected of sepsis. For sepsis, a single reference test providing error-free classification of the target condition is not available. (3-5) One proposed solution for such situations is to use an expert panel, in which a group of experts assess multiple relevant pieces of information (e.g. patient characteristics, diagnostic tests, follow-up information, and response to treatment) to make a final target condition classification. (6)

Typically, experts are asked to provide a dichotomous classification of the target condition (i.e. target condition present or absent) for each patient. When there is disagreement between experts, a final decision on target disease status of an individual can be made through majority vote, or deliberation until consensus is reached. (7) After obtaining the final target condition classification for each patient, measures of diagnostic accuracy (e.g. sensitivity, specificity) of the index test or clinical decision rule can be calculated in the traditional way. (4, 8) In this traditional approach, any remaining uncertainty among experts about the dichotomous target condition classification that may be present in certain patients, is basically ignored.

We recently showed in a simulation study that forcing dichotomous target condition classification by expert panels, thus ignoring any uncertainty in target disease classification, introduces bias in estimates of index test diagnostic accuracy. (9) This study looked at various scenarios in which we varied the number and accuracy of tests available to the expert panel, as well as the target disease prevalence.

Eliciting probabilistic estimates of patient target disease status from expert panels may be a more sound alternative to dichotomous target disease

classification. This will allow the panel to provide the likelihood of presence of the target condition ("probabilistic approach"), rather than the dichotomous classification of presence or absence of the target condition ("dichotomous approach"). Although probabilistic estimates of target condition presence may provide less biased estimates of diagnostic accuracy of the index tests, there is still the methodological challenge of calculating diagnostic accuracy measures for the index test, as the contingency table cannot be constructed in the traditional way.

The goal of this paper is to demonstrate how probabilistic estimates for presence of the target condition can be used to calculate diagnostic accuracy outcomes. Furthermore, we aim to explore and investigate how diagnostic accuracy measures of index tests differ between the traditional dichotomous approach, and the approach using probabilistic estimates of presence of the target condition. We will use a real-life diagnostic accuracy study as a case study: the SPACE study for prediction of sepsis in suspected patients in the emergency room.

Methods

Our aim is to compare and investigate difference in estimates of diagnostic accuracy between the dichotomous and two different probabilistic approaches when using an expert panel to obtain the final diagnosis. First we will briefly illustrate how the dichotomous approach is used to for traditional calculation of diagnostic accuracy of an index test. Next we explain what steps need to be taken to calculate diagnostic accuracy when using estimates for probability that the target condition is present. Then we describe how that information was acquired and used for our illustrative example, the SPACE study.

Dichotomous approach

The dichotomous approach is the most commonly used method for calculating diagnostic accuracy of an index test when using an expert panel as the reference standard. The basic concept is that each expert (preferably from different medical disciplines) in the panel provides the probability that the target condition is present. Final target condition classification is often

based majority vote, or on consensus between experts through deliberation. (5) Positive and negative index test results can be cross-classified with the final dichotomous target condition classification in a contingency table, allowing for calculation of diagnostic accuracy measures in the traditional way. (Table 1)

Table 1. Hypothetical example of how the contingency table for calculating diagnostic accuracy of an index test can be constructed, using the dichotomous target condition classification by the expert panel. TP = true positive; FP = false positive; TN = true negative; FN = false negative.

Patient	Expert panel dichotomous target condition classification	Index test result	TP	FP	TN	FN
1	Present	Positive	1	-	-	-
2	Absent	Negative	-	-	1	-
3	Absent	Positive	-	1	-	-
4	Present	Positive	1	-	-	-
5	Absent	Negative	-	-	1	-
6	Present	Negative	-	-	-	1

	Disease present	Disease absent	
Index test positive	∑TP = 2	∑FP = 1	**3**
Index test negative	∑FN = 1	∑TN = 2	**3**
	3	**3**	

Probabilistic approach

In the probabilistic approach experts are instructed to give the probability that the target condition is present for each individual. This allows the experts to express any remaining uncertainty regarding the target condition status. When a probability of target condition presence of 80% is provided to each individual in a group of 100 patients, the (true) number of patients having the target condition would be 80. Of course, in absent of a gold standard, it is not exactly known which patients in the total group have the target condition. Our assumption is that the probabilities of an expert panel are not systematically biased, in other words, the estimated probabilities of the panel are well calibrated.

Reconstructing the true contingency table based on the probabilistic estimates of target condition presence requires additional assumptions and calculations. The total number of positive and negative index test results are directly observed. The total number of patients with and without sepsis can be calculated by summing the expert panel estimates of the target condition being present, assuming that these panel estimates are well calibrated. Therefore, the row and column totals of the contingency table are known. The critical issue now becomes what the distribution of these row and column totals is across the remaining cells (i.e. TP, FP, TN, FN) of the contingency table. (Table 2) Multiple combinations are still possible, even when the row and column totals are known.

Two methods to estimate the expected contingency table, each using different assumptions, will be described in the following sections: the probabilistic weighting approach and the probabilistic diagnostic odds ratio (DOR) approach. For our case study, both approaches will be demonstrated to estimate the contingency table.

Table 2. Hypothetical example of how the contingency table for calculating diagnostic accuracy of an index test can be constructed, using the probability of disease as provided by the expert panel. TP = true positive; FP = false positive; TN = true negative; FN = false negative.

Patient	Probability that target condition is present	Index test result	TP	FP	TN	FN
1	0.8	1	?	?	?	?
2	0.2	0	?	?	?	?
3	0.1	0	?	?	?	?
4	0.7	1	?	?	?	?
5	0.5	0	?	?	?	?
6	0.9	1	?	?	?	?

	Disease present	Disease absent	
Index test positive	?	?	**3**
Index test negative	?	?	**3**
	3.2	**2.8**	

Probabilistic weighting approach

This method uses weighting of the index test result for each individual based on the probability that the target condition is present according to the expert panel. For example, the first case from Table 2 has 0.8 probability that the target condition is present, and a positive index test result. Using the weighting approach, that positive test result should be taken as 0.8 true positive and 0.2 false positive index test result. Doing this for the observed positive and negative index test results in all patients, and summing the total, would provide an estimate of the expected contingency table. (Table 3) This also preserves the previously mentioned row and column totals from the observed empirical data.

Table 3. The estimated expected contingency table for the hypothetical example based on probabilistic estimates of target condition presence using the weighting method.

	Disease present	Disease absent	
Index test positive	∑TP = 2.4	∑FP = 0.6	**3**
Index test negative	∑FN = 0.8	∑TN = 2.2	**3**
	3.2	**2.8**	

This method assumes that the expert panel provides the best possible estimate of target condition presence, and the index test result will not change the estimate by the expert panel for a given patient. In other words, the information from the index test being evaluated is already incorporated in the probabilistic estimates of target condition presence as given by the panel, either directly (the index test results were available to the panel) or indirectly (index test information is captured in the result of other test and follow-up information).

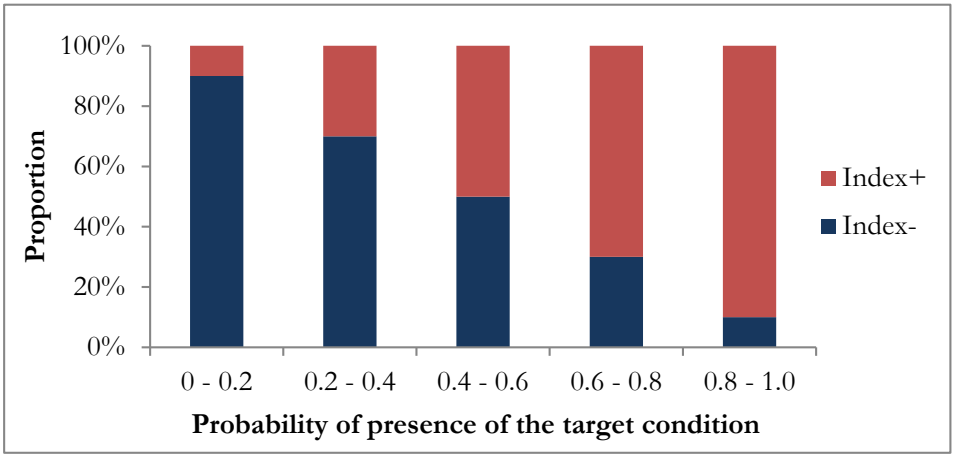
Probabilistic diagnostic odds ratio approach

To estimate the most likely values for the individual cells in the contingency table given the observed row and column totals, one additional parameter is required. One logical parameter would be the diagnostic odds ratio (DOR). The DOR is a single measure of diagnostic accuracy that indicates how the ratio of positive versus negative index test results differs between patients with and without the target condition, by taking the ratio of these ratios. An accurate index test will produce many (true) positive index test results and few (false) negative results among patients with the target condition, while among patients without the target condition there will be few (false) positive index test results and many (true) negative results. Dividing ratios of an informative diagnostic test on each other will produce a high DOR, whereas in an uninformative diagnostic test, these ratios will be the same among patients with and without the target, producing a DOR close to 1.

Figure 1 shows individual patients stratified according to their probability of the target condition being present as determined by the expert panel, together with the number of positive and negative index test results for each of these strata. Given an informative index test, the ratio of positive to negative index test results should increase as the probability of presence of the target condition increases. For example, consider the subgroup with a probability of presence of the target condition 0 – 0.2. If (on average) 10% of the individuals in this subgroup have the target condition, ideally only 10% of the index test results in this group should be positive. In the subgroup of 0.20 – 0.40, one would expect positive results in 30% of individuals. If there is a high linear correlation between the ratio of positive and negative index test results and the probability of target condition presence (as is the case in Figure 1), the index test will have a high DOR.

By performing a logistic regression with the index test result (positive or negative) as outcome and the probability of presence of the target condition provided by the expert panel as a covariate, an estimate of the DOR can be calculated. This DOR can then be combined with the observed row and column totals to estimate the best fitting contingency table by using an optimization program or script. (e.g. Excel GRG non-linear solver or nleqslv package in R) See Appendix for further details. Once the contingency table has been estimated, measures of diagnostic accuracy such as sensitivity, specificity, PPV, and NPV can be calculated in the traditional way.

Figure 1. Example showing the expected ratio of positive to negative index test results as a function of the probability of the target condition being present as determined by the expert panel.



Case study

The SPACE study (Sepsis in ACutely ill patients in the Emergency room) was taken as a case study to illustrate the use of the probabilistic approach for calculating diagnostic accuracy estimates. The primary aim of the SPACE study was to investigate what the diagnostic value of three clinical decision rules is for diagnosing sepsis in suspected patients presenting at the emergency department.

In short, data were prospectively collected from adult patients suspected of an infection presenting at the emergency department of the University Medical Center Utrecht between January 2018 and April 2018 for the internal medicine department. Suspected infection was defined by the treating physician in the ED as either the working diagnosis of infection or the differential diagnosis stated in the clinical chart. Patients were included in a consecutive series. No additional exclusion criteria were used.

SIRS (Systemic Inflammatory Response Syndrome) and the qSOFA (quick Sequential Organ Failure Assessment) are the two clinical decision rules that will be assessed with regard to their diagnostic accuracy for diagnosing

sepsis. Both can be calculated based on readily available measurements from clinical practice. SIRS consist of tachycardia (heart rate >90 beats/min), tachypnea (respiratory rate >20 breaths/min), fever or hypothermia (temperature >38 or <36 °C), and leukocytosis, leukopenia, or bandemia (white blood cells >1,200/mm3, <4,000/mm3 or bandemia ≥10%). (10) qSOFA score can be calculated based on respiratory rate ≥22/min, change in mental status, or systolic blood pressure ≤100 mmHg were present. (11) Both SIRS and qSOFA were considered positive if at least two criteria were present, as suggested by previous publications. (12) Alongside the two clinical decision rules, diagnostic accuracy of clinical bedside judgement (CBJ) for diagnosing sepsis in patients presenting at the ED was evaluated. This was assessed by an automated system in the patient record asking the treating physician whether or not sepsis was present at the moment of ED visit.

An expert panel was used as a reference standard to obtain the final diagnosis whether sepsis was present or absent. The panel consisted of at least two experts, taken from a pool of physicians involved in internal infectious diseases, emergency medicine, acute internal medicine, intensive care medicine, and general internal medicine. Experts were provided with all clinically relevant information, before, during, and after the patient was admitted to the hospital. They were asked to give a dichotomous (yes / no) answer to the question: was sepsis present at time of admission to the emergency department? In addition, experts were requested to provide the likelihood that the patient had sepsis at time of admission to the emergency department on a scale from 0 to 10, with 0 representing absolute certainty that the patient did not have sepsis, and 10 representing absolute certainty that the patient did have sepsis. Experts could also give feedback on their main motivation for providing a specific likelihood. CBJ, SIRS and qSOFA scores were not directly provided to the experts in the panel, however SIRS and qSOFA could be manually calculated if so desired.

Likelihoods provided by the expert panel were interpreted as probability estimates for sepsis status. For example, a likelihood of 6 was seen as a 0,6 probability of sepsis being present for a given individual. Probability of sepsis presence reported by at least two experts was weighted to provide a

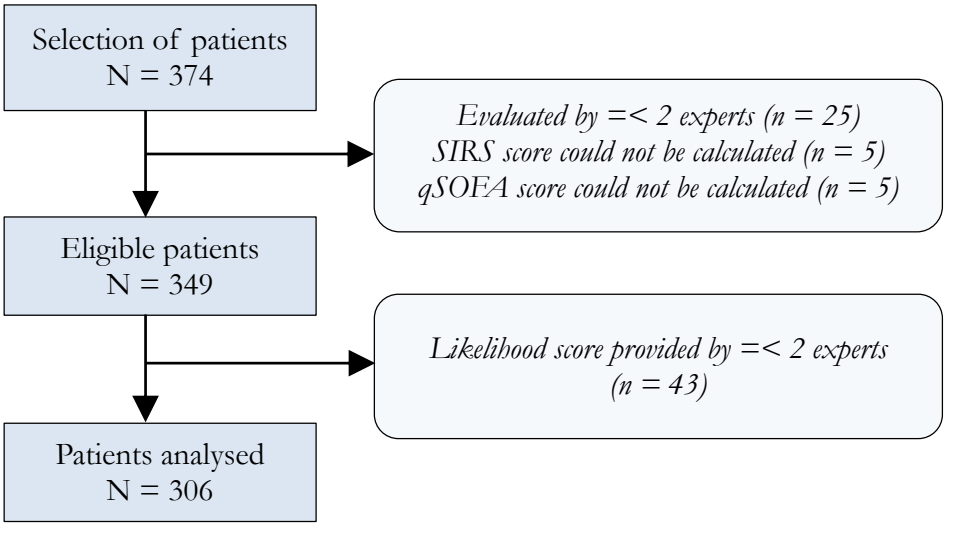
single probability of sepsis being present in an individual according to the expert panel.

If one or more measurements required for calculating SIRS or qSOFA was missing, the minimum score was determined, which was taken as the dichotomous index test result. If the minimum score of SIRS or qSOFA could not be determined, CBJ was missing, or there was no more than majority vote or likelihood estimate from the expert panel, multiple imputation would be used to account for the missing data. If any of these variables were missing in less than 5% of patients, complete case analysis was performed.

Results

A total number of 374 patients suspected of sepsis presented at the emergency room during the study period. Minimum scores (allowing for dichotomous classification) for SIRS and qSOFA could be calculated for all except five cases, which were excluded from the analysis.

Figure 2. Flowchart of patients in the SPACE study. Note that the reasons for exclusion are not mutually exclusive.



In 306 of those patients at least two experts provided dichotomous and probabilistic estimates of sepsis presence. Diagnostic accuracy measures were calculated for this group. Figure 2 provides a flowchart with additional information.

Figure 3 shows the distribution of the weighted probability estimates for presence of sepsis in patients with at least two expert panel scores, and the proportion of positive (red) and negative (blue) SIRS, qSOFA, and CBJ test results. Probabilities of sepsis presence provided by the expert panel followed a skewed distribution, with a median of 0.30. There was considerable uncertainty (i.e. a probability of sepsis presence between 0.2 and 0.8) in 57% of patients.

Positive test results were more prevalent for all three index tests as probability of sepsis presence increased. SIRS test results were positive in 185 patients (60.5%), and were distributed across the range of probabilities that sepsis is present according to the expert panel. qSOFA was only positive in 18 patients (5.9%) suspected of sepsis, however these were mainly observed in the group with a probability of sepsis >0.5 according to the expert panel. Positive CBJ results were observed more frequent in 46 out of 306 patients (15%). Most were mainly observed from a probability of sepsis of 0.3 and higher. Below a threshold of 0.2 none of the 81 patients were positive on the qSOFA or CBJ tests.

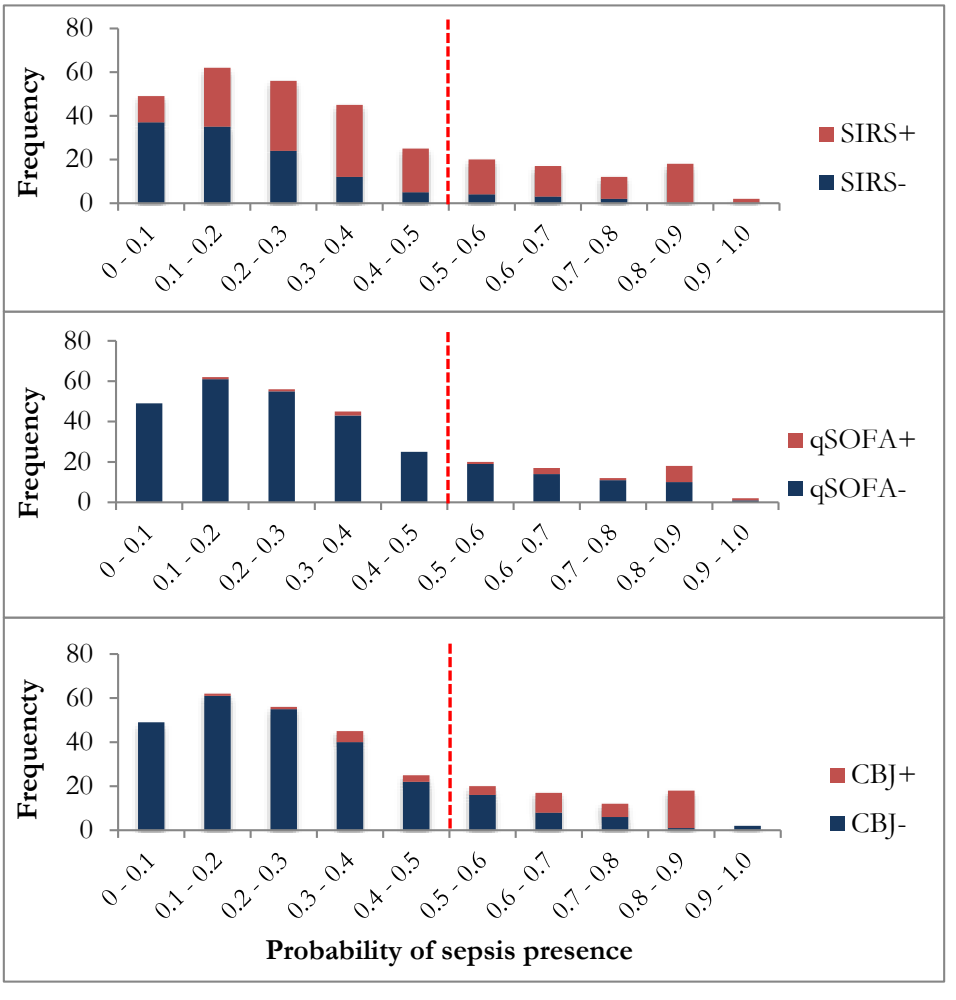
Comparison of diagnostic accuracy estimates between approaches

Three approaches were used to construct the contingency table: the dichotomous approach, the probabilistic weighting approach, and the probabilistic DOR approach. These can be used to calculate diagnostic accuracy measures for SIRS, qSOFA, and CBJ, which are given in tables 4, 5 and 6 respectively.

The dichotomous classification approach resulted in a much lower apparent prevalence (18%) compared to the apparent prevalence from the probabilistic approach (35%) because of the skewed distribution of the probabilities of sepsis presence given by the panel. This led to a higher number of true positives and false negatives, and to a lower number of false

positives and true negatives. However, the absolute impact on diagnostic accuracy measures differed between index tests. Note that the total number of positive expert panel classification (either from dichotomous or probabilistic classification) was equal across all contingency tables.

Figure 3. Distribution of weighted expert panel probabilities presence of sepsis, and the frequency of positive (red) and negative (blue) index test results. The intuitive threshold for dichotomisation of sepsis is depicted by the red dotted line.



In the dichotomous approach, SIRS was demonstrated to have relatively high sensitivity (91%) and NPV (96%), whereas specificity (46%) and PPV (27%) were considerably lower. (Table 4) The probabilistic weighting approach resulted in decreased sensitivity (74%) and NPV (77%) and increased PPV (43%). When the probabilistic DOR approach was used, diagnostic accuracy of SIRS was higher on all four diagnostic accuracy measures when compared to the traditional dichotomous approach.

Contingency tables and diagnostic accuracy estimates for qSOFA are provided in table 5. Dichotomous sepsis classification resulted in high specificity (98%) and NPV (85%), moderate PPV (67%), and poor sensitivity (22%). The probabilistic weighting approach resulted in both sensitivity and NPV of qSOFA dropping by 10% and 18% respectively when compared to the dichotomous approach. Specificity and PPV remained largely unaffected. The probabilistic DOR approach led to a 5% and 16% reduction of sensitivity and NPV, but resulted in near perfect specificity and PPV of qSOFA.

The diagnostic accuracy of CBJ is assessed in table 6. Using the dichotomous approach, CBJ had relatively good specificity (91%) and NPV (88%), with a sensitivity and PPV of 42% and 50% respectively. When the probabilistic weighting method was applied, sensitivity and NPV dropped by 16% and 19% compared to the dichotomous approach, while PPV increased and specificity was unaffected. Specificity and PPV increased to near perfect values when the probabilistic DOR approach was used, while NPV dropped to 76% and sensitivity was unaffected.

Table 4. Contingency table and diagnostic accuracy estimates for SIRS as an index test, constructed using dichotomous and probabilistic (weighting and DOR) approaches. DOR = diagnostic odds ratio; PPV = positive predictive value; NPV = negative predictive value.

	Dichotomous		Probabilistic (weighting)		Probabilistic (DOR)		
	Expert+	**Expert-**	**Expert+**	**Expert-**	**Expert+**	**Expert-**	
SIRS+	50	135	79.9	105.1	106.6	78.4	**185**
SIRS -	5	116	27.6	93.4	1.0	120.0	**121**
	55	**251**	**107.5**	**198.5**	**107.5**	**198.5**	
Sens.	91%		74%		99%		
Spec.	46%		47%		60%		
PPV	27%		43%		58%		
NPV	96%		77%		99%		

Table 5. Contingency table and diagnostic accuracy estimates for qSOFA as an index test, constructed using dichotomous and probabilistic (weighting and DOR) approaches. DOR = diagnostic odds ratio; PPV = positive predictive value; NPV = negative predictive value.

	Dichotomous		Probabilistic (weighting)		Probabilistic (DOR)		
	Expert+	**Expert-**	**Expert+**	**Expert-**	**Expert+**	**Expert-**	
qSOFA+	12	6	12.4	5.6	17.9	0.1	**18**
qSOFA -	43	245	95.1	192.9	89.6	198.4	**288**
	55	**251**	**107.5**	**198.5**	**107.5**	**198.5**	
Sens.	22%		12%		17%		
Spec.	98%		97%		100%		
PPV	67%		69%		100%		
NPV	85%		67%		69%		

Table 6. Contingency table and diagnostic accuracy estimates for CBJ as an index test, constructed using dichotomous and probabilistic (weighting and DOR) approaches. DOR = diagnostic odds ratio; PPV = positive predictive value; NPV = negative predictive value.

	Dichotomous		Probabilistic (weighting)		Probabilistic (DOR)		
	Expert+	**Expert-**	**Expert+**	**Expert-**	**Expert+**	**Expert-**	
CBJ+	23	23	27.7	18.3	45.3	0.8	**46**
CBJ -	32	228	79.8	180.2	62.3	197.7	**260**
	55	**251**	**107.5**	**198.5**	**107.5**	**198.5**	
Sens.	42%		26%		42%		
Spec.	91%		91%		100%		
PPV	50%		60%		98%		
NPV	88%		69%		76%		

Discussion

This study was performed to demonstrate how probabilistic estimates for the presence of the target condition can be incorporated in diagnostic research using expert panels as a reference standard. Using the SPACE study we illustrated how probabilistic estimates can explicitly be applied for calculating diagnostic accuracy measures of the index test under study rather than the traditional approach in which this uncertainty is ignored.

The expert panel expressed considerable uncertainty about the final diagnosis of sepsis in over half (57%) of study participants, despite having access to all diagnostic test results, follow-up information, and information on response to treatment. This underlines the difficult nature of diagnosing sepsis.

Prevalence was notably different between the dichotomous and probabilistic approaches, attributable to the skewed distribution of probabilities that sepsis was present. Both the probabilistic weighting and DOR approach resulted in significantly different diagnostic accuracy estimates when compared to the traditional method of dichotomous sepsis classification.

A simulation study has previously shown that dichotomous classification of the target condition by an expert panel leads to biased estimates of sensitivity and specificity. (9) Similar biases have been demonstrated for composite reference standards. (13-15) The results of our previous simulation study indicated that bias in diagnostic accuracy estimates of the index test is likely when an expert panel has substantial remaining uncertainty about target disease classification. In other words, if more patients are classified surrounding the threshold for dichotomisation of the target disease (i.e. a probability for the presence of sepsis of 0.5), the reference standard is likely to misclassify individuals, leading to bias of index test sensitivity and specificity estimates. Remaining uncertainty was clearly an issue in our empirical sepsis study, and therefore bias in diagnostic accuracy estimates based on the dichotomous approach is to be expected.

To acknowledge the significant uncertainty expressed by the expert panel, we applied a probabilistic weighting and probabilistic DOR approach for incorporating this uncertainty when calculating the diagnostic accuracy of SIRS, qSOFA, and CBJ. These two approaches resulted in widely different estimates of diagnostic accuracy, attributable to the underlying assumptions of each method. The probabilistic weighting approach assumes that the expert panel has all information available to make the most accurate estimates of the probability of target condition classification; in other words, the index test adds no new information. Hence, if an expert panel is uncertain about the presence of the target condition in a group of study participants (i.e. the panel does not have perfect accuracy), the accuracy of the index test will always be bounded by this imperfection. The index test result may be provided to the expert panel to ensure that this assumption holds, although this may introduce the risk of incorporation bias. (4, 16, 17) Incorporation bias will occur if the panel members overestimate (or underestimate) the value of the new test under evaluation because of high expectations or hype surrounding the new test. Probabilities of presence of the target condition elicited from the expert panel are then biased, and the association with index test results artificially too high. To avoid this bias, the results of the new test under evaluation are often withheld from the expert panel.

The probabilistic DOR approach assumes that the results of the index test could provide additional diagnostic information to an expert panel diagnosis. This means that information of index test results is not (fully) captured by the other information provided to the expert panel. Whether this assumption holds, is difficult to judge. Alternative methods may be sought after, such as latent class analysis, which explores correlations between the results of an index test and various other sources of information (e.g. patients' history, clinical examination, imaging, laboratory or function tests, severity scores) and relates them to an unknown (latent) outcome. (6, 18, 19) In latent class analysis, none of the single pieces of information (neither stand-alone or as a combination) can be promoted to the status of a reference standard. In this regard the aim of the latent class approach is similar to that of expert panel, as they use the same relevant pieces of information to determine diagnostic accuracy of an index test. The advantage of the latent class approach, being a statistical model, is that there is no risk of incorporation bias, as this approach it is not influenced by high expectations or hype with regard to the new index. The index test can thus safely be added to the model, allowing for estimation of its diagnostic accuracy. The drawback of the latent class approach is that it does not necessarily create a clinically relevant definition of the target condition. Furthermore, results of latent class analysis are directly subject to assumptions made during modelling.

A more fundamental question may arise: is the diagnostic accuracy framework still a useful concept when there is substantial uncertainty regarding the final diagnosis, or should alternative ways of validation of test results be explored? Abandoning the diagnostic test accuracy paradigm means index test results are related to other relevant clinical features related to the presence of absence of the target condition. Validation is an alternative process to evaluate a medical test in the absence of a gold reference standard. In this context, validity refers to whether the index test can identify clinically meaningful cases of the target condition. (6, 20, 21) Validation of a test can best be understood as a gradual process whereby the relationship between the test and various outcomes (such as prognosis or impact of treatment on test positive cases) is assessed. One might also check

whether findings are consistent with intuitive mechanics of the target condition, such as spreading patterns in infectious diseases. (22) In this way a body of evidence is generated, increasing the degree of confidence we can place on inferences about patients with positive and negative test results.

This paper should be viewed as a starting point for how to use probabilistic estimates of presence of the target condition elicited from an expert panel. There are however still several challenges for future research. First, it was assumed that the expert panel is well calibrated, meaning that the mean probability estimate of sepsis for a given patient provided by at least two experts in the panel, resembles the true probability of sepsis being present. The fact that these two experts can, and often will, differ in their individual probability estimates, points out that there is remaining uncertainty surrounding the average point estimate of the panel. This uncertainty could have been taken into account by, for example using a Monte-Carlo simulation based on the distribution of individual expert probability estimates of sepsis. (23)

Furthermore, researchers should be wary about selective missing values of probability estimates of target condition presence by experts. Not only might these values be selectively missing within patients, but also within experts (across patients). Ignoring these missing estimates may bias the mean value of probability that the target condition is present, which may subsequently affect diagnostic accuracy estimates of the index test. Multiple imputation could prove to be a valid alternative to complete case analysis (24), although further (simulation) studies should be performed to assess whether this is the case.

There are several other considerations researchers may have regarding the use of probabilistic estimates of target condition presence. There may be a relevant spectrum within the same target condition, in which less and more severe forms can be distinguished. (26) Diagnostic accuracy of the index test may differ between these subtypes of the target condition. Sample size may also affect the observed point estimates of diagnostic accuracy. Normally exact binomial distribution should be used to provide confidence intervals,

but given that probabilistic estimates of presence of the target condition are given on a continuous scale from 0 to 1, this may not be suitable.

In this paper we have used a case study to demonstrate how probabilistic estimates of presence of sepsis, elicited from an expert panel, can be used to calculate diagnostic test accuracy measures of three clinical decision rules. Considerable uncertainty surrounding sepsis classification was observed in over half of all study participants. Dichotomisation resulted in significantly different diagnostic accuracy estimates of the index test compared to the probabilistic weighting and DOR approach. The latter two also produced markedly different estimates. As both approaches have different assumptions, it is critical to match the analysis with the design and set-up how experts derived the probabilities of target condition presence. If we cannot, we should consider abandoning the diagnostic accuracy framework in settings where there is substantial uncertainty on target condition status in a considerable proportion of study participants.

References

1. Medam S, Zieleskiewicz L, Duclos G, Baumstarck K, Loundou A, Alingrin J, et al. Risk factors for death in septic shock: A retrospective cohort study comparing trauma and non-trauma patients. Medicine. 2017;96(50):e9241.

2. Sherwin R, Winters ME, Vilke GM, Wardi G. Does Early and Appropriate Antibiotic Administration Improve Mortality in Emergency Department Patients with Severe Sepsis or Septic Shock? The Journal of emergency medicine. 2017;53(4):588-95.

3. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62(8):797-806.

4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1):W1-12.

5. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10(10):e1001531.

6. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health technology assessment. 2007;11(50):iii, ix-51.

7. Bertens LC, van Mourik Y, Rutten FH, Cramer MJ, Lammers JW, Hoes AW, et al. Staged decision making was an attractive alternative to a plenary approach in panel diagnosis as reference standard. J Clin Epidemiol. 2015;68(4):418-25.

8. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ. 2002;324(7335):477-80.

9. Jenniskens K, Naaktgeboren CA, Reitsma JB, Hooft L, Moons KGM, van Smeden M. Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: A simulation study. J Clin Epidemiol. 2019;111:1-10.

10. Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. Journal of thoracic disease. 2017;9(4):943-5.

11. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-74.

12. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10.

13. Schiller I, van Smeden M, Hadgu A, Libman M, Reitsma JB, Dendukuri N. Bias due to composite reference standards in diagnostic accuracy studies. Statistics in medicine. 2016;35(9):1454-70.

14. Walter SD, Macaskill P, Lord SJ, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Statistics in medicine. 2012;31(11-12):1129-38.

15. Dendukuri N, Schiller I, de Groot J, Libman M, Moons K, Reitsma J, et al. Concerns about composite reference standards in diagnostic research. BMJ. 2018;360:j5779.

16. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen JJAoim. Sources of variation and bias in studies of diagnostic accuracy. A systematic review. 2004;140(3):189-202.

17. Handels RL, Wolfs CA, Aalten P, Bossuyt PM, Joore MA, Leentjens AF, et al. Optimizing the use of expert panel reference diagnoses in diagnostic studies of multidimensional syndromes. BMC neurology. 2014;14:190.

18. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statistics in medicine. 1999;18(22):2987-3003.

19. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KG, de Groot JA. Latent class models in diagnostic studies when there is no reference standard--a systematic review. American journal of epidemiology. 2014;179(4):423-31.

20. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use: Oxford University Press, USA; 2015.

21. Bland JM, Altman DGJB. Validating scales and indexes. 2002;324(7337):606-7.

22. Ewer K, Deeks J, Alvarez L, Bryant G, Waller S, Andersen P, et al. Comparison of T-cell-based assay with tuberculin skin test for diagnosis of Mycobacterium tuberculosis infection in a school tuberculosis outbreak. 2003;361(9364):1168-73.

23. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health economics. 2005;14(4):339-47.

24. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine. 2010;29(28):2920-31.

25. Brenner H. How independent are multiple 'independent' diagnostic classifications? Statistics in medicine. 1996;15(13):1377-86.

26. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ. 2016;353:i3139.

Appendix

Constraints that ought to be fulfilled when estimating values for individual cells of the contingency table. Numbers are derived from the hypothetical example. TP = true positive; FP = false positive; TN = true negative; FN = false negative; DOR = diagnostic odds ratio.

Estimated	Constraint	Observed	Example
TP + FP	=	∑Index test positive	TP + FP = 3
TN + FN	=	∑Index test negative	TN + FN = 3
TP + FN	=	∑Disease present	TP + FN = 3.2
TN + FP	=	∑Disease absent	TN + FP = 2.8
TP, FP, TN, FN	>=	0	∑TP = 2.4; ∑FP = 0.6; ∑FN = 0.8; ∑TN = 2.2
DOR	=	DOR (Logistic regression)	(TP / FP) / (FN / TN) = 4.5

"Not a single diagnosis has ever cured anyone"

Self-quoted

Chapter 4

Overdiagnosis across medical disciplines: a scoping review

K. Jenniskens

J.A.H. de Groot

J.B. Reitsma

K.G.M. Moons

L. Hooft

C.A. Naaktgeboren

BMJ Open. 2017 Dec 27;7(12):e018448

Abstract

Objective: To provide insight into how and in what clinical fields overdiagnosis is studied, and give directions for further applied and methodological research.

Design: Scoping review

Data sources: Medline up to August 2017

Study selection: All English studies on humans, in which overdiagnosis was discussed as a dominant theme.

Data extraction: Studies were assessed on clinical field, study aim (i.e. methodological or non-methodological), article type (e.g. primary study, review), the type and role of diagnostic test(s) studied, and the context in which these studies discussed overdiagnosis.

Results: From 4896 studies, 1851 were included for analysis. Half of all studies on overdiagnosis were performed in the field of oncology (50%). Other prevalent clinical fields included mental disorders, infectious diseases and cardiovascular diseases accounting for 9%, 8% and 6% of studies respectively. Overdiagnosis was addressed from a methodological perspective in 20% of studies. Primary studies were the most common article type (58%). The type of diagnostic tests most commonly studied were imaging tests (32%), although these were predominantly seen in oncology and cardiovascular disease (84%). Diagnostic tests were studied in a screening setting in 43% of all studies, but as high as 75% of all oncological studies. The context in which studies addressed overdiagnosis related most frequently to its estimation, accounting for 53%. Methodology on overdiagnosis estimation and definition provided a source for extensive discussion. Other contexts of discussion included definition of disease, overdiagnosis communication, trends in increasing disease prevalence, drivers and consequences of overdiagnosis, incidental findings and genomics.

Conclusions: Overdiagnosis is discussed across virtually all clinical fields and in different contexts. The variability in characteristics between studies and lack of consensus on overdiagnosis definition indicate the need for a uniform typology to improve coherence and comparability of studies on overdiagnosis.

Introduction

Overmedicalisation is the broad overarching term describing the use of "too much medicine". (1) It encompasses various concepts such as disease mongering, misdiagnosis, overutilization, overdetection and overtreatment. Initiatives relating to these concepts have begun to flourish on a global scale under the 'Choosing Wisely' initiative and in national programs such as Slow Medicine (Italy, the Netherlands and Brazil), Quaternary Prevention (Belgium) and Do not do (UK). (2, 3) A subcategory of the aforementioned concepts is overdiagnosis. This has become an even more popular term especially over the last two decades. (4-9) Furthermore, an annual conference going by the name of "Preventing Overdiagnosis", dedicated to issues surrounding this concept, has been gaining popularity ever since its start in 2013, demonstrating a growing interest in the topic. (10) In this scoping review we will focus specifically on overdiagnosis.

Defining overdiagnosis is challenging and diverse definitions exist. (11, 12) In a narrow sense, overdiagnosis describes individuals receiving a diagnosis with a condition that would never have become symptomatic before the end of the individual's life. (5, 7) However, overdiagnosis has also been described as giving a diagnosis that would not yield a net benefit. (1) These definitions are not similar, and thus may lead to different interpretations of (the extent of) overdiagnosis. Consequently, the mechanisms leading to overdiagnosis may also differ. Labelling an individual with a blood pressure over a certain threshold as hypertensive, and thus "diseased", is conceptually different than not knowing whether one should diagnose an individual with a very small potentially malignant growth as having cancer, and thus "diseased". Providing definitions in combination with mechanisms of overdiagnosis for a typology is challenging and source of extensive discussion. (13-17)

The range of overdiagnosis drivers is also extensive. It, amongst others, includes technological developments that detect smaller abnormalities than ever before which might not become clinically manifest. Furthermore, the use of large scale screening programs, inappropriate application of diagnostic criteria, legal incentives, cultural believes (i.e. that we should do

everything in our power to find and treat disease) and commercial or professional interests have driven overdiagnosis. (6, 18-20)

Consequences of overdiagnosis may be serious and can be subdivided in negative effects on patient health and additional costs within the health care system. (21) Health effects include impaired quality of life and early loss of life due to side-effects or complications of unnecessary subsequent testing or treatment. Incorrectly labelling of individuals as patients may also lead to stigmatization, impacting psychological well-being and indirectly exert social effects through eligibility for health benefits. In monetary terms, overdiagnosis can result in unwarranted usage of (follow-up) tests, treatment and healthcare facilities and services.

Despite the increasing number of publications on overdiagnosis, ranging from discussions on overdiagnosis definition to estimating its impact, a scoping analysis on overdiagnosis is still lacking. In the present study, we provide an overview of research that has been performed across medical disciplines surrounding the topic of overdiagnosis. Not only will we give insight into how and in what clinical fields overdiagnosis is studied, but also provide directions for further applied and methodological research to investigate the mechanisms and impact of overdiagnosis, and to generate directions for reducing or preventing overdiagnosis.

Methods

PubMed was searched on August 2017 for published articles using keywords related to overdiagnosis, overdetection, overscreening, insignificant disease, overtesting, overmedicalisation, pseudodisease, inconsequential disease, and quaternary prevention, by using the following query:

overdiagnos[tw] OR over diagnos*[tw] OR overdetect*[tw] OR over detect*[tw] OR "insignificant disease"[tw] OR overscreen*[tw] OR over screen*[tw] OR overtest*[tw] OR over test*[tw] OR overmedical*[tw] OR over medical*[tw]OR "pseudodisease"[tw] OR "pseudo disease"[tw] OR "inconsequential disease"[tw] OR "Quaternary prevention"[tw]*

These terms were chosen as they were believed to capture most concepts related to overdiagnosis, generating a representative set of articles. All English articles on humans where the full text was available were included. Articles in which overdiagnosis was a dominant theme were included. Overdiagnosis was considered a dominant theme when a paper clearly addressed overdiagnosis as an issue being investigated or discussed. For example, a study on the adoption of a new threshold guideline for prostate-specific antigen screening for prostate cancer was considered to have a dominant overdiagnosis theme. In contrast, a study that used overdiagnosis as a buzzword and merely suggested in the discussion that overdiagnosis might possibly play a role or have occurred, was excluded. Studies with overdiagnosis as a dominant theme were included regardless of which definition of overdiagnosis the authors adopted.

The titles and abstracts of the included studies were then screened. Included studies were assessed using (a list of) prespecified criteria. These criteria were established by screening the first 200 studies of the search query. They included clinical field, study aim, article type, type of diagnostic test, whether this was a screening test, and the context in which overdiagnosis was discussed. These criteria are described below (see further details in the Appendix). Articles were assessed based solely on title and abstract. If an abstract was unavailable (e.g. opinion pieces), the full text was scanned.

Clinical field

The clinical field to which the study belonged was determined using the ICD-10 classification. When a study addressed more than one clinical field or did not address overdiagnosis within a specific clinical field, but discussed overdiagnosis on a more general level, they were included in the separate category "No specific clinical field".

Study aim

Two study aims were distinguished: 1) studies focusing on *how* overdiagnosis should be studied. These are studies with a methodological aim. Examples are studies looking into how overdiagnosis estimations are affected by the methods used, or studies providing a framework for the definition of overdiagnosis. Simulation studies using mathematical models for estimating

the extent of overdiagnosis were also classified as methodological studies. Studies not addressing the aforementioned concepts, but rather provide, for example, a qualitative overview of the (possible) impact of overdiagnosis in a certain field, or calculate overdiagnosis estimates from empirical data, were considered to have 2) a non-methodological aim.

Article type

Studies were classified using four article types: primary studies, narrative reviews, systematic reviews or commentaries. Primary studies used data collected from trials, observational studies or generated using simulation models. Narrative reviews described a broad oversight on overdiagnosis. These included editorials, opinion pieces, interviews and overviews. Systematic reviews stated a specific hypothesis and tested this using a systematic approach to gather existing literature. If a systematic approach was lacking, these studies were scored as narrative reviews. Studies were considered commentaries when they, replied to previously published papers.

Type of diagnostic test

Diagnostic tests were categorized into six types: imaging, medical examination, biomarker, histology, prediction model or various. Whenever a study looked into a combination of two tests, both types were scored. For example, an image guided biopsy would be scored as both an imaging and histologic diagnostic test. If three or more diagnostic tests were addressed within a study, or overdiagnosis was addressed in a general context without any diagnostic test in particular, this was scored under "Various tests".

Screening

When studies focused on a test used for screening groups of asymptomatic individuals, this was scored as a screening study. Studies that did not explicitly state that the diagnostic test was studied in the context of screening, were scored as a non-screening.

Overdiagnosis context

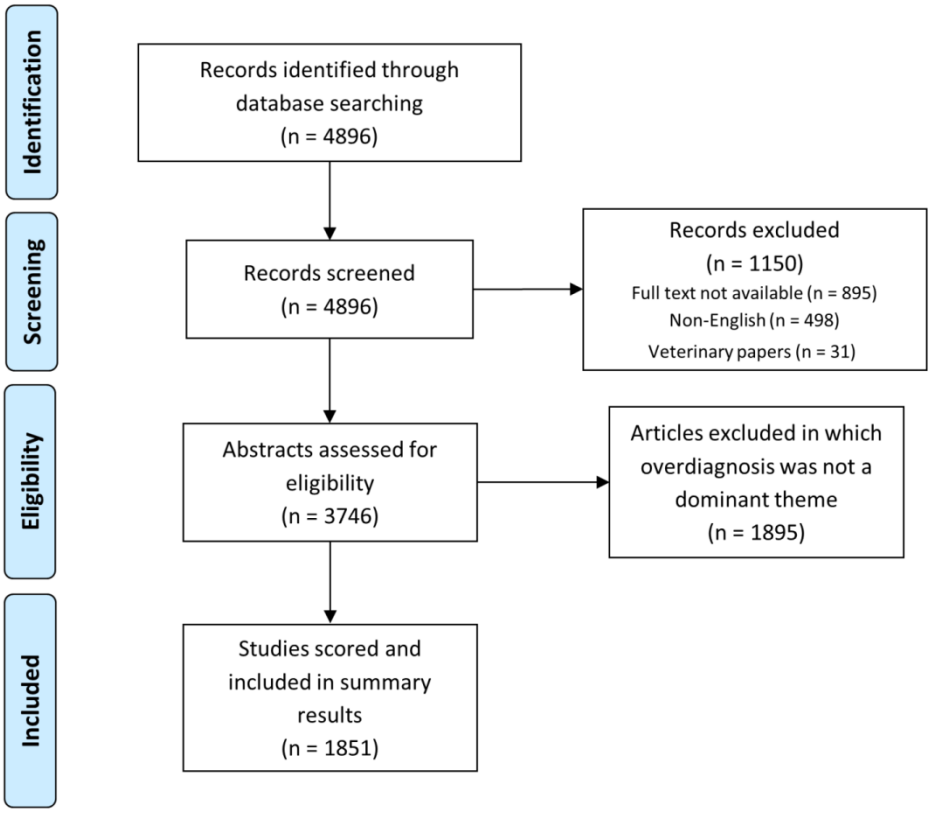
To assess the context in which studies discussed overdiagnosis five categories were defined: estimating extent of overdiagnosis, disease definition, overdiagnosis communication, incidental findings, and genomics.

The first category, estimating extent of overdiagnosis, relates to all articles giving a quantified estimate of overdiagnosis. Disease definition revolves around the setting of thresholds to define the absence or presence of a disease or to distinguish between two subcategories of a certain disease (e.g. progressive and non-progressive forms). Overdiagnosis communication relates to studies aimed at assessing and improving the understanding of overdiagnosis in the general public, and improving overdiagnosis dissemination by the healthcare professionals. Studies addressing abnormalities found of an unrelated condition during either diagnostic testing or surgery were scored as studies on incidental findings. Spurious findings on genome wide screening tests were scored in the overdiagnosis context of genomics.

Results

The PubMed search resulted in a total number of 4896 studies identified. After application of the inclusion criteria 3746 studies were assessed for eligibility on title and abstract. Studies in which overdiagnosis was a dominant theme yielded 1851 studies. (Figure). Table 1 provides a summarized view of the characteristics of the total number of studies, the four largest clinical fields, all other remaining clinical fields and studies not related to a specific clinical field.

Figure. Flow diagram of article selection for further review and scoring



Clinical field

Papers on overdiagnosis were found in all clinical fields, but were mainly published within oncology (50%), in which breast (34%), prostate (24%) and lung cancer (14%) ranked as most prevalently studied. Other clinical fields addressing overdiagnosis included mental disorders (9%), infectious diseases (8%) and cardiovascular disease (6%). Within these fields, studies were predominantly looking into bipolar disorder, malaria and pulmonary embolism (PE), respectively. (22-27)

Study aim

Studies addressing methodological issues consisted of 20%. The majority of these studies were performed within the field of oncology. However, non-methodological studies were the most common study aim used across

Table 1. Characteristics of papers in which overdiagnosis was a dominant them, with results shown for the total number of articles, the four largest clinical fields and studies not addressing a specific clinical field

	Total (n = 1851)	Oncological disorders (n = 920)	Mental disorders (n =171)	Infectious diseases (n = 143)	Cardiovascular disorders (n = 105)	Other clinical fields (n = 390)	No specific clinical field (n = 122)
Study aim							
Methodological	**20%**	30%	11%	4%	10%	4%	34%
Non-methodological	**80%**	70%	89%	96%	90%	96%	66%
Article type							
Primary study	**58%**	55%	53%	85%	61%	69%	27%
Narrative review	**24%**	22%	32%	9%	24%	22%	52%
Systematic review	**9%**	12%	8%	1%	10%	5%	11%
Commentary	**9%**	11%	8%	6%	6%	4%	10%
Diagnostic test							
Imaging	**32%**	48%	3%	4%	47%	19%	7%
Medical examination	**17%**	3%	58%	26%	26%	30%	4%
Biomarker	**15%**	16%	3%	29%	10%	16%	3%
Histology	**13%**	17%	0%	21%	2%	11%	2%
Prediction model	**3%**	4%	1%	2%	3%	4%	1%
Various	**21%**	13%	35%	18%	12%	20%	84%

	Total (n = 1851)	Oncological disorders (n = 920)	Mental disorders (n =171)	Infectious diseases (n = 143)	Cardiovascular disorders (n = 105)	Other clinical fields (n = 390)	No specific clinical field (n = 122)
Screening							
Yes	**43%**	75%	5%	10%	15%	10%	20%
No	**57%**	25%	95%	90%	85%	90%	80%
Overdiagnosis context							
Overdiagnosis estimation	**53%**	57%	22%	63%	65%	60%	16%
Disease definition	**15%**	8%	46%	13%	14%	22%	8%
Overdiagnosis communication	**3%**	5%	2%	0.7%	0%	0.8%	3%
Incidental findings	**0.8%**	0.8%	0%	0%	1%	1%	2%
Genomics	**0.4%**	0.3%	0%	0%	1%	0%	3%
Other*	**28%**	29%	30%	24%	19%	16%	67%

**Subcategories in this category include: overdiagnosis definition, drivers and consequences of overdiagnosis, and trend studies suggesting overdiagnosis*

all clinical fields, accounting for 80% of the total number of articles. These notably included studies using empirical data to assess the occurrence or estimate overdiagnosis for a specific disease.

Article type

Primary studies (58%) were the most common article type discussing overdiagnosis. Of all included studies narrative, systematic reviews and commentaries represented 24%, 9% and 9% respectively. From all studies that addressed a specific clinical field, the proportion of systematic reviews and commentaries was relatively high within oncology.

Type of diagnostic test

Imaging was the most often encountered diagnostic test, accounting for 32% of all studies. Biomarkers (15%), histology (13%) and medical examination (17%) were approximately equally often found. Prediction models were less common (3%). The proportion not related to one particular diagnostic test of interest was 21%. Distributions of diagnostic tests varied significantly depending on the clinical field. Imaging was most prevalent in oncology where it accounted for 48% of diagnostic tests, mostly related to breast (53%) and lung cancer screening (21%). Within the field of mental disorders medical examination was often seen in the form of application of the DSM (Diagnostic and Statistical Manual of Mental Disorders) as diagnostic tool. Biomarkers and histology were seen relatively more frequent as diagnostic tests for infectious diseases when compared to other clinical fields.

Screening

Diagnostic testing was studied in the context of screening in 43% of studies. There was however a skewed distribution between clinical fields. Within oncology, 75% of all studies were related to screening, whereas for mental disorders, infectious diseases and cardiovascular diseases this was 15% or lower.

Overdiagnosis context

The context in which overdiagnosis was most frequently discussed related to its estimation (53%). Only within the field of mental disorders was disease definition more frequently discussed than overdiagnosis estimation (46% vs

22%). Descriptions and example studies on each of the five predefined categories can be found in table 2. The majority of studies discussing overdiagnosis (72%) were classifiable in one of these categories. Studies that did not fall within any of the five categories were scored in a separate "Other" category (28%). Results for each of these overdiagnosis contexts are discussed below.

Table 2. Descriptions and examples of context of overdiagnosis discussion

Overdiagnosis context	Description	Example	Ref.
Overdiagnosis estimation	Providing a quantitative estimate of overdiagnosis	*Estimation of overdiagnosis in low-dose computed tomography screening for lung cancer*	(28)
Disease definition	Setting thresholds to define the absence or presence of a disease, or distinguishing between two subcategories within a disease	*Current definitions of airflow obstruction and attention deficit hyperactivity disorder yield overdiagnosis in primary care*	(29)
Overdiagnosis communication	Assessing and improving the understanding of overdiagnosis in the general public, and improving overdiagnosis dissemination by the healthcare professionals	*Assessing what the general public thinks is meant by the term 'overdiagnosis'*	(30)
Incidental findings	An abnormality found of an unrelated condition during either diagnostic testing or surgery	*Relevance of incidental findings when screening for a disorder in the abdominal area using multi-detector contrast-enhanced CT*	(31)
Genomics	Spurious genetic abnormalities	*Implications of genetic screening for common cancers in children*	(32)

Overdiagnosis estimation

The most common context of discussion relates to overdiagnosis estimation, accounting for 53% of all studies. These articles could be divided into two groups. The first were studies attempting to estimate the degree of overdiagnosis in their respective clinical fields. (79%) These often described the impact of implementation or a threshold shift of a diagnostic or screening intervention on the rate of overdiagnosis. Notable examples of this are prostate-specific antigen testing for prostate cancer and mammography for breast cancer. (33-38) However several articles estimated overdiagnosis in symptomatic conditions, such as incorrect diagnosis by untrained clinicians in patients presenting with malaria-like symptoms, leading to false-positives and unnecessary treatment. (26, 27) This should rather be considered misdiagnosis (incorrect diagnosis of a symptomatic person with a condition they do not have (1)) due to inaccuracy of clinical tests used in practice leading to false-positives, incorrect disease labels, and overtreatment. The second group represented studies that report methodological approaches for *how* one should estimate overdiagnosis. (21%) Differences regarding definitions used, measurement, study design and methods for estimation can lead to different results (39), hence there is often a large spread in these estimates, resulting in controversy regarding the true impact of overdiagnosis in the field.

Disease definition

In 15% of all studies disease definition was addressed. A relatively high proportion of these studies was addressed in the context of mental disorders (28%). Common topics included application of DSM for bipolar disorder, depression and attention deficit hyperactivity disorder, (40, 41) and physician diagnosis of attention deficit hyperactivity disorder or asthma, which were related to misdiagnosis rather than actual overdiagnosis. (42-44) The other major contributor was in oncology (25%), where the main issue was the transition of benign to malignant growths. Examples of such pre-disease conditions are ductal carcinoma in situ, early stage prostate tumours and papillary thyroid carcinoma. (45-47)

Overdiagnosis communication

Communication about overdiagnosis with patients or the public accounted for 3% of all 1851 publications. This mainly involved the people's understanding of the concept of overdiagnosis, and whether they perceived it to be an issue. (30, 48, 49) Other articles dealt with communication of overdiagnosis between the patient and the treating physician, (50, 51) or the development and effectiveness of decision aids. (52, 53)

Other contexts

Scientific literature on overdiagnosis in genomics and incidental findings were found only sporadically (0.4% and 0.8%). The term overmedicalisation was frequently used in literature to describe medicalisation of normal life events, such as birth, adolescence and death. Quaternary prevention was mostly used to describe the action being taken to prevent overmedicalisation. One of the most commonly observed topics in the other category was drivers and consequences of overdiagnosis. (18, 21, 54, 55) These were often mentioned alongside in narrative reviews on overdiagnosis. Furthermore, trend studies were common, describing the possibility of overdiagnosis based on a rapid increase in the number of diagnoses, without any significant decrease in the mortality rate. These studies did not provide an exact overdiagnosis estimate, but rather an indication that overdiagnosis might be occurring or increasing, based on historic data. Another context in which overdiagnosis was commonly addressed, especially in the last couple of years, was its definition. These studies aim at formulating accurate and appropriate definitions of overdiagnosis as well as related terminology (e.g. overmedicalisation, overdetection, disease mongering). In addition, some have attempted defining broad overall classifications to provide guidance for distinction between different overdiagnosis subtypes. (13, 16)

Discussion

This scoping review provides insight in the current landscape of overdiagnosis. There is great diversity in study characteristics across medical disciplines and in the contexts in which overdiagnosis is discussed. Some characteristics correlate with specific clinical fields, with, for example,

screening occurring predominantly in oncological studies and medical examination being the most prevalently used diagnostic test for mental disorders.

Overdiagnosis is discussed in a variety of contexts, however three could be distinguished which invoked significant debate: 1) differences in overdiagnosis definition, 2) differences in methods used, leading to varying overdiagnosis estimates, and 3) typologies for overdiagnosis.

Overdiagnosis definitions

The definition of overdiagnosis has been topic of discussion for some time. In a narrow sense it refers to a diagnosis that does not result in a net benefit for an individual. (1) This can be viewed within an individual or on a group level, where benefits (early detection of clinically relevant disease) are weighted against the deficits (overdiagnosis and its associated consequences). However, not all included studies give a clear definition, but implicitly use the definition of overdiagnosis as a diagnosis of a "disease" in an asymptomatic individual, that will never go on to cause symptoms or early death. (7) This definition is particular to the screening-context, but does not apply to a large portion of the studies found in this review that are on testing symptomatic individuals, for example those with mental disorders. Others have used the relation between pathology and symptoms as a measure of overdiagnosis. (56, 57) In the latter there is no doubt there is a clear abnormality, however it is uncertain whether smaller forms of this abnormality still significantly correlate with future clinically relevant disease. Ultimately, the question would be how or even if we should treat these individuals. These examples of definitions demonstrate the heterogeneity and complexity of the concept of overdiagnosis, and have led to the discussion regarding the extent or even the existence of overdiagnosis. Which definition researchers use for overdiagnosis needs to be reported completely to be able to judge the applicability of the results.

Methods for overdiagnosis estimation

Another discussion revolves around variation in estimates of overdiagnosis. Major trials such as the European Randomized Study of Screening for Prostate Cancer (ERSPC), the National Lung Screening Trial (NLST), the

Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, and the Malmö breast cancer screening trial, often form the basis for these discussions. (58-61) These trials look into the effects of cancer screening programs. The ERSPC did not provide an overdiagnosis in prostate cancer screening in their initial publication (62), but did provide an estimate of 41% in their 2014 publication. (58) However, this was obtained through modelling, and not calculated directly from the observed data. The NLST merely states that overdiagnosis is presumably not large, as the number of breast cancers diagnosed between the two screening arms is comparable. (59) And the PLCO and Malmö breast cancer screening trials did not state anything about overdiagnosis. (60, 61) The scientific community reacted by using different methods to provide overdiagnosis estimates for these trials. The rate of overdiagnosis that is estimated depends on various features such as the definitions and measurements used, study design and context and estimation approaches applied. (12, 39, 63-67) The latter can be divided in lead-time (the time between screening detection and clinical presentation) and excess incidence approach (excess number of cases between a screening and non-screening group), each of which has its merits and issues, and requires assumptions to be made. Ultimately, the variety in methodology used has resulted in variation in overdiagnosis estimates, and significant controversy between studies. (11, 67, 68)

Overdiagnosis typologies

Several studies have provided overviews and acknowledged that finding a singular definition of overdiagnosis may not be feasible. However, providing an overdiagnosis classification, aimed at describing subtypes of overdiagnosis, could prove to be useful. Some efforts have been made to create such a typology, however this is challenging as definitions vary widely and classifications can be made over different axes. Hence, this is a complex issue which should be addressed in a systematic manner. A comprehensive typology could aid researchers in their communication as was already suggested in a paper by Moynihan et al in 2012. (6) A recent paper by Rogers described the use of maldetection (issues with our understanding of what 'truly' disease is) and misclassification (an implicit or explicit threshold shift resulting in overdiagnosis). (13) Shortly after, Carter et al described the

concepts of predatory, tragic and misdirected overdiagnosis. (17) Other work by Hofmann takes a more sociological and philosophical point of view. In his 2017 publication, indicative, measurable and observable phenomena are used to describe the different stages in which a phenomenon develops into a clinical manifestation. (16) In oncology a tumour-patient classification has been described, relating to tumours that are regressive, non-progressive or truly malignant disease. (69) Although these works provide great improvement in our understanding of the issues at hand, they do not give further guidance as to how these concepts should be used in clinical research.

To our knowledge, this is the first scoping review performed on the subject of overdiagnosis. It provides broad insight in the available research on specific topics within overdiagnosis. To appreciate the findings in this review, the following limitations should be considered. First, studies were excluded when they did not have full text available. This may have led to exclusion of a selection of relevant articles, but not a systematic exclusion of a particular range of overdiagnosis studies. The same holds true for the lack of search criteria for iatrogenic disease, overtreatment, and overutilisation. The issue in identifying studies discussing overdiagnosis, is that there are no clear selection criteria to find these. Terminologies used to describe overdiagnosis differ between studies, are widely spread and search filters in medical databases are lacking. Hence, our goal was not to perform a comprehensive search. Instead, we aimed at finding a large representative of papers discussing overdiagnosis.

Second, unexpectedly, studies on genomics and incidental findings (or incidentalomas) were largely missed. Forward reference checking revealed that some of the papers not found in our search may use other terminology for describing overdiagnosis, such as the "prevalence of significant findings" or "diagnostic value". Using our search strategy these articles were unfortunately omitted and not included in this review. When researchers are interested particularly in this subset, the information in this review might not suffice.

In summary, overdiagnosis is a topic discussed over medical disciplines, and in a wide array of contexts, from conceptual ideas in definition to practical

issues for clinicians in daily practice. The various characteristics of studies looking at overdiagnosis suggest that there may be different (and sometimes multiple) underlying mechanisms through which it may manifest itself. A lack of consensus on what is called overdiagnosis hampers communication between researchers, physicians, patients, and policy makers. The use of overdiagnosis to describe misdiagnosis will dilute its actual meaning, result in linguistic confusion, and counterproductive discussion, and should thus be avoided. Providing clarity on the mechanisms that lead to overdiagnosis will aid researchers communicate their results, especially with regard to overdiagnosis estimates. Future methodological studies should focus on establishing a framework to aid clinicians and researchers in understanding the different subtypes of overdiagnosis, their consequences, and provide guidance for selecting appropriate study designs and methods that match the research question of interest.

References

1. Carter SM, Rogers W, Heath I, Degeling C, Doust J, Barratt A. The challenge of overdiagnosis begins with its definition. BMJ. 2015;350:h869.

2. ABIM foundation. Choosing Wisely Around the World 2015 [04-01-2017]. Available from: http://www.choosingwisely.org/resources/updates-from-the-field/choosing-wisely-around-the-world/.

3. Otte JA. Less is More Medicine [09-05-2017]. Available from: http://www.lessismoremedicine.com/projects/.

4. Welch GH. Overdiagnosed: Making People Sick in the Pursuit of Health2010.

5. Black WC. Overdiagnosis: An underrecognized cause of confusion and harm in cancer screening. J Natl Cancer Inst. 2000;92(16):1280-2.

6. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. BMJ. 2012;344:e3502.

7. Welch HG, Black WC. Overdiagnosis in cancer. J Natl Cancer Inst. 2010;102(9):605-13.

8. Etzioni R, Penson DF, Legler JM, di Tommaso D, Boer R, Gann PH, et al. Overdiagnosis due to prostate-specific antigen screening: lessons from U.S. prostate cancer incidence trends. J Natl Cancer Inst. 2002;94(13):981-90.

9. Pohl H, Welch HG. The role of overdiagnosis and reclassification in the marked increase of esophageal adenocarcinoma incidence. J Natl Cancer Inst. 2005;97(2):142-6.

10. Preventing Overdiagnosis Conference [04-01-2014]. Available from: http://www.preventingoverdiagnosis.net/.

11. Bae JM. Overdiagnosis: epidemiologic concepts and estimation. Epidemiol Health. 2015;37:e2015004.

12. Bach PB. Overdiagnosis in lung cancer: different perspectives, definitions, implications. Thorax. 2008;63(4):298-300.

13. Rogers WA, Mintzker Y. Getting clearer on overdiagnosis. J Eval Clin Pract. 2016;22(4):580-7.

14. Hofmann BM. Conceptual overdiagnosis. A comment on Wendy Rogers and Yishai Mintzker's article "Getting clearer on overdiagnosis". J Eval Clin Pract. 2016.

15. Rogers WA, Mintzker Y. Response to Bjorn Hofmann: Clarifying overdiagnosis without losing conceptual complexity. J Eval Clin Pract. 2016.

16. Hofmann B. Defining and evaluating overdiagnosis. J Med Ethics. 2016.

17. Carter SM, Degeling C, Doust J, Barratt A. A definition and ethical evaluation of overdiagnosis. J Med Ethics. 2016.

18. Paris J, Bhat V, Thombs B. Is Adult Attention-Deficit Hyperactivity Disorder Being Overdiagnosed? Can J Psychiatry. 2015;60(7):324-8.

19. Pathirana T, Clark J, Moynihan R. Mapping the drivers of overdiagnosis to potential solutions. BMJ. 2017;358:j3879.

20. Hofmann BM. Too much technology. BMJ. 2015;350:h705.

21. Doust J, Glasziou P. Is the problem that everything is a diagnosis? Aust Fam Physician. 2013;42(12):856-9.

22. Winters BS, Solarz M, Jacovides CL, Purtill JJ, Rothman RH, Parvizi J. Overdiagnosis of pulmonary embolism: evaluation of a hypoxia algorithm designed to avoid this catastrophic problem. Clin Orthop Relat Res. 2012;470(2):497-502.

23. Suh JM, Cronan JJ, Healey TT. Dots are not clots: the over-diagnosis and over-treatment of PE. Emerg Radiol. 2010;17(5):347-52.

24. Bruchmuller K, Margraf J, Schneider S. Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. J Consult Clin Psychol. 2012;80(1):128-38.

25. Bonati M, Reale L. Reducing overdiagnosis and disease mongering in ADHD in Lombardy. BMJ. 2013;347:f7474.

26. Harchut K, Standley C, Dobson A, Klaassen B, Rambaud-Althaus C, Althaus F, et al. Over-diagnosis of malaria by microscopy in the Kilombero Valley, Southern Tanzania: an evaluation of the utility and cost-effectiveness of rapid diagnostic tests. Malar J. 2013;12:159.

27. Mwanziva C, Shekalaghe S, Ndaro A, Mengerink B, Megiroo S, Mosha F, et al. Overuse of artemisinin-combination therapy in Mto wa Mbu (river of mosquitoes), an area misinterpreted as high endemic for malaria. Malar J. 2008;7:232.

28. Patz EF, Jr., Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemagi MC, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. JAMA Intern Med. 2014;174(2):269-74.

29. Schermer TR, Smeele IJ, Thoonen BP, Lucas AE, Grootens JG, van Boxem TJ, et al. Current clinical guideline definitions of airflow obstruction and COPD overdiagnosis in primary care. Eur Respir J. 2008;32(4):945-52.

30. Moynihan R, Nickel B, Hersch J, Doust J, Barratt A, Beller E, et al. What do you think overdiagnosis means? A qualitative analysis of responses from a national community survey of Australians. BMJ Open. 2015;5(5):e007436.

31. Sconfienza LM, Mauri G, Muzzupappa C, Poloni A, Bandirali M, Esseridou A, et al. Relevant incidental findings at abdominal multi-detector contrast-enhanced computed tomography: A collateral screening? World J Radiol. 2015;7(10):350-6.

32. Hall AE, Chowdhury S, Pashayan N, Hallowell N, Pharoah P, Burton H. What ethical and legal principles should guide the genotyping of children as part of a personalised screening programme for common cancer? J Med Ethics. 2014;40(3):163-7.

33. Pelzer AE, Colleselli D, Bektic J, Schaefer G, Ongarello S, Schwentner C, et al. Over-diagnosis and under-diagnosis of screen- vs non-screen-detected prostate cancers with in men with prostate-specific antigen levels of 2.0-10.0 ng/mL. BJU Int. 2008;101(10):1223-6.

34. Heijnsdijk EA, de Carvalho TM, Auvinen A, Zappa M, Nelen V, Kwiatkowski M, et al. Cost-effectiveness of prostate cancer screening: a simulation study based on ERSPC data. J Natl Cancer Inst. 2015;107(1):366.

35. Arnsrud Godtman R, Holmberg E, Lilja H, Stranne J, Hugosson J. Opportunistic testing versus organized prostate-specific antigen screening: outcome after 18 years in the Goteborg randomized population-based prostate cancer screening trial. Eur Urol. 2015;68(3):354-60.

36. Beckmann K, Duffy SW, Lynch J, Hiller J, Farshid G, Roder D. Estimates of over-diagnosis of breast cancer due to population-based mammography screening in South Australia after adjustment for lead time effects. J Med Screen. 2015;22(3):127-35.

37. Seigneurin A, Labarere J, Francois O, Exbrayat C, Dupouy M, Filippi M, et al. Overdiagnosis and overtreatment associated with breast cancer mammography screening: A simulation study with calibration to population-based data. Breast. 2016;28:60-6.

38. Gunsoy NB, Garcia-Closas M, Moss SM. Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the United Kingdom. Br J Cancer. 2014;110(10):2412-9.

39. Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. Ann Intern Med. 2013;158(11):831-8.

40. Phelps J, Ghaemi SN. The mistaken claim of bipolar 'overdiagnosis': solving the false positives problem for DSM-5/ICD-11. Acta Psychiatr Scand. 2012;126(6):395-401.

41. Sciutto MJ, Eisenberg M. Evaluating the evidence for and against the overdiagnosis of ADHD. J Atten Disord. 2007;11(2):106-13.

42. Garcia-Rio F, Soriano JB, Miravitlles M, Munoz L, Duran-Tauleria E, Sanchez G, et al. Overdiagnosing subjects with COPD using the 0.7 fixed ratio: correlation with a poor health-related quality of life. Chest. 2011;139(5):1072-80.

43. Guder G, Brenner S, Angermann CE, Ertl G, Held M, Sachs AP, et al. "GOLD or lower limit of normal definition? A comparison with expert-based diagnosis of chronic obstructive pulmonary disease in a prospective cohort-study". Respir Res. 2012;13(1):13.

44. Aaron SD, Vandemheen KL, Boulet LP, McIvor RA, Fitzgerald JM, Hernandez P, et al. Overdiagnosis of asthma in obese and nonobese adults. CMAJ. 2008;179(11):1121-31.

45. Evans AJ, Pinder SE, Ellis IO, Wilson AR. Screen detected ductal carcinoma in situ (DCIS): overdiagnosis or an obligate precursor of invasive disease? J Med Screen. 2001;8(3):149-51.

46. Van der Kwast TH, Roobol MJ. Defining the threshold for significant versus insignificant prostate cancer. Nat Rev Urol. 2013;10(8):473-82.

47. Vaccarella S, Dal Maso L, Laversanne M, Bray F, Plummer M, Franceschi S. The Impact of Diagnostic Changes on the Rise in Thyroid Cancer Incidence: A Population-Based Study in Selected High-Resource Countries. Thyroid. 2015;25(10):1127-36.

48. Hersch J, Jansen J, Barratt A, Irwig L, Houssami N, Howard K, et al. Women's views on overdiagnosis in breast cancer screening: a qualitative study. BMJ. 2013;346:f158.

49. Moynihan R, Nickel B, Hersch J, Beller E, Doust J, Compton S, et al. Public Opinions about Overdiagnosis: A National Community Survey. PLoS One. 2015;10(5):e0125165.

50. van Agt H, Fracheboud J, van der Steen A, de Koning H. Do women make an informed choice about participating in breast cancer screening? A survey among women invited for a first mammography screening examination. Patient Educ Couns. 2012;89(2):353-9.

51. Wegwarth O, Gigerenzer G. Less is more: Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. JAMA Intern Med. 2013;173(22):2086-7.

52. Bae JM. Development and application of patient decision aids. Epidemiol Health. 2015;37:e2015018.

53. Hersch J, Barratt A, Jansen J, Irwig L, McGeechan K, Jacklyn G, et al. Use of a decision aid including information on overdetection to support informed choice about breast cancer screening: a randomised controlled trial. Lancet. 2015;385(9978):1642-52.

54. Day M. Drug industry is partly to blame for overdiagnosis of bipolar disorder, researchers claim. BMJ. 2008;336(7653):1092-3.

55. Carneiro AV. Screening for coronary artery disease in assymptomatic adults is not recommended, so why is it still done? Rev Port Cardiol. 2004;23(12):1633-8.

56. Hoffman JR, Carpenter CR. Guarding Against Overtesting, Overdiagnosis, and Overtreatment of Older Adults: Thinking Beyond Imaging and Injuries to Weigh Harms and Benefits. J Am Geriatr Soc. 2017.

57. de Roos MA, van der Vegt B, de Vries J, Wesseling J, de Bock GH. Pathological and biological differences between screen-detected and interval ductal carcinoma in situ of the breast. Ann Surg Oncol. 2007;14(7):2097-104.

58. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Zappa M, Nelen V, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. Lancet. 2014;384(9959):2027-35.

59. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395-409.

60. Andriole GL, Crawford ED, Grubb RL, 3rd, Buys SS, Chia D, Church TR, et al. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. J Natl Cancer Inst. 2012;104(2):125-32.

61. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ. 1988;297(6654):943-8.

62. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009;360(13):1320-8.

63. Wu D, Perez A. A limited Review of Over Diagnosis Methods and Long Term Effects in Breast Cancer Screening. Oncol Rev. 2011;5(3):143-7.

64. Duffy SW, Lynge E, Jonsson H, Ayyaz S, Olsen AH. Complexities in the estimation of overdiagnosis in breast cancer screening. Br J Cancer. 2008;99(7):1176-8.

65. de Gelder R, Heijnsdijk EA, van Ravesteyn NT, Fracheboud J, Draisma G, de Koning HJ. Interpreting overdiagnosis estimates in population-based mammography screening. Epidemiol Rev. 2011;33:111-21.

66. Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulati R, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J Natl Cancer Inst. 2009;101(6):374-83.

67. Puliti D, Miccinesi G, Paci E. Overdiagnosis in breast cancer: design and methods of estimation in observational studies. Prev Med. 2011;53(3):131-3.

68. Davidov O, Zelen M. Overdiagnosis in early detection programs. Biostatistics. 2004;5(4):603-13.

69. Marcus PM, Prorok PC, Miller AB, DeVoto EJ, Kramer BS. Conceptualizing overdiagnosis in cancer screening. J Natl Cancer Inst. 2015;107(4).

Appendix

Criteria used for scoring of articles, and a description of specific in- and exclusion criteria per item

Criterion	Outcome	Description
Full-text available	Yes / No	Is a full-text available from PubMed?
Veterinary study	Yes / No	Is the paper a study with animals?
Overdiagnosis as a dominant theme	Yes / No	Is overdiagnosis discussed as a specific dominant theme **Include:** Prognostic / prediction studies relating to disease progression **Include:** Trend studies. Index test will often be not addressed **Include:** Active surveillance studies that assess what the impact is of having an in-between category, next to treat and do not treat **Exclude:** Studies in which no diagnostic method is evaluated **Exclude:** Erratum's **Exclude:** Case-studies (n = < 10) **Exclude:** Overview articles without a specific focus on diagnostics **Exclude:** Articles not mentioning overdiagnosis or only briefly commenting on it (particularly in the discussion) Example: Exclude article which states: "When Diagnostic test X is replaced with Diagnostic test Y sensitivity and specificity may be improved. As a result overdiagnosis of Disease Z may be reduced"
Clinical field	Bone & connective tissue	Examples: Myopathy, osteoporosis, dental problems
	Cancer	Examples: Prostate cancer, breast cancer, leukaemia **Exclude:** cervical cancer caused by HPV (=infection)
	Cardiovascular	Examples: Pulmonary embolism, angina
	Congenital	Examples: Down syndrome, hypospadias

Ear	Example: Tinnitus
Eye	Example: Gingivitis
Gastrointestinal	Examples: Crohn's disease, reflux disease, liver failure
Gynaecology & Obstetrics	Example: Preeclampsia
Immune system	Examples: Allergic reactions, autoimmune disorders, Heparin induced thrombocytopenia (HIT), PANDA's, Rheumatoid arthritis
Infection	Examples: Malaria, HIV, HPV, Clostridium difficile, pneumonia
Mental	Examples: ADHD, autism, depression, schizophrenia, bipolar disorder, (vascular) dementia **Include:** Diseases that are primarily psychiatric disorders and often result in impaired cognitive function **Exclude:** See neurological disorders
Metabolic	Examples: Diabetes, hypogonadism, hypothyroidism, growth related 'disorders', nutrition status
Neurological	Example: Multiple sclerosis, Parkinson's, Alzheimer **Include:** Diseases of the central / peripheral nervous system, that often have motorial implications **Exclude:** See mental disorders
Perinatal	Example: Malnutrition of the unborn child, child specific problems during pregnancy **Include:** disease in the unborn child
Respiratory	Examples: COPD, asthma, nasal disorders
Skin	Example: Eczema
Trauma	Examples: Car accidents, cuts, fractures, sprains, injury during surgery
Urogenital	Examples: Chronic kidney failure, kidney stones
No specific clinical field	Multiple clinical domains are assessed **OR** it is unclear if the paper focusses on a specific clinical domains

		Example: a methodological paper on how we should quantify overdiagnosis
Study aim	Methodological	Papers describing a theoretical framework for assessing overdiagnosis **Include:** Commentaries discussing the way overdiagnosis was determined in a different empirical primary study **Include:** Combination papers; Papers that are empirical, but also have a strong methodological focus on overdiagnosis **Include:** Modelling studies
	Non-methodological	Results from a primary study or assessment of outcomes by a review / overview paper
Article type	Commentary	A comment, reply or rebuttal on a previously published paper or commentary
	Narrative review	A paper giving a broad oversight of a specific topic, often from one particular authors view **Include:** editorials **Include:** opinion pieces **Include:** interviews **Include:** guidelines **Exclude:** Overviews that only refer to 1 or 2 accuracy studies, without further discussion on the topic of overdiagnosis
	Primary paper	Consists of a collection of original primary data collected by the researcher
	Systematic review	Collection and synthesis of available evidence on a topic. **Include:** Systematic assessments / meta-analyses of various articles within a specific domain **Exclude:** General discussions and exposes about a subject without a clear structural approach
Type of diagnostic test	Biomarker	Any measurement of chemicals in the human body as well as genotyping **Include:** immunohistochemistry (even though this may be assessed via microscopy in some cases) **Include:** Rapid diagnostic test for malaria

	Histology	Qualitative visual assessment of a target tissue through biopsy under a microscope (or similar devices) **Exclude:** Rapid diagnostic test for malaria (biomarker) **Exclude:** Scope (medical examination)
	Imaging	Any form of digital visualisation of the human body, such as MRI, CT, EKG, EEG, etc **Exclude:** Scope (medical examination)
	Medical examination	(Quick) medical tests that are performed directly by the clinician, either with or without specific medical equipment **Include:** Endoscopy, colonoscopy, spirometry, reflex test, exploratory surgery, DSM-V assessment, psychological evaluations, skin prick tests (for allergy), blood pressure measurement **Include:** Assessment of medical history of the patient by a clinician, such as age, gender, smoking habits, exercise pattern, etc
	Prediction model	List of predictors used in a prediction model **Exclude:** Overall assessments using multiple tests (="none") **Exclude:** Modelling studies that evaluate one particular index test, while using input on transition predictions in the rest of that model **Note:** Other index tests cannot be checked with a prediction model, since they will be part of that model
	None	Not one specific test is studied (so a broad range of tests or no specific one is addressed) **Include:** Overview papers that only discuss the general topic of overdiagnosis **Include:** Papers discussing various tests (hence there is no specific index test)

Screening	Yes / No	Is the primary focus of the study on diagnosis or detection in asymptomatic patients? **Include:** Screening is mentioned multiple times and explicitly **Exclude:** Screening as an example in an overview / review paper **Exclude:** Prognostic studies in patients that already received diagnosis
Overdiagnosis context	Overdiagnosis estimation	Overdiagnosis relating to the effect that a diagnostic test has on the number of excess cases found **Include:** Overdiagnosis mentioned in the results **Include:** Accuracy studies quantifying false-positive findings or % of overdiagnosis **Include:** Modelling papers that quantify overdiagnosis **Exclude:** Comparison of two diagnostic tests, without explicit quantification / assessment of overdiagnosis **Exclude:** Misdiagnosis / misclassification (= disease definition) **Exclude:** Overview papers that only briefly mention results from other primary studies **Exclude:** Overview papers that mention some quantitative results of overdiagnosis, but predominantly have a more broad discussion in general (=other)
	Disease definition	Overdiagnosis as a result of shifting the disease definition in terms of biomarker threshold or criteria in a scoring list **Include:** Misclassification / misdiagnosis **Include:** Papers assessing pathologic / biologic / mechanistic background of the disease in context with overdiagnosis. *However be critical whether these directly link particular biologic subclassifications of a disease to overdiagnosis*

	Overdiagnosis communication	Overdiagnosis as subject of communication between clinicians and/or patients **Include:** Studies that assess overdiagnosis communication to patients before or after diagnostic tests **Include:** Studies assessing people's general understanding of the concept of overdiagnosis
	Incidental findings	Overdiagnosis as a coincidental finding resulting from diagnostic testing of an unrelated illness
	Genomics	Overdiagnosis resulting from genome (screening) assessments, determining high-risk groups
	Other	Overdiagnosis that cannot be related to any of the categories above **Include:** Overview paper describing multiple aspects of overdiagnosis (e.g. accuracy, definition, litigation, methodology) **Include:** Studies looking at the downstream consequences of overdiagnosis (e.g. quality of life) **Include:** Studies looking at overall reasons for clinicians to overdiagnose (e.g. litigation risk, carefulness, unaware of negative consequences) **Include:** Trend studies **Include:** Studies on drivers and consequences of overdiagnosis

"There are two things in life: things you can change, and things you can't. The latter are not interesting"

Theo Jenniskens, my father

Chapter 5

A framework for overtesting, overdiagnosis, overtreatment, and related concepts in the era of 'too much medicine'

K. Jenniskens

L. Hooft

J.B. Reitsma

K.G.M. Moons

C.A. Naaktgeboren

Submitted

Abstract

Concepts related to 'too much medicine' remain a complex multifaceted issue, difficult to grasp and dissect. Although valuable descriptive frameworks have been proposed, these have not tackled the issues related to too much medicine across clinical domains, nor have they provided actionable strategies for reducing them. We provide a conceptual framework aimed at distinguishing uncertainty over thresholds and errors, two key mechanisms leading to 'too much medicine', and placing these in the clinical pathway of screening, diagnosis, prognosis and treatment of individuals. This allows researchers to evaluate concepts related to 'too much medicine' in the context of their own specific research, and facilitates communication between researchers, healthcare providers and patients. Based on the mechanism(s) at play, we provide strategies for reducing too much medicine.

Introduction

There is a rising awareness of the consequences of 'too much medicine' for individual patients and to the healthcare system as a whole. (1) 'Too much medicine' is the umbrella term to capsulize several concepts related to excessive and unnecessary use of healthcare services. Amongst these are overtesting, overdiagnosis, misdiagnosis, diagnostic error, overtreatment, and overutilization of medical activities. Communication on these topics, especially the widely used term overdiagnosis, is hampered by the myriad of definitions that exist, and our view is that there is an urgent need to be more specific about the underlying mechanisms. (2, 3)

While a single definition for 'too much medicine' might not be feasible, a descriptive framework may provide more insight into the many facets of the interrelated components. So far, most conceptual or methodological research in this area has focused on overdiagnosis and diagnostic error (4-7), specifically on providing guidance for widening disease definitions (8), methods for estimating the amount of overdiagnosis in a certain disease area (particularly in the field of oncology) (9-14), describing harms related to overdiagnosis (15), and creating frameworks that aim to provide better understanding of overdiagnosis. (7, 16-19) Though these frameworks are valuable for describing overdiagnosis as a general concept, most are only applicable only to a specific clinical field or context, (15, 19) and do not include other aspects of 'too much medicine' such as overtreatment or overtesting. (16, 17) Additionally, while links between the drivers for overdiagnosis and approaches for reducing it have been made (20, 21), they can be further elucidated and incorporated in a framework.

Building on preceding guidance and frameworks, we present a framework that includes various concepts related to 'too much medicine' and is applicable across clinical contexts. This allows researchers to disentangle whether the use of a test (either a screening, diagnostic, prognostic or monitoring test) and its downstream consequences (such as labelling of individuals and treatment decisions) leads to more harm than benefit. This broad starting point allows us to differentiate between the multiple, often simultaneously present, underlying mechanisms that drive 'too much

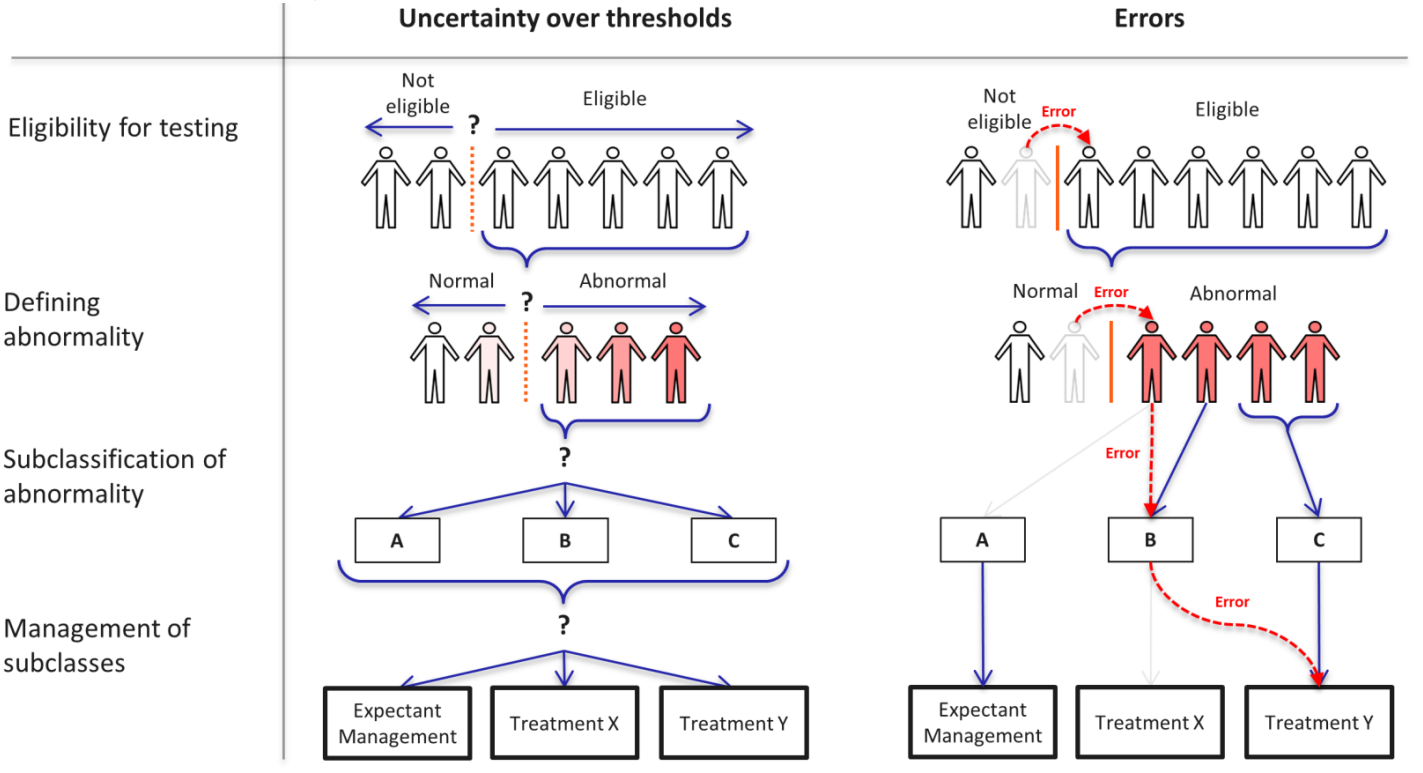
medicine'. Use of our framework will structure the thinking, communication and study of overtesting, overdiagnosis, overtreatment, and related concepts, and their consequences.

Conceptual framework

We propose a framework that maps different mechanisms of 'too much medicine' to stages of the clinical pathway (see Figure). This clinical pathway encompasses the entire clinical process starting as early as screening tests in (non-symptomatic) individuals up until management and treatment of patients diagnosed with the target condition. (22) We distinguish four key stages of the clinical pathway: eligibility for testing, presence of abnormality, determining sub-classification of abnormality, and choosing appropriate management for a subclass.

At each stage 'too much medicine' can arise from two main underlying mechanisms: 1. uncertainty over thresholds (e.g. screening, diagnostic, prognostic or treatment thresholds) and 2. errors (e.g. screening, diagnostic, prognostic or treatment decision errors). In the following sections we first provide a general description of these mechanisms, then we address them in context of the four key stages of the clinical pathway, and finally we illustrate them with clinical examples. This framework will facilitate describing and differentiating between, often interrelated and simultaneously occurring, underlying mechanisms leading to 'too much medicine', and provide guidance for actions that can be taken for reducing its harmful consequences.

Figure. Schematic framework describing uncertainty over thresholds and errors, as mechanisms that can occur at different stages of the clinical pathway. Uncertainties over thresholds are marked by questions marks and orange dotted lines. Colours given to individuals depict the gradual increase from no abnormality present (transparent) to abnormality clearly present (red). A, B, and C represent different diagnostic or prognostic subclasses of an abnormality. Treatments X and Y represent two treatments that are different with regard to, for example, clinical effectiveness, side-effects, and complications. Errors (red dotted arrows) can occur in individual patients after uncertainty over thresholds has been resolved. (i.e. there is no remaining uncertainty about the optimal threshold for maximum net benefit)



Mechanisms of too much medicine

Uncertainty over thresholds

The term 'threshold' is often used to refer to a numerical cut-off of a test, for example, what is considered high blood pressure? However, we use the term threshold in the broader sense, to refer to all the points along the clinical pathway at which a decision is made: screening, diagnosis, prognosis, and treatment. Uncertainty over thresholds is the situation in which there is no widely agreed upon threshold, or the chosen threshold is not evidence-based. In other words, there is uncertainty over thresholds when it is unclear which threshold results in the optimal harm/benefit ratio for a given person or targeted population. (23) The harms of 'too much medicine' introduced by uncertainty over thresholds can best be understood at the level of the healthcare system as a whole, as the question to be answered is: do the benefits outweigh the harms for the targeted group of individuals. Uncertainty over thresholds can be resolved through more high quality, adequately powered, empirical research into the long-term consequences (in terms of health outcomes and costs), and by summarizing this and all other available evidence in systematic reviews and evidenced based guidelines.

Errors

We define errors as any deviation from prevailing agreed upon thresholds. They can occur as diagnostic or prognostic errors (e.g. deviating or overruling the result of the prevailing 'gold standard') or management or treatment errors (e.g. deviating or overruling established treatment guidelines). Errors thus only exist when there are explicit agreed upon diagnostic or treatment thresholds. We broadly divide errors into human, technical, and system errors. Human errors are the result of incorrectly ordering, conducting, or interpreting information leading to incorrect decisions for the individual patient. Technical errors are the result of inaccurate provided by the instruments used (e.g. due to miscalibration or malfunctioning of tests). Finally, system errors occur when the (healthcare) system prevents certain tests or treatments being ordered or conducted (e.g. availability of only an inferior test or suboptimal treatment).

Uncertainty over thresholds and errors can both be understood at the level of the target population, however at the level of the individual patient this is more complex. Whether an error has occurred in a specific individual patient can be established by performing the reference standard in that same individual, or consulting the prevailing guidelines in case of a treatment decision. Whether a diagnosis or treatment leads to a net benefit for one specific individual can never be determined, as the alternative course of action (e.g. no diagnosis or an alternative treatment) was not taken for that individual. This relates to the concept of counterfactual thinking, which is considered complex and hard to interpret on an individual level. (24, 25)

Stages of the clinical pathway

In the following sections we describe how uncertainty over thresholds and errors occur at each stage of the clinical pathway, supported by clinical examples (Table 1 and the Appendix).

Eligibility for testing for abnormality

Uncertainty over thresholds as mechanism

The uncertainty in the first stage of the clinical pathway involves determining who is eligible for testing (i.e. for screening, diagnosis, prognosis, or monitoring). At one extreme, some screening tests may be applied universally and at the other extreme, some confirmatory, invasive, or expensive diagnostic or prognostic tests are only performed when there is a very high probability that the patient has a certain disease. The relevant question to ask is whether the benefits of testing a particular population outweigh the harms compared to not testing. To answer such a question, not only the direct harms of the test itself need to be considered, but also downstream consequences of the test results. A typical example can be found in the field of breast cancer, where there is uncertainty surrounding the appropriate starting age of screening. (Table 1, E1) (26, 27)

There is not only a trend within the public health and medical system towards broadening indication for screening and testing, but also within the general public, due to the rapid development in diagnostics and monitoring devices

that are available directly for the consumers, ranging from health monitoring apps for smartphones (28), to total body scans (29). Such technological developments are flooding the market before the key question on whether they lead to a net benefit on a population level, has been answered.

Incidental findings, defined as results pertaining to an abnormality other than the expected target disease, also relate to uncertainty over testing. With increasing resolution, these findings are becoming more common in various imaging tests (30, 31). While some incidental findings may be clinically relevant, others might be serendipitous, hence the net benefit of subsequent testing and/or treatment of incidental findings may be questionable. Uncertainty in incidental findings relates more to how far we should go when testing an individual, rather than who we should test.

Errors as mechanism

Errors of eligibility for testing occur when there are clear indication guidelines to decide who should be tested, but the tests are applied to an individual not having that indication. An example is when a physician is expected by a patient to perform a screening or diagnostic test regardless of whether there is an indication. (Table 1, E2) (32)

The question of interest to study this type of error is: why are more individuals being tested? Various drivers have been described that encourage the overuse of tests, including e.g. expectations of patients, the idea that 'more is better', the feeling of reassurance, consequences of undertesting, time constraints, and difficulty of keeping up with new evidence. (20) Additionally, there are few barriers to overtesting when tests are simple and non-invasive to use, little feedback is given on testing practices to physicians, or lack of financial disincentives. Tests may sometimes be bundled together on test request forms or software systems (i.e. a system error), and the physician is forced to perform a series of tests, despite the fact that combining these does not provide additional diagnostic information. (e.g. ASAT/ALAT tests for routine liver function) (33)

Defining abnormality

Uncertainty over thresholds as mechanism

Once a screening, diagnostic or prognostic test has been performed, we discriminate between individuals with and without abnormalities. In this stage of the clinical pathway we determine what we define as 'abnormality'. Thresholds to classify individuals can either be defined explicitly using numbers (e.g. mmHg for hypertension) or more implicitly (e.g. histopathological diagnosis). Because slight changes in treatment thresholds or diagnostic criteria can lead to a large change in absolute numbers of individuals being labelled as diseased, such changes are often a subject of serious debate. (8, 41) Uncertainty can arise when it is unknown which threshold for defining abnormality optimizes the net benefit for individuals in terms of patient outcomes and costs. However, it is important to consider that labels do not only have long-term impact via the treatment decisions they lead to, but that the diagnostic or prognostic label itself can have direct consequences for psychological well-being (i.e. being told you have an indolent tumour as opposed to cancer), eligibility for disability aids and reimbursement, and decision on further testing and management.

The widening disease definitions or prognostic classifications by altering thresholds results in inclusion of a group of individuals at low risk of unfavourable outcomes, in whom the net benefit of further testing and treatment may be limited. When considering altering a threshold for defining abnormality, the main research question to address is to quantify the net benefit of classifying those individuals with likely more mild conditions or prognosis as being diseased or indicated for treatment.

Examples of uncertainty over thresholds involve discussion about at what level high blood pressure needs to be classified and treated as hypertension (42, 43), and at what concentration of prostate specific antigen (PSA) warrants further testing in a prostate cancer screening setting? (44) These are both changes in explicit thresholds, though changes at implicit thresholds may be more insidious. Technological advances allow for detection of very low biomarker concentrations, and high resolution imaging allows small abnormalities to be observed. But are these clinically relevant?

Table 1. Examples of how uncertainty over thresholds and errors can lead to 'too much medicine' at different stages of the clinical pathway.

Clinical pathway	Mechanism	Example
Eligibility for testing	Uncertainty over thresholds	*E1. Breast cancer screening (34)* Whether extending the national breast cancer screening to start screening at age 40 instead of 50 will be beneficial.
	Errors	*E2. Head trauma in children (32)* Parental anxiety leading to cranial CT scans for minor blunt head trauma in children
Defining abnormality	Uncertainty over thresholds	*E3. Major depressive disorder (35)* The discussion about widening the definition by removal of the bereavement exclusion from the diagnostic criteria in the DSM-V
	Errors	*E4. Pulmonary Embolism (36)* Artefact on pulmonary CT angiography leading to diagnosis of pulmonary embolism
Subclassification of abnormality	Uncertainty over thresholds	*E5. Disruptive mood dysregulation disorder (DMDD) (37)* Whether the inclusion of DMDD as new diagnosis in the DSM-V as alternative for paediatric bipolar disorder will lead to better patient outcomes
	Errors	*E6. Alzheimer's disease (AD) / vascular dementia (VD) (38)* Structural brain imaging leading to misdiagnosis of VD as AD when compared to autopsy as a reference standard
Management of subclasses	Uncertainty over thresholds	*E7. Ductal carcinoma in situ (DCIS) (39)* Whether active surveillance is a valid alternative to surgical treatment for DCIS patients
	Errors	*E8. Malaria (40)* Despite having a negative culture and rapid diagnostic test, individuals still receive anti-malaria medication

For example, should small asymptomatic blood clots found in the lungs through imaging be labelled as pulmonary emboli? (45)

There can also be uncertainty over thresholds for defining the presence of abnormality in patients presenting with signs or symptoms. A threshold may shift depending on what society considers part of normal life. This makes abnormality a relative and dynamic concept. For example, it has been discussed whether prolonged bereavement is a natural reaction to a life-changing event, or a disease, as part of major depressive disorder? (Table 1, E3) (35)

Errors as mechanism

Errors at the level of defining abnormal result in incorrect diagnostic labels for individuals (when compared to a reference standard), consequently labelling too many individuals with a condition. In essence, errors by the test of interest at this level are false-positives, which can be detected in a research setting in which it is compared to a 'gold' reference standard. However, as this reference test is not performed in daily clinical practice (e.g. due to costs, risks, practical constraints), or its results are incorrectly interpreted, these false-positives will not be detected. Consequently, these individuals will receive a diagnostic label, subclass, and management, without any potential benefit. The key question that should be asked is what causes the errors in defining abnormality, and can we act upon those errors to reduce overdiagnosing abnormalities?

An example of classification error is patients suspected of scaphoid fractures who were diagnosed using bone scintigraphy scans. (46) Bone scintigraphy was found to overdiagnose individuals when compared to the prevailing "gold standard" of follow-up with clinical examination and repeated scans. Therefore, if scintigraphy scans would be implemented in clinical practice, overdiagnosis would occur. Other examples of classification errors include using a tourniquet cuff for longer than the advised duration inducing (pseudo)hyperkalaemia, leading to overdiagnosis of hypertension (47), or an artefact on a pulmonary CT angiography leading to an incorrect diagnosis of pulmonary embolism. (Table 1, E4) (36)

Subclassification of abnormalities

Uncertainty over thresholds as mechanism

Given the presence of an abnormality, thresholds are required to distinguish subclasses of that abnormality. The aim of subclassification is to identify and separate individuals based on the aetiology and prognosis of their abnormality. All these patients receive a diagnostic label, however the type of label they receive has implications for their prognosis and guides further patient management. Research questions in this area relate to the uncertainty over which thresholds for subclassification result in a balance between the harms and benefits of the treatments for each diagnostic or prognostic subcategory.

In oncology, for example, prognostic staging is used for abnormal growths, with higher stages corresponding to a more severe condition. However, there can be uncertainty over which thresholds provide valuable and accurate prognostic categories, usable to guide further patient management. (48) Another example in the field of obstetrics and gynaecology, relates to a change of diagnostic subclassifications by electronic foetal monitoring in the SOGC guidelines. This led foetuses with a non-reassuring foetal status to be further subclassified into 'abnormal' and 'atypical' tracings. (49, 50) Whether this subclassification results in improved patient management and clinical outcomes is unknown.

Uncertainty over subclassification of abnormalities can also refer to distinguishing between two different disorders with similar signs, symptoms or test results, but requiring different clinical management due to different underlying aetiologies. For example, there is uncertainty over subclassification of children previously diagnosed with bipolar disorder, whom should now be labelled with temper dysregulation syndrome with dysphoria. (Table 1, E5) (37) While in general there is consensus that these children have a condition and require intervention, the specific subclass of mental disorders to which they belong and type of intervention required (i.e. antipsychotics and mood stabilizers or psychosocial intervention) is up for debate.

Errors as mechanism

Errors at the level of subclassification share similar properties with errors occurring at the stage of defining abnormality. The test of interest can be compared to a reference standard to detect false-positives, however there is a reason that the reference test not performed, incorrectly performed, or wrongfully interpreted in daily clinical practice. The key research question is what the drivers are for this, and whether something can be done about it.

For example, in the case of patients suffering from some form of dementia, a follow-up question might be whether this is or is not Alzheimer related. When structural brain imaging was used to this extent, it was found to lead to overdiagnosis of vascular dementia when compared to autopsy results (the reference standard). (Table 1, E6) (38) This error is hard to resolve, as obviously autopsy results are not available at the time subclassification needs to be made. Another example relates to the overdiagnosis of borderline personality disorder as bipolar disorder in clinical practice, when compared to a research setting. (51) In this particular example, increasing availability and commercial stimuli of medication for bipolar disorder may be the underlying driver.

Management of subclasses

Uncertainty over thresholds as mechanism

Finally, after an abnormality has been identified, and a diagnostic or prognostic sub-classification has been made (if applicable), the next step is to determine the appropriate management. Especially in this component of the clinical pathway, there is a direct weighing of benefits (e.g. improving quality of life, prolonging life expectancy) and risks (e.g. complications, side-effects) of the available management options. The choice of treatment regimen, intensity or the decision to withhold treatment depends on both disease prognosis and probability of harms and benefits for a specific disease subclass. Uncertainty over thresholds for management decision relates to not knowing what management option provides the highest net benefit for a specific subclass. The main research question to be answered here is what is the optimal threshold to maximize harms and benefits of the treatments for each diagnostic and prognostic subclass of individuals.

For example, there can be uncertainty over whether ductal carcinoma in situ (DCIS) should be treated by surgical resection of the tumour, or monitored through active surveillance. (Table 1, E7) (39) Though surgical resection of the tumour would reduce the risk of recurrence, it does have risks related to complications, possible reconstructive surgery, and associated (healthcare) costs. Active surveillance would only require regular check-ups, though it may pose a higher risk of recurrence.

Errors as mechanism

Errors in management of a specific diagnostic or prognostic subclass often involves human errors. Doctors, either knowingly or unknowingly, provide management to an individual which is not recommended by prevailing guidelines. This deviation does not necessarily need to be wrong, as there may be specific circumstances, e.g. patient preferences, to specifically decide for another type of management than indicated. Individual patient preferences (e.g. for specific health related outcomes, willingness to pay) should be taken into account to provide personalized estimates of potential benefits and harms of treatment options. There is, however, also a trend in healthcare towards defensive medicine and "more is better". Research questions related to this error are aimed eliciting motivations for providing certain patient management, and looking at facilitators and barriers to use evidence-based guidelines, together with training healthcare professionals and patients the essential principles of "less-is-more" medicine and shared decision making.

An example of an error in management decision occurring can be found in the field of malaria treatment in resource limited settings. Despite having both a negative culture and rapid diagnostic test result, a significant number of individuals still received artemisinin-combination therapy against malaria. (Table 1, E8) (40)

Strategies for reducing 'too much medicine'

Possible directions for strategies to reduce 'too much medicine' may be informed by the underlying mechanism that is present (i.e. uncertainty over thresholds or errors). These are described in the following sections, and a

Table 2. Overview of mechanisms, their definitions, and possible strategies for reducing 'too much medicine'.

Mechanism	Definition	Strategies for reducing 'too much medicine'
Uncertainty over thresholds	Not knowing what the most appropriate threshold is for maximizing benefits and minimizing harms for a group of individuals	Solution lies in gathering evidence on optimal thresholds - Clinical trials - Observational studies - Modelling studies
Human errors	Humans that take discordant actions when compared to a reference standard or guideline	Solution lies in behaviour of doctors and patients - Uncover underlying drivers - Better communication of guidelines - Discuss motivations for deviating from guidelines - Training principles of "less-is-more" medicine
Technical errors	Devices or software programs that lead to discordant actions when compared to a reference standard or guideline	Solution lies in device or software - Optimize current tests - Implement better tests - Improve software system
System errors	The (healthcare) system leads to discordant actions when compared to a reference standard or guideline	Solution lies in (healthcare) system - Increase access to better diagnostic tests and treatments - Reduce financial or time constraints - Improve quality of guidelines

general overview is given in Table 2. It is important to note that uncertainty over thresholds and errors can occur simultaneously, hence multiple strategies may need to be used to maximize reduction of 'too much medicine'.

Reducing uncertainty over thresholds

Performing research on the impact of a threshold shift (e.g. who to test, who to diagnose, who to prognosticate, or who to treat) on patient relevant outcomes can help reduce uncertainty over what the optimal threshold is. (8) Impact of a shift in threshold should be measured by comparing the harms and benefits for one threshold to the other, and should incorporate all important downstream consequences, such as subsequent patient management, complications, and clinically relevant health events. The earlier in the clinical pathway (i.e. eligibility for testing) the more complex it is to capture all the downstream consequences. Ultimately, the result of a threshold shift is a ratio between the harms (psychological harm from unnecessary labelling, complications from treatment, medical expenses) and benefits (early detection and treatment of clinically relevant disease or prognostic classes) of the old and the new threshold. (15) Lowering thresholds will result in more (or earlier) detection of individuals with clinically relevant disease, but will inevitably increase overdiagnosis and overtreatment.

Several study designs can be considered to assess the harm/benefit ratio, each with specific strengths and limitations. (11, 52) Clinical trials or observational studies are useful for obtaining direct evidence through empirical estimates of overdiagnosis, and incremental cost-effectiveness ratios. However, when there is, for example, a significant amount of lead-time (i.e. the time between development of an abnormality and actual detection of the targeted condition), only costly randomized screening trials with significant follow-up time may be suitable as an empirical study design. (53) Alternatively, (decision analytic) modelling studies can be used to simulate long-term outcomes. (54-56) There is an ongoing debate on how to weigh harms and benefits. In some situations, a cost-effectiveness based approach may be combined with citizens' juries or multiple criteria decision analysis (MCDA) to ensure stakeholder support of a new threshold. (57, 58)

Reducing errors

While the solution to reducing uncertainty over thresholds is more high quality empirical or model-based research, the solutions to reducing errors are multifaceted. When human errors are the cause of 'too much medicine', it is helpful to identify whether the errors occur due to a lack of knowledge on guidelines, their attitude towards these guidelines, or practical issues in performing tests. This way interventions can be targeted specifically towards the problem at hand. Dialogue with physicians is key in unravelling the drivers for these errors. If, for example, knowledge about current guidelines is lacking, one might opt to stimulate communication and education of these guidelines to healthcare providers as a potential solution. However, if physicians knowingly deviate from those guidelines (sometimes rightfully so), solutions might want to be sought elsewhere.

Technical errors can be reduced by improving the performance of the diagnostic or prognostic test itself. Technical performance of a device can be increased by, for example, improving discrimination (i.e. sensitivity and specificity), consistency, performing software updates, or reducing artefacts. These aforementioned properties can be improved for the current index test, but alternatively a novel test with better performance may be developed.

For system errors it is important to realize that neither the healthcare professional, nor the device or test, is to blame. The professional may be restricted by the system if one is required to diagnose, prognosticate and treat the patient with limited resources and time. The (software) system may also only allow for packages of bundled tests to be ordered, resulting in unnecessary overtesting. (59) Solutions should be sought at the level of the organization, that may, for example, provide more time per patient, allocate more resources for guideline recommended tests and treatments, or ensure flexibility in software.

Concluding remarks

We have developed a framework that allows concepts related to 'too much medicine' to be described by using uncertainty over thresholds and errors as mechanisms. These mechanisms are outlined using four stages along the

clinical pathway, from testing to treatment. A variety of clinical examples have been given to demonstrate its applicability across clinical fields and contexts. Furthermore, the framework helps to provide guidance for strategies useful for assessing or reducing 'too much medicine' based on the underlying mechanisms. Use of this framework by researchers, healthcare professionals, policy makers, and others interested in 'too much medicine', including related concepts such as overdiagnosis and diagnostic error, will help facilitate communication and stimulate constructive discussion.

References

1. Carter SM, Rogers W, Heath I, Degeling C, Doust J, Barratt A. The challenge of overdiagnosis begins with its definition. BMJ. 2015;350:h869.

2. Bae JM. Overdiagnosis: epidemiologic concepts and estimation. Epidemiol Health. 2015;37:e2015004.

3. Brodersen JA-Ohoo, Schwartz LM, Heneghan C, O'Sullivan JWA-Ohoo, Aronson JK, Woloshin S. Overdiagnosis: what it is and what it isn't. (2515-4478 (Electronic)).

4. Institute of M, National Academies of Sciences E, Medicine. Improving Diagnosis in Health Care. Erin PB, Bryan TM, John RB, editors. Washington, DC: The National Academies Press; 2015.

5. Godlee F. Preventing overdiagnosis. BMJ : British Medical Journal. 2012;344.

6. Balogh E, Miller B, Ball J. Improving Diagnosis in Health Care. Washington (DC)2015.

7. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis (Berl). 2015;2(2):97-103. Epub 2016/03/10.

8. Doust J, Vandvik PO, Qaseem A, Mustafa RA, Horvath AR, Frances A, et al. Guidance for Modifying the Definition of Diseases: A Checklist. JAMA Intern Med. 2017.

9. Wu D, Perez A. A limited Review of Over Diagnosis Methods and Long Term Effects in Breast Cancer Screening. Oncol Rev. 2011;5(3):143-7.

10. Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. Ann Intern Med. 2013;158(11):831-8.

11. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. BMJ. 2015;350:g7773.

12. de Gelder R, Heijnsdijk EA, van Ravesteyn NT, Fracheboud J, Draisma G, de Koning HJ. Interpreting overdiagnosis estimates in population-based mammography screening. Epidemiol Rev. 2011;33:111-21.

13. Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulati R, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J Natl Cancer Inst. 2009;101(6):374-83.

14. Ripping TM, Ten Haaf K, Verbeek ALM, van Ravesteyn NT, Broeders MJM. Quantifying Overdiagnosis in Cancer Screening: A Systematic Review to Evaluate the Methodology. J Natl Cancer Inst. 2017;109(10).

15. Harris RP, Sheridan SL, Lewis CL, Barclay C, Vu MB, Kistler CE, et al. The harms of screening: a proposed taxonomy and application to lung cancer screening. JAMA Intern Med. 2014;174(2):281-5. Epub 2013/12/11.

16. Rogers WA, Mintzker Y. Getting clearer on overdiagnosis. J Eval Clin Pract. 2016;22(4):580-7.

17. Hofmann B. The overdiagnosis of what? On the relationship between the concepts of overdiagnosis, disease, and diagnosis. Med Health Care Philos. 2017.

18. Carter SM, Degeling C, Doust J, Barratt A. A definition and ethical evaluation of overdiagnosis. J Med Ethics. 2016.

19. Marcus PM, Prorok PC, Miller AB, DeVoto EJ, Kramer BS. Conceptualizing overdiagnosis in cancer screening. J Natl Cancer Inst. 2015;107(4).

20. Pathirana T, Clark J, Moynihan R. Mapping the drivers of overdiagnosis to potential solutions. BMJ. 2017;358:j3879.

21. Newman-Toker DE. A unified conceptual model for diagnostic errors: underdiagnosis, overdiagnosis, and misdiagnosis. Diagnosis (Berl). 2014;1(1):43-8.

22. Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Defining the clinical pathway in cochrane diagnostic test accuracy reviews. BMC medical research methodology. 2016;16(1):153. Epub 2016/11/12.

23. Schiff GD, Martin SA, Eidelman DH, Volk LA, Ruan E, Cassel C, et al. Ten Principles for More Conservative, Care-Full Diagnosis. Ann Intern Med. 2018;169(9):643-5. Epub 2018/10/05.

24. Hofmann B. Getting personal on overdiagnosis: On defining overdiagnosis from the perspective of the individual person. J Eval Clin Pract. 2018;24(5):983-7. Epub 2018/08/02.

25. Hofmann B. Diagnosing overdiagnosis: conceptual challenges and suggested solutions. Eur J Epidemiol. 2014;29(9):599-604.

26. Degeling C, Barratt A, Aranda S, Bell R, Doust J, Houssami N, et al. Should women aged 70-74 be invited to participate in screening mammography? A report on two Australian community juries. BMJ Open. 2018;8(6):e021174. Epub 2018/06/16.

27. Sardanelli F, Aase HS, Alvarez M, Azavedo E, Baarslag HJ, Balleyguier C, et al. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. European radiology. 2017;27(7):2737-43. Epub 2016/11/04.

28. Chen A. Why doctors are worried about the Apple Watch EKG. 2019 [29-01-2019]; Available from: https://www.theverge.com/2019/1/8/18172132/apple-watch-ekg-electrocardiogram-health-science.

29. Furtado CD, Aguirre DA, Sirlin CB, Dang D, Stamato SK, Lee P, et al. Whole-body CT screening: spectrum of findings and recommendations in 1192 patients. Radiology. 2005;237(2):385-94. Epub 2005/09/20.

30. Booth TC, Najim R, Petkova H. Incidental findings discovered during imaging: implications for general practice. The British journal of general practice : the journal of the Royal College of General Practitioners. 2016;66(648):346-7. Epub 2016/07/02.

31. O'Sullivan JW, Muntinga T, Grigg S, Ioannidis JPA. Prevalence and outcomes of incidental imaging findings: umbrella review. BMJ. 2018;361:k2387. Epub 2018/06/20.

32. Natale JE, Joseph JG, Rogers AJ, Mahajan P, Cooper A, Wisner DH, et al. Cranial computed tomography use among children with minor blunt

head trauma: association with race/ethnicity. Archives of pediatrics & adolescent medicine. 2012;166(8):732-7. Epub 2012/08/08.

33. Linthorst G. Interpretation of lab results - ALAT and ASAT. Amsterdam Medical Student journal. 2016;6:16-7.

34. Bjorndal A, Forsetlund L. Mammography Screening of Women 40-49. Oslo, Norway2007.

35. Regier DA, Kuhl EA, Kupfer DJ. The DSM-5: Classification and criteria changes. World Psychiatry. 2013;12(2):92-8.

36. Hutchinson BD, Navin P, Marom EM, Truong MT, Bruzzi JF. Overdiagnosis of Pulmonary Embolism by Pulmonary CT Angiography. AJR Am J Roentgenol. 2015;205(2):271-7.

37. Johnson K, McGuinness TM. Disruptive mood dysregulation disorder: a new diagnosis in the DSM-5. J Psychosoc Nurs Ment Health Serv. 2014;52(2):17-20.

38. Niemantsverdriet E, Feyen BF, Le Bastard N, Martin JJ, Goeman J, De Deyn PP, et al. Overdiagnosing Vascular Dementia using Structural Brain Imaging for Dementia Work-Up. J Alzheimers Dis. 2015;45(4):1039-43.

39. Groen EJ, Elshof LE, Visser LL, Rutgers EJT, Winter-Warnars HAO, Lips EH, et al. Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). Breast. 2017;31:274-83.

40. Mwanziva C, Shekalaghe S, Ndaro A, Mengerink B, Megiroo S, Mosha F, et al. Overuse of artemisinin-combination therapy in Mto wa Mbu (river of mosquitoes), an area misinterpreted as high endemic for malaria. Malar J. 2008;7:232.

41. Manrai AK, Patel CJ, Ioannidis JPA. In the Era of Precision Medicine and Big Data, Who Is Normal? JAMA. 2018;319(19):1981-2. Epub 2018/05/02.

42. Miyazaki K. Overdiagnosis or not? 2017 ACC/AHA high blood pressure clinical practice guideline: Consequences of intellectual conflict of interest. Journal of general and family medicine. 2018;19(4):123-6. Epub 2018/07/13.

43. 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/PCN A 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. Hypertension. 2018;71(6):1269-324. Epub 2017/11/15.

44. Pelzer AE, Colleselli D, Bektic J, Schaefer G, Ongarello S, Schwentner C, et al. Over-diagnosis and under-diagnosis of screen- vs non-screen-detected prostate cancers with in men with prostate-specific antigen levels of 2.0-10.0 ng/mL. BJU Int. 2008;101(10):1223-6.

45. Wiener RS, Schwartz LM, Woloshin S. When a test is too good: how CT pulmonary angiograms find pulmonary emboli that do not need to be found. BMJ. 2013;347:f3368. Epub 2013/07/04.

46. de Zwart AD, Beeres FJ, Rhemrev SJ, Bartlema K, Schipper IB. Comparison of MRI, CT and bone scintigraphy for suspected scaphoid fractures. European journal of trauma and emergency surgery : official publication of the European Trauma Society. 2016;42(6):725-31. Epub 2015/11/12.

47. Asirvatham JR, Moses V, Bjornson L. Errors in potassium measurement: a laboratory perspective for the clinician. N Am J Med Sci. 2013;5(4):255-9.

48. Oude Ophuis CM, Louwman MW, Grunhagen DJ, Verhoef K, van Akkooi AC. Implementation of the 7th edition AJCC staging system: Effects on staging and survival for pT1 melanoma. A Dutch population based study. Int J Cancer. 2017;140(8):1802-8. Epub 2017/01/22.

49. Ashmead G. Fetal Heart Rate Monitoring Update: The Good, the Bad, and the Atypical. The Female Patient. 2011;36:14-22.

50. Liston R, Sawchuck D, Young D, Society of O, Gynaecologists of C, British Columbia Perinatal Health P. Fetal health surveillance: antepartum and intrapartum consensus guideline. J Obstet Gynaecol Can. 2007;29(9 Suppl 4):S3-56. Epub 2007/10/20.

51. Zimmerman M, Ruggero CJ, Chelminski I, Young D. Psychiatric diagnoses in patients previously overdiagnosed with bipolar disorder. J Clin Psychiatry. 2010;71(1):26-31.

52. Noyes K, Holloway RG. Evidence from cost-effectiveness research. NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics. 2004;1(3):348-55. Epub 2005/02/18.

53. Puliti D, Miccinesi G, Paci E. Overdiagnosis in breast cancer: design and methods of estimation in observational studies. Prev Med. 2011;53(3):131-3.

54. van Luijt PA, Heijnsdijk EA, van Ravesteyn NT, Hofvind S, de Koning HJ. Breast cancer incidence trends in Norway and estimates of overdiagnosis. J Med Screen. 2017;24(2):83-91.

55. van Luijt PA, Heijnsdijk EA, de Koning HJ. Cost-effectiveness of the Norwegian breast cancer screening program. Int J Cancer. 2017;140(4):833-40.

56. Draisma G, De Koning HJ. MISCAN: estimating lead-time and over-detection by simulation. BJU Int. 2003;92 Suppl 2:106-11. Epub 2004/02/27.

57. Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making--An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2016;19(1):1-13. Epub 2016/01/23.

58. Wise J. Citizens' juries for health policy. BMJ. 2017;357:j2650. Epub 2017/06/04.

59. Bindraban RS, Ten Berg MJ, Naaktgeboren CA, Kramer MHH, Van Solinge WW, Nanayakkara PWB. Reducing Test Utilization in Hospital Settings: A Narrative Review. Annals of laboratory medicine. 2018;38(5):402-12. Epub 2018/05/26.

Appendix

Examples specific for human, technical, and system errors

Clinical pathway	Mechanism	Example
Eligibility for testing	Human errors	*Head trauma in children (32)* Anxiety of parents leading to an increased number of Cranial CT scans being ordered
	Technical errors	*Breast cancer screening (60)* Women of 70 years old not being invited to breast cancer screening due to a software failure
	System errors	*Atrial fibrillation (28)* People having free access to apps for smartwatches allowing for monitoring atrial fibrillation
Defining abnormality	Human errors	*Attention deficit hyper activity disorder (ADHD) (61)* Overestimating the role of gender in diagnosing ADHD
	Technical errors	*Pulmonary Embolism (36)* Artefacts on pulmonary CT angiography leading to overdiagnosis
	System errors	*Mental disorders (62)* Clinicians intentionally code mental disorders wrong in order to ensure treatment access and reimbursement
Subclassification of abnormality	Human errors	*Bipolar disorder (BD) / borderline personality disorder (BPD) (51)* BPD is misdiagnosed as BD in clinical practice when compared to a research setting
	Technical errors	*Lung carcinoid tumours (63)* Artefacts on histological staining lead to misclassification of cytological diagnosis
	System errors	*Hypo(mania) (64)* Guidelines for diagnosing (hypo)mania use unspecific criteria, open for interpretation

Clinical pathway	Mechanism	Example
Management of subclasses	Human errors	*Malaria (40)* Despite having a negative culture and rapid diagnostic test, individuals still receive anti-malaria medication
	Technical errors	*Incorrect dosage (65)** Excessive dose of medication being given due to lack of software feedback loop
	System errors	*Fee-for-service (66)* A fee for service system leads to overtreatment and overutilization of healthcare resources

* Although the origin is a human error, the solution can be found at the software level

"Most models are wrong,

the question is how wrong they have to be to not be useful"

George E.P. Box, statistician

Chapter 6

Decision analytic modelling was a valuable tool to assess the impact of a risk prediction model on health outcomes and healthcare costs before a randomized trial

K. Jenniskens*

G.R. Lagerweij*

C.A. Naaktgeboren

L. Hooft

K.G.M. Moons

J.M. Poldervaart

H. Koffijberg

J.B. Reitsma

**Authors contributed equally*

J Clin Epidemiol. 2019 Jul 19;115:106-115

Abstract

Objective: To demonstrate how decision analytic models (DAM) can be used to quantify impact of using a (diagnostic or prognostic) prediction model in clinical practice, and provide general guidance on how to perform such assessments.

Study design and setting: A DAM was developed to assess the impact of using the HEART score for predicting major adverse cardiac events (MACE). Impact on patient health outcomes and healthcare costs was assessed in scenarios by varying compliance with and informed deviation (ID) (using additional clinical knowledge) from HEART score management recommendations. Probabilistic sensitivity analysis was used to assess estimated impact robustness.

Results: Impact of using the HEART score on health outcomes and healthcare costs was influenced by an interplay of compliance with and ID from HEART score management recommendations. Compliance of 50% (with 0% ID) resulted in increased missed MACE and costs compared to usual care. Any compliance combined with at least 50% ID, reduced both costs and missed MACE. Other scenarios yielded a reduction in missed MACE at higher costs.

Conclusion: Decision analytic modelling is a useful approach to assess impact of using a prediction model in practice on health outcomes and healthcare costs. This approach is recommended before conducting an impact trial to improve its design and conduct.

Introduction

Diagnostic or prognostic prediction models can be used to support management decisions such as subsequent testing, treatment or lifestyle changes. Developed prediction models require external validation to ensure they have adequate predictive performance. (1-4) However good predictive performance does not imply that implementation in clinical practice will improve health outcomes or reduce healthcare costs. The impact of using risk prediction models in clinical practice on patient health and monetary outcomes can be evaluated in impact studies, such as comparative longitudinal (ideally (cluster) randomized) trials, in which care directed by the prediction model is compared to usual care. (5-10)

Impact studies for prediction models are infrequent, most likely due to their complexity, long follow-up, associated high costs, and lack of regulatory requirements. (7-9, 11-13) In addition, the benefits observed in such impact studies have typically been smaller than expected, or even lacking. (14-16) An approach using a decision analytic model (DAM) may prove useful, making use of evidence available at the time an impact study is being considered. A DAM could provide insight in the conditions under which a prediction model is likely to result in favourable health outcomes or costs when implemented in clinical practice.

Decision analytic modelling is a method that integrates multiple sources of evidence to assess the downstream cost-effectiveness of applying a prediction model in daily practice. (7-9, 17, 18) Constructing a DAM forces researchers to think about the pathway through which (multiple alternative) complex interventions can lead to health and monetary benefits; such as variation in the interplay between the model predictions and subsequent patient management based on these predicted risks. DAMs also allow for uncertainty on parameters, such as distribution of predicted probabilities or effectiveness of treatment, to be taken into account. Additionally, downstream effects of hypothetical scenarios can be analysed, by varying values of parameters for which there is little or no evidence. The results are then used to inform decisions for an individual patient or healthcare policy. DAMs have also been proposed and performed before conducting

longitudinal comparative trials to assess impact of (complex) therapeutic interventions and diagnostic tests (19-21), although they are still rare for diagnostic or prognostic prediction models. An explanation for this could be that using DAMs to assess impact is more complex for prediction models than for interventions, as the former would not only need to include accuracy of predictions, but also downstream effects of, for example, benefits and harms of subsequent tests. Additionally, lack of available evidence on compliance with management recommendations from a prediction model based on the predicted risk, and informed deviation from that compliance (i.e. whether there is incremental value of a clinician's experience on top of predictions provided by a model) may also explain the limited number of DAMs assessing impact of prediction models before conducting a formal large scale, long-term, costly, empirical impact study. Even though DAMs are particularly ideal to estimate the impact when evidence is lacking, namely by simulating multiple (hypothetical) scenarios.

In this paper we demonstrate how to assess the potential impact of a prediction model on patient health outcomes and healthcare costs using a DAM approach, specifically focusing on the effect of compliance with management recommendations. We will use the HEART score prediction model for diagnosis of major adverse cardiac events (MACE) in patients with chest pain as a case study. (22) This paper will conclude by providing generic guidance on how to perform a decision analytic model-based assessment to estimate the impact of using a prediction model in daily practice, and how the results of such DAM can inform the design and conduct of a subsequent prospective comparative prediction model impact study.

Methods

Case study

We compared implementation of the HEART score prediction model to usual care in a DAM as an example of how compliance with management recommendations from a prediction model influences the impact of that model on patients' health outcomes, healthcare costs, and cost-effectiveness. The HEART score provides an excellent example for illustrating the usefulness of a DAM, as model development (22) and several external

validations have shown that the HEART score can correctly predict and stratify patients according to their risk of having MACE (23-26), and HEART score predictions were categorized and linked to management recommendations. (Table 1) Although a randomized impact trial has recently been conducted for the HEART score prediction model (27), the DAM only used information from studies and data sources available before this trial was conducted. Note that the main of this paper was not to replicate the results from this impact trial (27), but rather to illustrate *how* a DAM can be used to assess the impact of a prediction model on patient health outcomes and healthcare costs.

The HEART score is a prediction model that uses routinely collected information from patient history and blood tests to predict MACE in patients presenting with chest pain at the emergency room, to generate a risk score ranging from 0 to 10. (Table 1) (22) The potential benefit of using the HEART score lies in its ability to stratify patients according to their risk of MACE and provide risk based management recommendations. Physicians are advised to promptly discharge low risk patients (i.e. HEART score ≤ 3), reducing utilization of healthcare resources, and providing additional diagnostic testing in higher risk patients (i.e. HEART score ≥ 4), to prevent unnecessary delay in treatment initiation. Non-invasive diagnostic testing for the intermediate HEART score category consisted of stress bicycle ECG, myocardial scintigraphy, coronary CT angiography, and cardiac MRI. Invasive diagnostic testing for the high HEART score category consisted of coronary angiography, in combination with any of the non-invasive tests.

Table 1. Overview of the HEART score prediction scores, categories, and their associated risk based management recommendations

HEART score	HEART score category	Management recommendation
0 – 3	Low	Discharge home
4 – 6	Intermediate	Non-invasive testing
7 – 10	High	Invasive testing

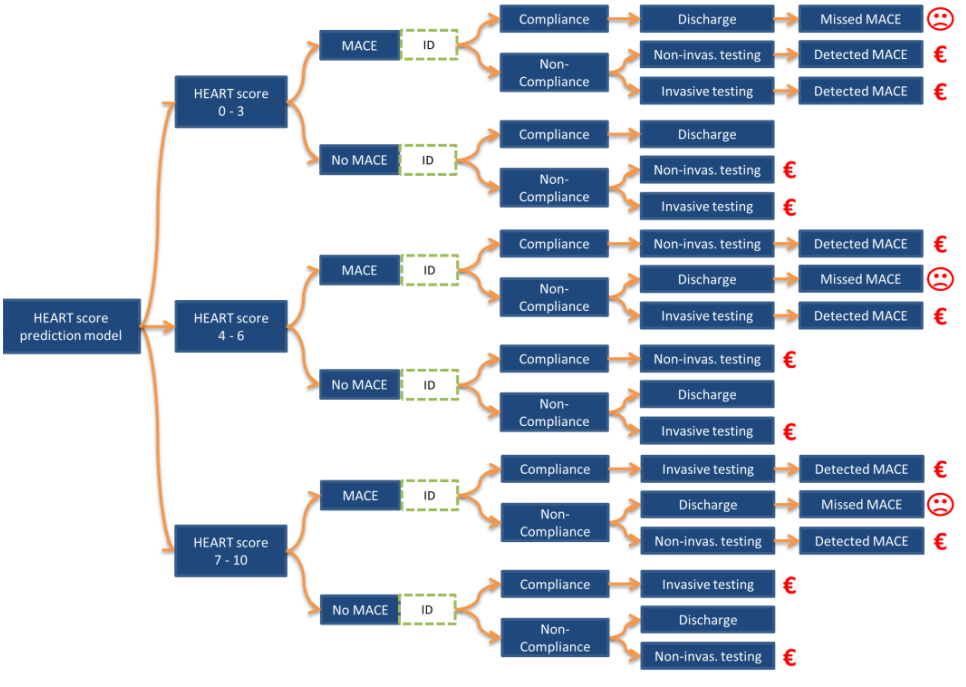
We evaluated the HEART score purely as a diagnostic instrument for MACE, meaning that in our model the HEART score and any subsequent actions do not have an impact on the total number of MACE. MACE found during diagnostic work-up (detected MACE) were considered a favourable outcome, whereas MACE in discharged patients (missed MACE) were considered an unfavourable outcome.

Structure of the decision analytic model

Figure 1 and Appendix A display the DAM comparing usual care to the HEART score strategy. In the usual care strategy HEART scores are not available to clinicians and are therefore not used to guide subsequent patient management decisions. In the HEART score strategy, we mimicked that clinicians at the emergency department would calculate the HEART score, and be given clear guidance on subsequent risk based patient management recommendations (see Table 1).

The DAM used an assistive (as opposed to a directive) prediction model approach, meaning physicians were not forced to comply with management recommendations. (8, 28) This allows for better mimicking actual implementation of the prediction model in clinical practice and thus provide more realistic and generalizable quantification of impact. The focus of this study is to quantify the impact of compliance with the HEART score predictions and subsequent management recommendations on patient relevant health outcomes and monetary outcomes. Accordingly, we varied the amount of compliance to prediction model's management recommendations (i.e. the percentage of patients in whom the specified management recommendation was followed) in several scenarios (see "Scenario analysis" paragraph).

Figure 1. Decision tree for using the HEART score prediction model for management decisions in patients presenting with chest pain at the emergency department. ID = Informed deviation of management recommendations corresponding to HEART score predictions. Euro signs and emoticons represent negative effects on costs and health outcomes respectively.



In the DAM clinicians could deviate from recommended management based on additional patient information (e.g. signs and symptoms) or clinical expertise, leading to more appropriate stratification of management given to patients, so-called informed deviation (ID). This ID was included as a variable in the DAM, defined as the proportion of patients for whom the initial management recommendations according to the prediction model were incorrect, in which physicians – informed by additional knowledge – correctly deviate from those recommendations. ID ranged from 0% (uninformative compliance; compliance is equal in patients with and without MACE) to 100% (fully informative compliance; patients with MACE follow a diagnostic pathway, patients without MACE are discharged). For an example of how ID influences management recommendations in the low

HEART score category, see Table 2. Introducing 50% ID to a scenario in which there is 80% compliance to management recommendations, would lead to an additional 40% of patients with MACE receiving testing, and an additional 10% of individuals without MACE being discharged.

Table 2. Illustration how compliance with, and informed deviation from HEART score management recommendations for the low HEART score category (0-3) influence the proportion of patients being discharged vs. receiving additional (non-)invasive testing. Asterisks denote the preferred course of action for patients with and without MACE.

		Scenario (HEART score 0-3)		
		Compliance 80% ID 0%	ID 50%	Compliance 80% ID 50%
MACE	Discharged	80%	-40% (80%*0.5)	40%
	Additional testing*	20%	+40% (80%*0.5)	60%
No MACE	Discharged*	80%	+10% (20%*0.5)	90%
	Additional testing	20%	-10% (20%*0.5)	10%

Input parameters for the decision analytic model

To operationalize the DAM, each parameter requires an input value. Three types of input parameters are considered. First, transition probabilities, which are the probabilities for transitioning from one (health) state to the next, are defined (marked in Figure 1 by the orange arrows). Secondly, we defined the main and other health outcomes. Finally, input values for the intended and unintended effects and costs of any subsequent tests, treatments, and conditions need to be determined. Input for most of these parameters in the "usual care" strategy was based on the observational data from the study by Nieuwets et al. (29) See Appendix B for an overview of all input parameters.

Transition probabilities

The distribution of the target patient population across HEART score categories and MACE rates per HEART score category were derived from development(22), and multiple external validation studies of the HEART score prediction model. (23, 24, 30) Values for compliance and ID were not available, and are further described in the "Scenario analysis" section. Transition probabilities and likelihood of receiving specific diagnostic tests (e.g. a stress bicycle ECG) in non-invasive and invasive diagnostic testing pathways, were derived from a study measuring consumption of healthcare resources in usual care. (29)

Health outcomes

The health outcome of interest was defined as the proportion of missed MACE, i.e. patients with MACE at 6 weeks who were (initially) discharged without any subsequent diagnostic work-up. MACE detected during or occurring after diagnostic workup was not included as adverse outcome, as this would have been found and managed accordingly in clinical practice. MACE was defined as occurrence of one or more of the following events or interventions: acute myocardial infarction (both ST- and non-ST-segment elevation), unstable angina, percutaneous coronary intervention (PCI), coronary artery bypass grafting (CABG), significant stenosis (>50%) managed conservatively, and death due to any cause. (31)

Healthcare costs

Calculation of the HEART score relies on readily available predictors, hence no extra costs are associated with collection of these predictors when compared to usual care. Costs of MACE were calculated based on a weighted average of costs and probability of each individual MACE component, derived from scientific literature. (29, 32-35) Costs of non-invasive and invasive testing pathways in a specific HEART score category were calculated by taking the average number of times a specific diagnostic test was used per patient in that pathway, and multiplying it by its unit costs. (29) Summing the average cost for all diagnostic tests in each of the pathways yielded the total costs of diagnostic testing. Similarly, the average number of admission and re-admission days were calculated for each of the diagnostic pathways. Complication rates in non-invasive and invasive testing pathways

were not explicitly included in the model, however the expected frequency of severe complications for procedures included in the DAM is low (36-38) and expected costs of complications are largely captured by the number of (re)admission days.

Analyses

Scenario analysis was performed, comparing hypothetical scenarios in which compliance and ID were varied. Furthermore, a probabilistic sensitivity analysis was performed, in which a cohort was run through a series of simulations to take into account uncertainty surrounding the parameters in the DAM. A time horizon of 6 weeks was taken for the analyses, for which discounting was not deemed necessary.

Scenario analysis

Scenario analysis focused on comparing different compliances to HEART score predictions and corresponding management recommendations, combined with varying degrees of ID from those compliances. The influence of compliance on missed MACE and costs was investigated in three different scenarios: low (50%), medium, (75%), and full (100%) compliance. Furthermore, four scenarios were defined for ID: no (0%), low (25%), medium (50%), and high (75%) ID.

For each scenario the incremental proportion of missed MACE, healthcare costs, and cost per missed MACE will be given per HEART score category and for all HEART score categories combined, as compared with usual care. Cost-effectiveness planes will be provided to give insight in the distribution of missed MACE and healthcare costs in the presence of parameter uncertainty.

Probabilistic sensitivity analysis

Monte Carlo simulation was used to assess the robustness of expected health outcomes and healthcare costs based on uncertainty surrounding the different parameters. A series of 10,000 simulations was run per scenario, each with a patient population of 200,000, reflective of the annual Dutch population visiting the emergency department with chest pain. (39)

Parameter uncertainty was reflected by calculating standard errors, and defining appropriate statistical distributions for each parameter. Beta and Dirichlet distributions were used to account for uncertainty in transition probabilities. Gamma distributions were used for uncertainty surrounding costs (see Appendix B).

Results

In usual care, the average proportion of patients with missed MACE was estimated at 0.016 (95% confidence interval 0.007 – 0.027), or an average of 16 MACE in discharged patients per 1,000 individuals presenting with chest pain at the emergency department. The average cost per patient in usual care was €2,870. (29)

Scenario analysis

The impact of compliance and ID on the number of missed MACE (i.e. effects), costs and cost-effectiveness was investigated in various scenarios. Negative values for missed MACE indicate a decrease and positive numbers an increase in missed MACE compared to usual care. The values in Table 3, 4 and 5 are coloured green to indicate a beneficial effect, or red to indicate an unbeneficial effect of using the HEART score in practice.

Missed MACE

Table 3 shows the average difference in missed MACE (per person) for each of the HEART score categories and for the total patient population, as compared to usual care. Maybe somewhat surprisingly at first sight, the low HEART score category shows an increase in the proportion of missed MACE as compliance increases, whereas in the intermediate and high HEART categories there is an inverse relation. This can be explained by the different management recommendations associated with each HEART score category. Higher compliance in the low HEART score category, obviously leads to more patients being discharged, running the risk of missing more MACE in these patients. On the other hand, compliance in the intermediate and high HEART score categories, automatically implies more diagnostic testing, reducing the risk of missing MACE. ID counteracts the higher

Table 3. Average difference in missed MACE per person between the HEART score strategy and usual care. A negative number represents a reduction in missed MACE. The average number of missed MACE per patient in usual care was 0.016.

	Compliance	Informed deviation (ID)			
		0%	25%	50%	75%
HEART score 0-3	50%	0.006	0.004	0.001	-0.001
	75%	0.011	0.007	0.004	0.000
	100%	0.016	0.011	0.006	0.001
HEART score 4-6	50%	0.002	-0.005	-0.011	-0.018
	75%	-0.011	-0.015	-0.018	-0.021
	100%	-0.025	-0.025	-0.025	-0.025
HEART score 7-10	50%	0.003	-0.003	-0.009	-0.014
	75%	-0.009	-0.011	-0.014	-0.017
	100%	-0.020	-0.020	-0.020	-0.020
Total	50%	0.004	-0.001	-0.006	-0.011
	75%	-0.002	-0.005	-0.009	-0.012
	100%	-0.008	-0.010	-0.012	-0.014

proportion of missed MACE in the low HEART score category, and further reduces missed MACE in the intermediate and high categories.

Costs

Table 4 shows the average difference in costs per patient between the HEART score strategy and usual care. Costs declined for the low and intermediate HEART score category when compliance and ID increased. A different pattern is observed when the high HEART score category is taken into consideration, where a higher compliance led to higher costs. In the total patient population an ID of at least 50% reduced costs of the HEART score strategy compared to usual care.

Table 4. Average difference in costs per patient between the HEART score strategy and usual care. A negative number represents a reduction in costs. The average cost per patient in usual care was €2,870.

	Compliance	Informed deviation (ID)			
		0%	25%	50%	75%
HEART score 0-3	50%	€ 26	-€ 55	-€ 137	-€ 219
	75%	-€ 148	-€ 186	-€ 224	-€ 262
	100%	-€ 323	-€ 317	-€ 312	-€ 306
HEART score 4-6	50%	€ 7	-€ 114	-€ 235	-€ 356
	75%	-€ 102	-€ 196	-€ 289	-€ 383
	100%	-€ 211	-€ 278	-€ 344	-€ 410
HEART score 7-10	50%	€ 537	€ 613	€ 689	€ 764
	75%	€ 1,465	€ 1,309	€ 1,153	€ 996
	100%	€ 2,393	€ 2,005	€ 1,617	€ 1,228
Total	50%	€ 98	€ 23	-€ 51	-€ 126
	75%	€ 125	€ 43	-€ 38	-€ 119
	100%	€ 151	€ 64	-€ 24	-€ 112

Costs / missed MACE ratio

To gain insight in the monetary investment required to reduce missed MACE, the ratio for the difference in costs and missed MACE between the HEART strategy and usual care is calculated. Table 5 shows the results for the different scenarios of compliance and ID, exhibited for the total patient population. HEART score strategy is considered cost-effective when there are less costs and fewer missed MACE compared to usual care (marked in green in Table 5). HEART score strategy is considered not cost-effective when there are both extra costs and more missed MACE compared to usual care (marked in red in Table 5). When missed MACE could be reduced at higher costs, cost-effectiveness depends on the willingness to pay for reducing missed MACE (marked in black in Table 5).

Table 5. Ratios of the average difference in cost and missed MACE between the HEART score strategy and usual care. Cost-effective scenarios are marked in green, where numbers represent the reduction in costs to prevent one missed MACE. Not cost-effective scenarios are marked in red, where numbers represent the increase in costs for one extra missed MACE. Other scenarios are marked in black, where cost-effectiveness depends on the willingness to pay for preventing missed MACE. Numbers represent the increase in costs to prevent one missed MACE.

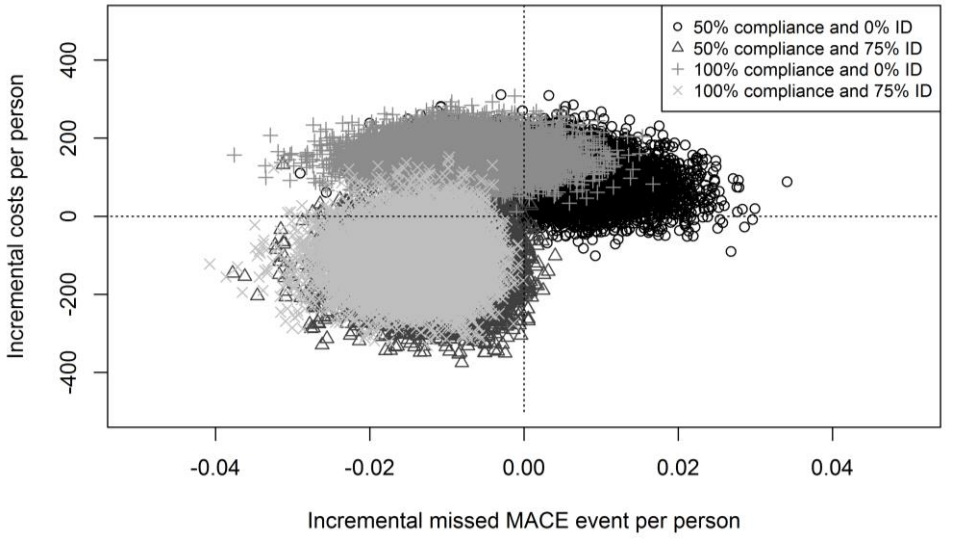
	Compliance	Informed deviation (ID)			
		0%	**25%**	**50%**	**75%**
Total	**50%**	€ 25,946	€ 21,749	€ 8,614	€ 11,648
	75%	€ 64,113	€ 8,099	€ 4,292	€ 9,738
	100%	€ 19,751	€ 6,576	€ 2,092	€ 8,228

The impact of introducing the HEART score strategy on cost per missed MACE depended greatly on the interplay between compliance and ID. For scenarios with a compliance of at least 50% combined with at least 50% ID, costs and missed MACE were both reduced, resulting in a promising (i.e. cost-effective) strategy.

Probabilistic sensitivity analysis

Figure 2 shows the incremental cost-effectiveness plane of four scenarios (compliance of 50% / 100% and ID of 0% / 75%) of the HEART score strategy compared to usual care. In the scenario with 50% compliance with (and 0% ID from) management recommendations, 68% of simulations resulted in an outcome that would be considered not cost-effective. This means that there are both more missed MACE, and higher costs for the HEART score strategy compared to usual care. When 100% compliance (and 0% ID) was assumed, there was a reduction in missed MACE, but in all simulations costs per patient were higher. For both the 50% compliance / 75% ID and 100% compliance / 75% ID scenarios, 94% of the simulations resulted both in a reduction in missed MACE and cost savings.

Figure 2. Incremental cost-effectiveness planes for 10,000 simulations (each symbol represents one simulation) comparing the HEART score strategy to usual care for the following scenarios: 50% compliance with 0% ID; 50% compliance with 75% ID; 100% compliance with 0% ID; 100% compliance with 75% ID. Note that negative numbers indicate a more desirable outcome (less missed MACE and / or reduction in costs).



Discussion

We have shown how a DAM can be used to estimate the potential health-economic impact of using a diagnostic or prognostic prediction model in practice, using only data and information available before performing a costly, long-term, randomized impact trial. We illustrated this for various hypothetical scenarios if the HEART score prediction model was to be implemented in clinical practice.

Generating a DAM for impact assessment of a prediction model forces researchers to think about its main goals (e.g. reducing the primary outcome, reducing side-effects, optimizing diagnostic and treatment pathways) and how it aims to achieve these goals. DAMs can help demonstrate under what conditions (e.g. amount of required compliance and deviation of model adherence) a prediction model is likely to have the desired impact on health outcomes and/or costs. If it is unlikely that these conditions are going to be

satisfied, then one should consider whether investment in a large-scale prediction model impact trial is justified. (7-9, 40) Should those conditions be deemed plausible, a pilot study or qualitative assessments with experts in the field might be considered to gain more insight and reduce parameter uncertainty. This information can then be used to update the DAM, allowing researchers to re-assess the prediction model's expected consequences. Researchers should ensure using representative and valid input parameters for their DAM, preventing goal-oriented model construction and assessment. In general DAMs should be used for optimizing, the design and conduct of an upcoming impact study. (9, 41)

A DAM can be developed for any type of prediction model to evaluate its potential impact. Table 6 provides a concise overview on how to conduct a model-based impact assessment of a prediction model. The first step is designing the DAM, for which different structures can be chosen, such as a decision tree, Markov model, or micro-simulation model. Next, parameter estimates should be collected, such as probabilities (e.g. transition probability between (health) states), health outcomes (e.g. quality of life), and costs (e.g. cost of diagnostic tests). Feasibility of creating a DAM depends on availability of these data. Analyses can then be run for different scenarios, typically by varying parameters with the greatest uncertainty surrounding them (e.g. compliance and ID for our case study). Alternatively, scenarios can look at other cut-offs for stratifying patients into risk categories. Robustness of outcome measures can be assessed by using Monte Carlo simulation, varying parameter estimates based on the uncertainty surrounding them. In the final step, the results of the DAM can guide the decision on whether a trial to study the impact of a prediction model is warranted. If so, a DAM could provide directions for a pilot study or qualitative assessment before the trial to help optimize its design and conduct. More details on how to develop and analyse DAMs can be found in literature. (17, 42)

Table 6. Guidance for a model-based impact assessment of prediction models, before data on clinical impact have become available. Solid arrows mark the logical sequence in which the steps should be taken. Dotted arrows allow researchers to make adapt and adjust decisions in previous steps, based on newly available information.

Steps	Methods & Sources	Examples from the case study
Design of the decision analytic model	• Decision tree (50) • Markov model (49) • Micro-simulation model (51) • New & current clinical pathway • Guidelines • Protocol / design paper • Expert opinion	➢ Reference (31) ➢ Figure 1 ➢ Appendix A
Collecting parameter estimates and uncertainty	• Development / validation studies • Medical consumption studies • Electronic Health Records • Expert opinion	➢ Reference(22-24) ➢ Reference (29) ➢ Appendix B
Comparative (scenario) analysis	• Probabilistic sensitivity analysis (52) • Best, reasonable, worst case scenarios (53) • One/two/multi-way sensitivity analysis & threshold analysis (17)	➢ Table 3,4 & 5 ➢ Figure 2
Impact trial recommen-dations	• Pilot study • Sample size calculation • Education on prediction model usage • Application of the prediction model in a specific cohort	

To provide insight in the validity of our DAM it is worthwhile to compare its results to those of the impact trial that was performed by Poldervaart et al (27). Unfortunately, health outcomes could not be compared, because a different health outcome was used in the impact trial compared to the DAM (any MACE vs. missed MACE). Furthermore, cost data in the trial were collected over a 3-month time horizon, different from the 6-week time horizon used in literature available before the trial. Still, the impact of non-compliance in our DAM study can be translated to the actual HEART impact trial. The DAM showed that non-compliance without ID in patients with a low HEART score had a detrimental effect on potential cost savings, which was also the main finding of the HEART impact trial: substantial non-compliance in the low HEART score category led to small differences in total cost reduction. This information could have been known before the impact trial, and hence could have been used to support a more efficient design and conduct by, for example, assessing potential compliance of physicians beforehand using interviews, or performing a pilot study.

Few other DAM-based assessments have been previously performed that assess the potential impact of prediction models before an empirical impact study has been executed. One study assessed the value of a prediction model for predicting shoulder pain in patients with early stage oral cavity squamous cell carcinoma after surgical removal of lymph nodes. (43) Although the analysis did focus on specific scenarios regarding the accuracy of predictions, compliance or additional clinical expertise on top of the prediction model were not evaluated. DAM assessments have also been used for headroom analysis, a method that is used to assess the likelihood of potential cost-effectiveness of an intervention, often at very early stage of development, for a given willingness to pay threshold. (19, 44-48) These analyses also make use of data before implementation of an innovation to assess potential benefit. Although a headroom approach is feasible for prediction models, to our knowledge there are no articles on this topic described in literature.

This is one of the first examples in which a DAM was applied for impact assessment of implementing a prediction model in daily practice, using solely data available before conducting a trial. This method can be applied using data that is commonly available after prediction model development and

validation, or can be retrieved from (hospital) databases. Compared to a clinical trial, DAM assessments require a fraction of the time and cost, and could help improve design and conduct of an impact trial, reducing research waste.

There are a few considerations to fully appreciate the findings of the impact assessment in this paper. Use of healthcare resources in our model was based on the first 6 weeks of medical consumption. (29) It is likely that negative consequences from MACE will last beyond this timeframe. Markov chain modelling could account for these long-term effects, however reliable data for these effects were lacking. (49) Out-of-hospital costs, such as general practitioner visits, medication usage, and non-medical costs (e.g. labour productivity losses, traveling expenses) were not included in the assessment. Although these are likely to influence the incremental costs and health from a societal perspective, the general conclusions will likely be similar.

We viewed the HEART score purely as a diagnostic tool, which implies that using the HEART score cannot prevent MACE. It can only optimize correct stratification of patients and streamline subsequent management. Because MACE is not prevented, the natural outcome is missed MACE, associated with poorer outcome and additional costs. Others have argued that the HEART score can also be used to predict future MACE, opening the opportunity to prevent it. This would of course lead to a rather different DAM. We chose not to do this because HEART was designed for use in an acute care setting, where patients present with chest pain, which is clearly a diagnostic setting.

A decision analytic model is ideal for assessing the expected impact of using a prediction model in clinical practice on patient health outcomes and healthcare costs, using solely data available before conducting an empirical long-term randomized impact study. With the results of such DAMs one can decide whether an empirical impact trial is still deemed necessary, and if so, under what conditions such prediction model is likely to show favourable results. Efforts can then be directed at improving the use of the prediction model by clinicians and on improving the trial design. In general, DAMs can provide insight in the mechanism through which a prediction model and its risk based management recommendations can lead to desired results, and

expose potentials flaws in mechanistic pathways, allowing researchers to adapt the design of an empirical trial beforehand. Ultimately model-based impact assessments have the potential to reduce research waste, by more efficient selection of prediction models in which an empirical impact trial is warranted.

References

1. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7. Epub 2015/05/20.

2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. Epub 2013/02/09.

3. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56(9):826-32. Epub 2003/09/25.

4. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605. Epub 2009/05/30.

5. Dharmarajah B, Thapar A, Salem J, Lane TR, Leen EL, Davies AH. Impact of risk scoring on decision-making in symptomatic moderate carotid atherosclerosis. The British journal of surgery. 2014;101(5):475-80. Epub 2014/03/13.

6. van Montfort P, Willemse JP, Dirksen CD, van Dooren IM, Meertens LJ, Spaanderman ME, et al. Implementation and Effects of Risk-Dependent Obstetric Care in the Netherlands (Expect Study II): Protocol for an Impact Study. JMIR research protocols. 2018;7(5):e10066. Epub 2018/05/08.

7. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606. Epub 2009/06/09.

8. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8. Epub 2012/03/09.

9. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagnostic and Prognostic Research. 2018;2(1):11.

10. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73. Epub 2015/01/07.

11. 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. Epub 2011/10/18.

12. van Giessen A, Peters J, Wilcher B, Hyde C, Moons C, de Wit A, et al. Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2017;20(4):718-26. Epub 2017/04/15.

13. Wallace E, Uijen MJ, Clyne B, Zarabzadeh A, Keogh C, Galvin R, et al. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. BMJ Open. 2016;6(3):e009957. Epub 2016/03/24.

14. Kappen TH, Moons KG, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, van Klei WA. Impact of risk assessments on prophylactic antiemetic prescription and the incidence of postoperative nausea and vomiting: a cluster-randomized trial. Anesthesiology. 2014;120(2):343-54. Epub 2013/10/10.

15. Huang DT, Yealy DM, Filbin MR, Brown AM, Chang CH, Doi Y, et al. Procalcitonin-Guided Use of Antibiotics for Lower Respiratory Tract Infection. N Engl J Med. 2018;379(3):236-49. Epub 2018/05/22.

16. Snooks H, Bailey-Jones K, Burge-Jones D, Dale J, Davies J, Evans B, et al. Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC). Southampton (UK)2018.

17. Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the economic evaluation of health care programme. 3rd ed. Oxford: Oxford University Press; 2005.

18. IJzerman M, Koffijberg H, Fenwick E, Krahn M. Emerging Use of Early Health Technology Assessment in Medical Product Development: A Scoping Review of the Literature. PharmacoEconomics. 2017;35(7):727-40. Epub 2017/04/23.

19. McAteer H, Cosh E, Freeman G, Pandit A, Wood P, Lilford R. Cost-effectiveness analysis at the development phase of a potential health technology: examples based on tissue engineering of bladder and urethra. Journal of tissue engineering and regenerative medicine. 2007;1(5):343-9. Epub 2007/11/27.

20. Greving JP, Buskens E, Koffijberg H, Algra A. Cost-effectiveness of aspirin treatment in the primary prevention of cardiovascular disease events in subgroups based on age, gender, and varying cardiovascular risk. Circulation. 2008;117(22):2875-83. Epub 2008/05/29.

21. Miquel-Cases A, Steuten LM, Retel VP, van Harten WH. Early stage cost-effectiveness analysis of a BRCA1-like test to detect triple negative breast cancers responsive to high dose alkylating chemotherapy. Breast. 2015;24(4):397-405. Epub 2015/05/06.

22. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation. 2008;16(6):191-6. Epub 2008/07/31.

23. Backus BE, Six AJ, Kelder JC, Bosschaert MA, Mast EG, Mosterd A, et al. A prospective validation of the HEART score for chest pain patients at the emergency department. International journal of cardiology. 2013;168(3):2153-8. Epub 2013/03/08.

24. Backus BE, Six AJ, Kelder JC, Mast TP, van den Akker F, Mast EG, et al. Chest pain in the emergency room: a multicenter validation of the HEART Score. Crit Pathw Cardiol. 2010;9(3):164-9. Epub 2010/08/31.

25. Streitz MJ, Oliver JJ, Hyams JM, Wood RM, Maksimenko YM, Long B, et al. A retrospective external validation study of the HEART score among patients presenting to the emergency department with chest pain. Internal and emergency medicine. 2018;13(5):727-48. Epub 2017/09/13.

26. Oliver JJ, Streitz MJ, Hyams JM, Wood RM, Maksimenko YM, Long B, et al. An external validation of the HEART pathway among Emergency Department patients with chest pain. Internal and emergency medicine. 2018. Epub 2018/03/08.

27. Poldervaart JM, Reitsma JB, Backus BE, Koffijberg H, Veldkamp RF, Ten Haaf ME, et al. Effect of Using the HEART Score in Patients With Chest Pain in the Emergency Department: A Stepped-Wedge, Cluster

Randomized Trial. Ann Intern Med. 2017;166(10):689-97. Epub 2017/04/25.

28. Kappen TH, van Loon K, Kappen MA, van Wolfswinkel L, Vergouwe Y, van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. J Clin Epidemiol. 2016;70:136-45. Epub 2015/09/25.

29. Nieuwets A, Poldervaart JM, Reitsma JB, Buitendijk S, Six AJ, Backus BE, et al. Medical consumption compared for TIMI and HEART score in chest pain patients at the emergency department: a retrospective cost analysis. BMJ Open. 2016;6(6):e010694. Epub 2016/06/18.

30. Six AJ, Cullen L, Backus BE, Greenslade J, Parsonage W, Aldous S, et al. The HEART score for the assessment of patients with chest pain in the emergency department: a multinational validation study. Crit Pathw Cardiol. 2013;12(3):121-6.

31. Poldervaart JM, Reitsma JB, Koffijberg H, Backus BE, Six AJ, Doevendans PA, et al. The impact of the HEART risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. BMC cardiovascular disorders. 2013;13:77. Epub 2013/09/28.

32. Soekhlal RR, Burgers LT, Redekop WK, Tan SS. Treatment costs of acute myocardial infarction in the Netherlands. Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation. 2013;21(5):230-5. Epub 2013/03/05.

33. Bekelman JE, Halpern SD, Blankart CR, Bynum JP, Cohen J, Fowler R, et al. Comparison of Site of Death, Health Care Utilization, and Hospital Expenditures for Patients Dying With Cancer in 7 Developed Countries. JAMA. 2016;315(3):272-83. Epub 2016/01/20.

34. Mahoney EM, Wang K, Cohen DJ, Hirsch AT, Alberts MJ, Eagle K, et al. One-year costs in patients with a history of or at risk for atherothrombosis in the United States. Circulation Cardiovascular quality and outcomes. 2008;1(1):38-45. Epub 2008/09/01.

35. Osnabrugge RL, Magnuson EA, Serruys PW, Campos CM, Wang K, van Klaveren D, et al. Cost-effectiveness of percutaneous coronary intervention versus bypass surgery from a Dutch perspective. Heart. 2015;101(24):1980-8. Epub 2015/11/11.

36. Baim D, Grossman W. Complications of cardiac catheterization. Baltimore Williams & Wilkins; 1996.

37. Johnson LW, Lozner EC, Johnson S, Krone R, Pichard AD, Vetrovec GW, et al. Coronary arteriography 1984-1987: a report of the Registry of the Society for Cardiac Angiography and Interventions. I. Results and complications. Catheterization and cardiovascular diagnosis. 1989;17(1):5-10. Epub 1989/05/01.

38. Wyman RM, Safian RD, Portway V, Skillman JJ, McKay RG, Baim DS. Current complications of diagnostic and therapeutic cardiac catheterization. Journal of the American College of Cardiology. 1988;12(6):1400-6. Epub 1988/12/01.

39. van der Zalm P. HEART-score biedt arts extra handvat. Zorgvisie; 2016 [31-10-2018]; Available from: https://www.zorgvisie.nl/heart-score-biedt-arts-extra-handvat/.

40. Lilford RJ, Chilton PJ, Hemming K, Girling AJ, Taylor CA, Barach P. Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. BMJ. 2010;341:c4413.

41. Tuffaha HW, Gordon LG, Scuffham PA. Value of information analysis in healthcare: a review of principles and applications. Journal of medical economics. 2014;17(6):377-83. Epub 2014/03/22.

42. ISPOR. Synthesizing Health Care Evidence - Modelling/Decision Analysis Methods. 2018 [20-12-2018]; Available from: https://www.ispor.org/heor-resources/more-heor-resources/outcomes-research-guidelines-index.

43. Govers TM, Rovers MM, Brands MT, Dronkers EAC, Baatenburg de Jong RJ, Merkx MAW, et al. Integrated prediction and decision models are valuable in informing personalized decision making. J Clin Epidemiol. 2018. Epub 2018/09/01.

44. Girling A, Lilford R, Cole A, Young T. Headroom Approach to Device Development: Current and Future Directions. International journal of technology assessment in health care. 2015;31(5):331-8. Epub 2015/12/24.

45. Girling A, Young T, Brown C, Lilford R. Early-stage valuation of medical devices: the role of developmental uncertainty. Value in health : the

journal of the International Society for Pharmacoeconomics and Outcomes Research. 2010;13(5):585-91. Epub 2010/04/24.

46. Van Nimwegen KJ. Feasibility of the Headroom Analysis in Early Economic Evaluation of Innovative Diagnostic Technologies With no Immediate Treatment Implications. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2014;17(7):A550. Epub 2014/11/01.

47. Cosh E, Girling A, Lilford R, McAteer H, Young T. Investing in new medical technologies: A decision framework. Journal of commercial biotechnology. 2007;13(4):263-71.

48. Miquel-Cases A, Retel VP, Lederer B, von Minckwitz G, Steuten LM, van Harten WH. Exploratory Cost-Effectiveness Analysis of Response-Guided Neoadjuvant Chemotherapy for Hormone Positive Breast Cancer Patients. PLoS One. 2016;11(4):e0154386. Epub 2016/04/29.

49. Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. PharmacoEconomics. 1998;13(4):397-409. Epub 1998/03/08.

50. Ryder HF, McDonough C, Tosteson AN, Lurie JD. Decision Analysis and Cost-effectiveness Analysis. Seminars in spine surgery. 2009;21(4):216-22. Epub 2009/12/01.

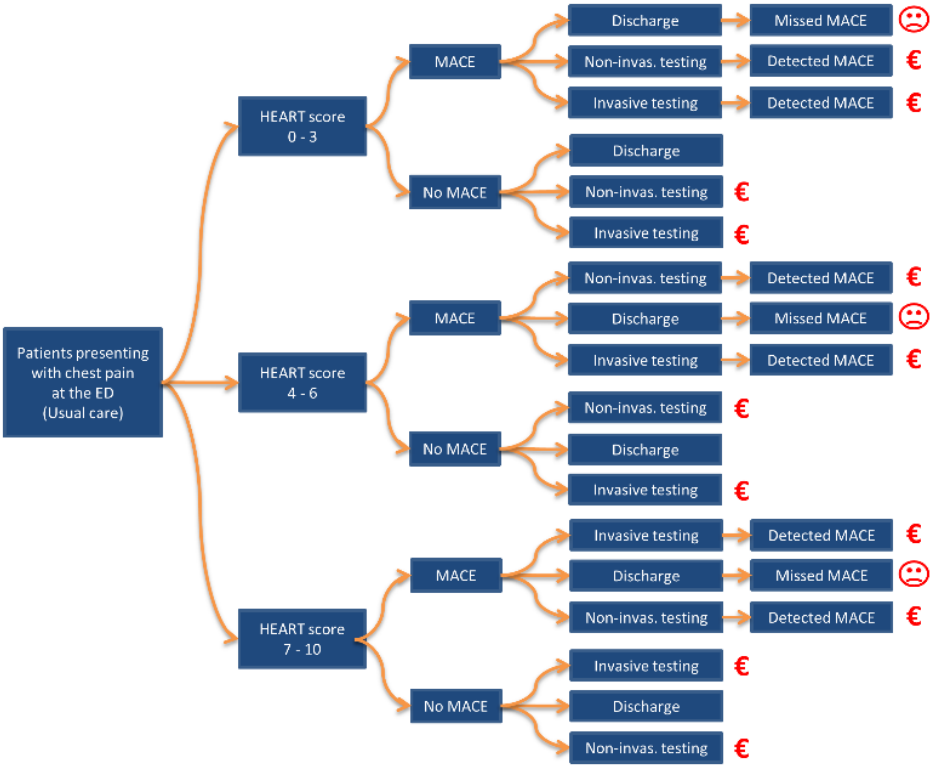
51. Zucchelli E, Jones AM, Rice N. The evaluation of health policies through dynamic microsimulation methods. International Journal of Microsimulation. 2012;5(1):2-20.

52. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health economics. 2005;14(4):339-47. Epub 2005/03/01.

53. Briggs AH, Gray AM. Handling uncertainty when performing economic evaluation of healthcare interventions. Health technology assessment. 1999;3(2):1-134. Epub 1999/08/17.

Appendix A

The decision tree for the usual care strategy. HEART scores are given in this decision tree, as they can be calculated, but in current practice they are not provided to clinicians, nor are the management recommendations attached to these scores. Compliance and ID are not present in the usual care strategy, as there is no HEART score to comply to.



Appendix B

Input parameters used for the decision analytic model for evaluation of the HEART score

Parameter	Transition probability	Standard error	Distribution	Source
HEART score 0-3	**0.400**	**0.019**	**Dirichlet**	**(29)**
Probability at least one MACE < 6 weeks	**0.020**	**0.009**	**Beta**	**(29)**
• Compliance				
➢ Discharged	1.000	-	-	Conse-quential
➢ Average # of MACE	1.000	0.000	Gamma	(29)
• Non-compliance				
➢ Non-invasive testing	0.750	0.217	Beta	(29)
➢ Average # of MACE	1.333	0.333	Gamma	(29)
➢ Invasive testing	0.250	0.217	Beta	(29)
➢ Average # of MACE	1.000	0.000	Gamma	(29)
Probability of no MACE < 6 weeks	**0.980**	**0.009**	**Beta**	**(29)**
• Compliance				
• Discharged	1.000	-	-	Conse-quential
• Non-compliance				
• Non-invasive testing	0.948	0.021	Beta	(29)
• Invasive testing	0.052	0.021	Beta	(29)
HEART score 4-6	**0.444**	**0.020**	**Dirichlet**	**(29)**
Probability at least one MACE < 6 weeks	**0.236**	**0.025**	**Beta**	**(29)**
• Compliance				
➢ Non-invasive testing	1.000	-	-	Conse-quential
➢ Average # of MACE	1.194	0.067	Gamma	(29)
• Non-compliance				
➢ Discharged	0.226	0.075	Beta	(29)
➢ Average # of MACE	1.000	0.000	Gamma	(29)
➢ Invasive testing	0.774	0.075	Beta	(29)

Parameter	Cost	Standard error	Distribution	Source
➢ Average # of MACE	1.167	0.078	Gamma	(29)
Probability of no MACE < 6 weeks	**0.764**	**0.025**	**Beta**	**(29)**
• Compliance				
➢ Non-invasive testing	1.000	-	-	Conse-quential
• Non-compliance				
➢ Discharged	0.806	0.041	Beta	(29)
➢ Invasive testing	0.194	0.041	Beta	(29)
HEART score 7-10	**0.156**	**0.014**	**Dirichlet**	**(29)**
Probability at least one MACE < 6 weeks	**0.590**	**0.049**	**Beta**	**(29)**
• Compliance				
➢ Invasive testing	1.000	-	-	Conseq uential
➢ Average # of MACE	1.333	0.094	Gamma	(29)
• Non-compliance				
➢ Discharged	0.077	0.052	Beta	(29)
➢ Average # of MACE	1.000	0.000	Gamma	(29)
➢ Non-invasive testing	0.923	0.052	Beta	(29)
➢ Average # of MACE	1.458	0.104	Gamma	(29)
Probability of no MACE < 6 weeks	**0.410**	**0.049**	**Beta**	**(29)**
• Compliance				
➢ Invasive testing	1.000	-	-	Conse-quential
• Non-compliance				
➢ Discharged	0.132	0.055	Beta	(29)
➢ Non-invasive testing	0.868	0.055	Beta	(29)

Parameter	Cost	Standard error	Distribution	Source
HEART score 0-3				
• Non-invasive testing	€ 558	€ 87	Gamma	(29)
• Invasive testing	€ 2,900	€ 850	Gamma	(29)
HEART score 4-6				
• Non-invasive testing	€ 1,458	€ 152	Gamma	(29)
• Invasive testing	€ 5,729	€ 648	Gamma	(29)
HEART score 7-10 non-invasive				
• Non-invasive testing	€ 2,701	€ 287	Gamma	(29)
• Invasive testing	€ 6,145	€ 448	Gamma	(29)
MACE	€ 5,484	€ 291	Gamma	Weight. average
• Acute myocardial infarction (AMI)	€ 5,800			(32)
• Percutaneous coronary intervention (PCI)	€ 6,850			(35)
• Coronary artery bypass grafting (CABG)	€ 5,621			(35)
• Conservative management	€ 315			(34)
• Death	€ 2,552			(33)

"We never really save a life; we only postpone death"

Karl Claxton, health economist

Chapter 7

When Is Pursuing an Innovative Idea Worthwhile? A Model-Based Approach

A. Kluytmans

J. Deinum

K. Jenniskens

A.E. van Herwaarden

J. Gloerich

A.J. van Gool

G.J. van der Wilt

J. Grutters

Clin Chem Lab Med. 2019 Jul 9.

Abstract

Background: With seemingly unlimited technological possibilities yet limited budgets, clinicians face the challenge of which novel ideas to pursue and which to lay aside. Although health economic modelling methods may support innovation decisions, they are not yet widely known or used. Our aim was to illustrate early health economic modelling to clinicians by applying its methods to the case of diagnosing primary aldosteronism (PA) in patients with hypertension.

Methods and Findings: we developed a cohort state-transition model to simulate diagnosis, treatment, and long-term health outcomes for patients aged ≥40 years with resistant hypertension suspected of PA. We included relevant literature and Dutch costing data and took a lifetime, societal perspective on costs and health effects (quality-adjusted life-years, QALYs). In our model we compared the current aldosterone-to-renin ratio test for diagnosing PA to a hypothetical new test. During a patient's lifetime, a perfect diagnostic test would yield 0.027 QALYs and increase costs by €43. At a cost-effectiveness threshold of €20,000 per QALY, the maximum price for this perfect test to be cost-effective is €498 (95% CI: €275 - €808). The value of the perfect test was most strongly influenced by the sensitivity of the current biomarker test. Threshold analysis showed the novel test needs a sensitivity of at least 0.9 and a specificity of at least 0.7 to be cost-effective.

Conclusions: Applying a model-based approach to determine the added value of a clinical innovation in PA diagnostics, we demonstrated there was room for improvement while indicating a maximum price per test, supporting the conclusion that early health economic modelling is useful and feasible in clinical practice to determine the cost-effectiveness of novel ideas prior to extensive development activities and clinical implementation. More applications of early modelling through collaborations between health economists and clinical experts will illustrate the benefits and help further the accessibility of early health economic modelling in dealing with innovation.

Introduction

In clinical medicine, the past decades have been an era of health technology innovation (5). Inventions such as genome sequencing, PET scans, stents, and biologicals have considerably improved the diagnosis and treatment of numerous medical conditions. However undeniable the advantages of health technology innovation are (6), the advances coincide with a steep rise in health care expenditure (7, 8). Around the globe, studies predict that health care costs will consume around one-third of family incomes by the year 2040 (9, 10), which would challenge the accessibility of health care and social solidarity (11). In addition, it is not always clear whether innovations provide 'true value for money' (12). The need for decision-making in innovation is therefore evident.

Efficient methods to assess the added clinical value of novel concepts – prior to committing considerable resources to their development – could help decide which new ideas to pursue and which to lay aside. Health technology assessments or health economic modelling provide such methods. They are increasingly being applied at the early stages of technology development to inform decisions regarding the further development of potentially innovative ideas (13). Even when there is no concrete data on the innovation, methodologies such as headroom analysis afford estimates of its potential by assessing the room for improvement in current care practices (14-16).

Novel ideas are often proposed by scientists or clinicians who tend to be unfamiliar with health economic modelling concepts. Using the example of diagnosing primary aldosteronism (PA), we will illustrate how such early modelling methods can help clinicians and developers make evidence-informed decisions on whether or not to pursue a new diagnostic approach.

In the field of PA, new concepts are being considered out of concern for the current diagnostic method, which centres around the aldosterone-to-renin ratio (ARR) test. First, renin's variability is a cause for concern (17). The different recommended cut-off values for the ARR in the Endocrine Society Clinical Practice Guideline introduce further variability (18). Second, for reliable use as a screening test – the Japan Endocrine Society even recommends to screen all hypertensive patients for PA – the test's sensitivity

should be perfect to avoid false negatives (19). Doubts regarding the ARR's suitability are evident from the many studies that investigated its diagnostic properties, indicating a repeated desire to evaluate the test (20-29). Third, optimization of the PA diagnosis is an active field of research, also implying concerns regarding current diagnostic methods, as is illustrated by ENSAT-HT and PRIMAL, two ongoing clinical trials (30, 31), and the Berge et al (2015) and Rehan et al (2015) studies investigating the use of liquid chromatography-mass spectrometry (LC-MS) (32, 33).

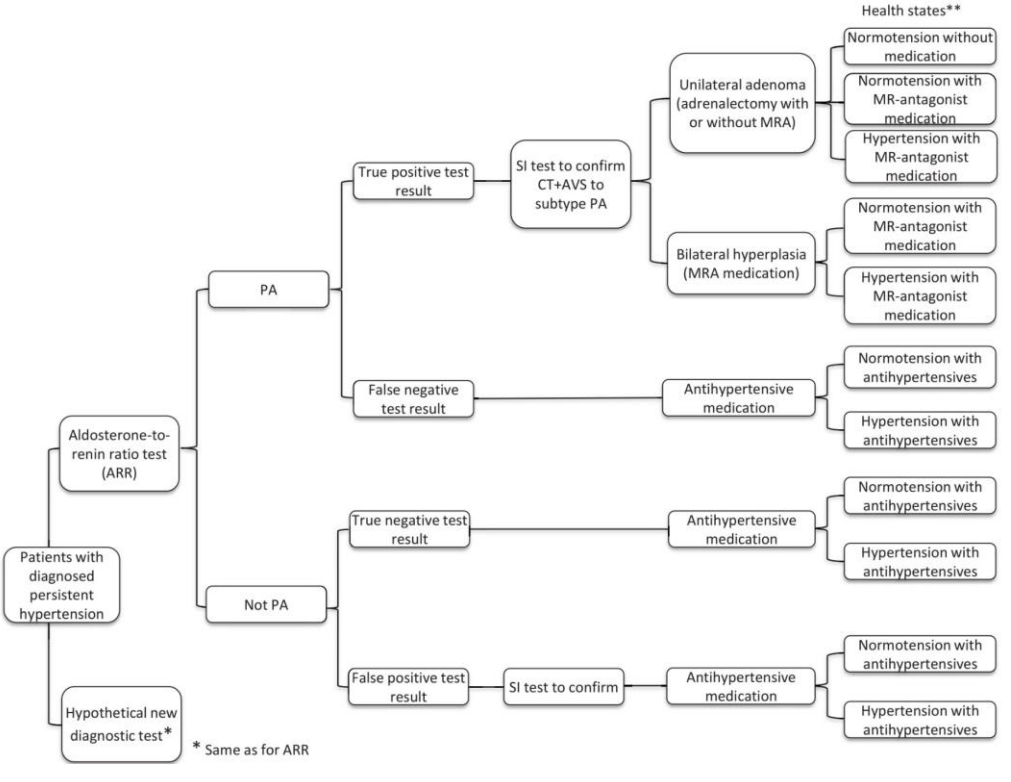
We believe novel tests for PA, such as a targeted LC-MS protein test, are highly relevant for an early health economic assessment prior to their (further) development and clinical validation to thus estimate their potential value. Through decision modelling, we analysed the room for improvement in current PA testing and determined what properties, in terms of sensitivity, specificity, and costs, a new test should have in order to be clinically valid as well as cost-effective.

Methods

Model Overview and Validation

To simulate the expected costs and health benefits of various diagnostic strategies, we constructed a health economic model (see Appendix A) that consists of a decision tree for the diagnostic part (Figure 1) and a cohort state-transition model (see Appendix A) to simulate the long-term health consequences (Appendix B). The model was conceptualized and optimized based on the literature and the input of various experts in the areas of hypertension, laboratory medicine, and technology assessment, and developed and analysed using Microsoft Office Excel 2016.

Figure 1. Decision tree for the diagnosis of primary aldosteronism (PA), comparing the current aldosterone-to-renin ratio test to a hypothetical diagnostic test, leading to one of five health states that form the starting point of the long-term part of the model that simulates yearly costs and health effects (see Appendix B).



Target Population

The target population consists of patients diagnosed with resistant hypertension who are suspected of PA. We assume that prior attempts to treat their hypertension with antihypertensive medication have failed and that patients are being monitored by a hypertension specialist. In our model, the starting age of our patient cohort is 40, with a lifetime model being created to reflect a lifelong perspective on costs and effects.

Model Structure

In the decision tree, the patients are divided into those with and those without PA, where patients with PA can have a (false) negative or (true) positive test result and those without PA a (true) negative or (false) positive test result. Every positive test result is confirmed by a saline-infusion (SI) test (which is considered the criterion reference test for PA) (18). In case of a positive SI test, a CT scan plus adrenal vein sampling (AVS) is used to subtype PA. Following their diagnosis, the patients receive treatment: adrenalectomy and/or medication for those patients with a confirmed PA and continuation of antihypertensive medication for those not diagnosed with PA. Following treatment, patients either become normotensive, with or without continued use of their medication, or remain hypertensive.

The long-term part of the model is shown in Appendix B. All patients may develop cardiovascular complications (i.e., stroke, coronary artery disease, atrial fibrillation, or heart failure). The risk is lowest for the normotensive group, higher for the hypertensive, and highest for the PA groups (34). Once affected by a cardiovascular complication, patients move to a post-complication health state which is associated with higher health care costs and a lower quality of life compared to their starting health states. At every cycle, patients in the cohort may die from cardiovascular or other causes.

The decision tree and state-transition model are equally structured for the two strategies (ARR test and hypothetical test), but the parameters that are assigned to the model differ (see model input).

Assumptions

We assume that SI tests are always conducted to confirm the PA diagnosis and that they are 100% accurate. The assumed success rate for antihypertensive medication is based on a 50% treatment adherence rate (35). Due to scarce data on the long-term development of hypertension in this population, we assume that once patients are assigned to normotension or hypertension health states, they will remain normotensive or hypertensive until they develop cardiovascular complications or die. Regarding cardiovascular complications, our modelled cohort could only develop one complication during the remaining life-span due to scarce data on the

probability and consequences of developing multiple cardiovascular complications within this specific population.

Model input

All model inputs and their sources are listed in Appendix C.

Transition probabilities

In the diagnostic part of our model, we assume a .15 prevalence of PA, which rate was reported by Jansen et al (2014) and fits in with the literature on PA in patients with essential hypertension (28, 36). Because prevalence estimates differ across health- care settings, we varied this parameter to investigate the relationship between PA prevalence and the headroom for a diagnostic test (see Analyses) (37). For the aldosterone-to-renin ratio test, we assume a sensitivity of .89 and a specificity of .96 based on a meta-analysis conducted by Li et al (2016) (29). For confirmed PA patients we assume a probability of .47 of having an aldosterone-producing adenoma that qualifies for surgery after the Shah et al 2014 study (38). The outcomes of adrenalectomy are based on a recent study by Williams et al (2017), who reported that 37% of their patients became normotensive without further need for medication and another 47% following antihypertensive medication, while 16% remained hypertensive even with antihypertensive medication (39). For all the patients that did not have an aldosterone-producing adenoma suitable for surgery, we assume that the probability to become either normotensive or remain hypertensive is equal to the patients' adherence to their medication regimen, which Azizi et al (2016) estimated to be .5 (35).

In the long-term part of our model, we based the probability for our patients to develop one of four cardiovascular complications on publically available data on normotension and hypertension issued by the Dutch Ministry of Health. For the PA patients, odds ratios published by Monticone et al (2018) were used to calculate their increased risk of cardiovascular events compared to hypertensive patients (34). The probability of dying from cardiovascular causes was derived from various studies, where we distinguished between the year of the incident and all subsequent years (40-43). The probability of

dying from other causes was based on publically available data from Statistics Netherlands (44).

Utilities

To model health effects in terms of quality-adjusted life-years (QALYs, see Appendix A), we retrieved utilities (see Appendix A) for every health state from the relevant literature. QALYs were discounted at a rate of 1.5% annually to reflect the Dutch time preference for health effects (45).

Costs

In the model, costs were determined from a societal perspective and based on Dutch sources, comprising one-time costs of the various tests, the yearly costs of medication, and those associated with the four cardiovascular events. All costs were converted to 2016-prices and discounted at a rate of 4% (45, 46).

Analyses

Based on the transition probabilities pertaining to the PA diagnosis, the cohort data was divided across health states (See Figure 1 and Appendix B). Using yearly cycles, we calculated the costs and QALYs proportional to either remaining in the starting health states or a progression to cardiovascular events or death. We next computed the total accumulated costs and QALYs per patient at 10 years, 20 years, and lifetime, and compared the costs and QALYs of the current test with those of the hypothetical test. We used these per-patient totals in our subsequent headroom (14-16), sensitivity (3, 4), and threshold analyses.

Our headroom analysis (see Appendix A) estimates the (financial) room for improvement in the current diagnostic pathway of patients with PA by comparing the ARR to a hypothetical, perfectly accurate test with a sensitivity and specificity of 1. The difference in QALYs between the current and the perfect diagnostic procedure is known as the effectiveness gap, or the health effects foregone by imperfections in the current protocol. Headroom, then, is calculated by monetizing the effectiveness gap using a cost-effectiveness threshold – in our case €20,000 per QALY, following the Dutch recommendations for diagnostic and preventive interventions – and

adding or subtracting any differences in costs between the current and perfect diagnostic strategy. The resulting headroom can be interpreted as the maximum price at which a perfectly accurate PA test could be considered cost-effective. Given the lifelong consequences of cardiovascular complications, a lifetime horizon on costs and health benefits is most appropriate. We will also report the main results at 10 and 20 years after the diagnosis.

Many of the parameters in our model suffer from a degree of uncertainty. A PSA (see Appendix A) was conducted using 10,000 samples from beta distributions. Standard errors for probabilistically varied parameters are listed in Appendix C. Results of the PSA are reported using the percentile method, yielding ranges similar to 95% confidence intervals. Further, a univariate sensitivity analysis was conducted to investigate the individual impact of the following parameters: the sensitivity and specificity of the ARR (varied from 0.22 to 1 and 0.56 to 1 respectively, based on the ranges reported in a meta-analysis by Li et al (29)), the prevalence of PA (varied from 5 to 25%), and the cost-effectiveness threshold (varied from €20,000 to €80,000). The mentioned ARR sensitivity and specificity values were selected because of the heterogeneous figures reported in the literature and the PA prevalence rate and cost-effectiveness threshold based on potential differences across health care settings.

Since a novel test for PA is unlikely to be perfect and costless, besides the headroom analysis we also performed a multivariate threshold analysis to investigate trade-offs between a novel test's sensitivity, specificity, and price. In this analysis sensitivity and specificity were simultaneously varied, both ranging from 0 to 1.

Results

Headroom analysis

The average QALYs per patient of the ARR strategy with a lifetime horizon were 21.249 versus 21.276 for the perfect diagnostic strategy (Table 1). The average costs were €17,779 versus €17,822. Compared to the current test, a

Table 1. Results of the headroom analysis as averages per patient. Headroom was calculated with a cost-effectiveness threshold of €20,000 per QALY.

Horizon	Current test		Perfect test		Increments		Headroom*
	QALYs	Costs	QALYs	Costs	QALYs	Costs	(95% CI)
10-years	7.721	€6,867	7.724	€6,915	0.003	€48	€19 (-€9 - €62)
20-years	13.692	€12,245	13.703	€12,280	0.011	€35	€181 (€90 - €320)
Lifetime	21.249	€17,779	21.276	€17,822	0.027	€43	€498 (€275 - €808)

perfect test would yield 0.027 QALYs – or nearly 10 days in perfect health – at a cost increase of €43. At our cost-effectiveness threshold of €20,000 per QALY gained, the resulting headroom is 0.027*€20,000-€43, or €498 (95% CI: €275 to €808). The QALY gain – and therefore the headroom – decreases for the 20- and 10-year time horizons.

Univariate sensitivity analyses

The lifetime headroom results were further investigated whereby key parameters of interests were varied one by one (Table 2). The sensitivity of the current test had the largest impact on the headroom estimate, followed by the cost-effectiveness threshold. PA prevalence had some impact on the headroom, indicating that an improved test would be most valuable in care settings with relatively many PA patients. The impact of the specificity of the current test on the headroom results was negligible.

Multivariate threshold analysis

The results of the multivariate sensitivity analysis, in which the sensitivity and specificity of the novel test were varied simultaneously, are presented in Figure 3. To exemplify, a novel PA test for hypertensive patients with a sensitivity of .90 and a specificity of 1 might cost €51 more than the current test and still be considered cost-effective compared to the current test. Note that for the current test we assumed a sensitivity of .89 and a specificity of .96 (29). Figure 3 shows that a novel test for diagnosing PA may be worth the extra expenditure when its sensitivity is at least .90 and its specificity is at least .70.

Figure 2. Tornado plot for the four univariate sensitivity analyses, lifetime perspective on costs and effects. †The four parameters of interest are listed on the y-axis, with their ranges specified in brackets. The x-axis contains the headroom values. The bars show the lifetime headroom result across the range of the parameter. For example, when the sensitivity of the current ARR test is varied in the model, the headroom is near zero when the current care has a sensitivity of 1 and reaches €3,500 when the sensitivity of the current ARR test is as low as .22. The base-case headroom result of €498 per patient is taken as a reference value and represented by the red line.

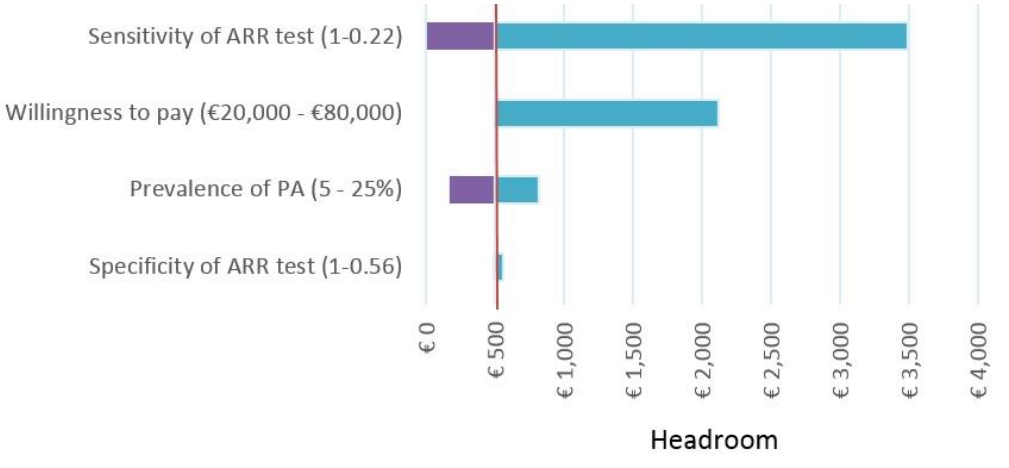


Figure 3. Results of the multivariate threshold analysis. The figure contains model results indicating the monetary loss or gains associated with a hypothetical new test for different specificity (y-axis) and sensitivity (x-axis) combinations. The red area indicates a net loss for the new test and the green area its additional value compared to the current test.

Specificity of the new test \ Sensitivity of the new test	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	-€ 4,122	-€ 3,675	-€ 3,228	-€ 2,781	-€ 2,334	-€ 1,887	-€ 1,440	-€ 993	-€ 546	-€ 99	€ 348
0.1	-€ 4,107	-€ 3,660	-€ 3,213	-€ 2,766	-€ 2,319	-€ 1,872	-€ 1,425	-€ 978	-€ 531	-€ 84	€ 363
0.2	-€ 4,092	-€ 3,645	-€ 3,198	-€ 2,751	-€ 2,304	-€ 1,857	-€ 1,410	-€ 963	-€ 516	-€ 69	€ 378
0.3	-€ 4,077	-€ 3,630	-€ 3,183	-€ 2,736	-€ 2,289	-€ 1,842	-€ 1,395	-€ 948	-€ 501	-€ 54	€ 393
0.4	-€ 4,062	-€ 3,615	-€ 3,168	-€ 2,721	-€ 2,274	-€ 1,827	-€ 1,380	-€ 933	-€ 486	-€ 39	€ 408
0.5	-€ 4,047	-€ 3,600	-€ 3,153	-€ 2,706	-€ 2,259	-€ 1,812	-€ 1,365	-€ 918	-€ 471	-€ 24	€ 423
0.6	-€ 4,032	-€ 3,585	-€ 3,138	-€ 2,691	-€ 2,244	-€ 1,797	-€ 1,350	-€ 903	-€ 456	-€ 9	€ 438
0.7	-€ 4,017	-€ 3,570	-€ 3,123	-€ 2,676	-€ 2,229	-€ 1,782	-€ 1,335	-€ 888	-€ 441	€ 6	€ 453
0.8	-€ 4,002	-€ 3,555	-€ 3,108	-€ 2,661	-€ 2,214	-€ 1,767	-€ 1,320	-€ 873	-€ 426	€ 21	€ 468
0.9	-€ 3,987	-€ 3,540	-€ 3,093	-€ 2,646	-€ 2,199	-€ 1,752	-€ 1,305	-€ 858	-€ 411	€ 36	€ 483
1	-€ 3,972	-€ 3,525	-€ 3,078	-€ 2,631	-€ 2,184	-€ 1,737	-€ 1,290	-€ 843	-€ 396	€ 51	€ 498

Discussion

We applied headroom and threshold analyses to illustrate how early health economic modelling can support decisions regarding health care innovations. Focusing on a hypothetical novel test to support the diagnosis of PA in patients with hypertension, we found some headroom for improving the current diagnostic protocol. Additional research is needed to determine whether a headroom of roughly €500 per patient provides realistic opportunities for the R&D of innovations such as LC-MS techniques. Headroom analyses can similarly be applied to other indications and target populations to obtain an estimate of the potential value of new health technologies – diagnostic or otherwise – before committing resources to their further development and testing. Our analyses may be especially useful in the early stages of biomarker innovation (47).

Decisions to adopt (or decline) innovations are particularly complicated if there is debate about the assessment of current health care practices. While most studies investigating the current ARR test for diagnosing PA in patients with essential hypertension found that its specificity and sensitivity are high, some studies showed that they may be as low as .56 and .22, respectively (26, 28). Through our model-based approach we were able to simulate what the headroom would be given these discrepancies and show that uncertainty regarding ARR's sensitivity would be most influential.

Several choices and assumptions we made in the modelling process may challenge the acceptability of our headroom analysis. Due to the lack of data, we assumed that patients would either remain normotensive or hypertensive without allowing for transitions between the two. Because in reality hypertensive patients can become normotensive with a reduced risk of developing cardiovascular events, our headroom may be overestimated. We also assumed that false negatives would never be identified as having PA, while in reality the persistence of hypertension will likely prompt further testing and, in some cases, the eventual diagnosis of PA. Although a delayed diagnosis is costly as well, our assumption may have resulted in an overestimation of the headroom in that it underestimates the current detection rate of PA. Likewise, the model may overestimate the risk of

cardiovascular events in PA patients because they were computed relative to the risks in a hypertensive population that is likely to also contain undiagnosed PA patients.

Another limitation is that we reduced the complex and multifaceted nature of the diagnostic process to one specific test and assumed that all other tests in the diagnostic pathway are perfectly accurate. While debatable (48), doing so allowed us to single out the effectiveness gap of the ARR test itself and ignore the inaccuracies of tests that are beyond the scope of our headroom analysis. Finally, we based the success rate of the medication regimen in false negatives and non-PA patients on a 50% treatment adherence. If, in reality, the success rate is lower, our headroom is underestimated and, if it is higher, overestimated.

We acknowledge that these – and other – assumptions may affect the support for our model. However, our goal was to illustrate how an exploratory analysis could illustrate whether or not a (diagnostic) innovation might be worthwhile and why. Our model can easily be adapted to incorporate different assumptions or model inputs, for example when new data becomes available.

As such, headroom analysis is one important step in supporting the decision to pursue or abandon a new health care approach. It can also help steer the development of innovation and maximize its added value by assessing how the innovation should be (further) developed or applied to maximize its potential benefit. Also, when multiple innovative approaches compete for the same resources, headroom analysis may help prioritize proposals. What these uses have in common, is that headroom analysis helps identify those problems whose solution would have the biggest impact. In doing so, it helps foster efficient innovation such that it maximizes societal benefit.

We believe early health economic modelling methods such as headroom analysis can help scientists and clinicians decide on innovations before committing considerable resources. On a larger scale, models like ours could help increase the efficiency and yield of innovations. More applications of

early modelling through collaborations between health economists and clinical experts will illustrate their benefits and help further their accessibility.

References

1. Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M. Conceptualizing a Model: A Report of the ISPOR-SMDM Modelling Good Research Practices Task Force-2. Value in Health. 2012;15(6):804-11.

2. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods. 2010;52(1):99-105.

3. Briggs AH, Claxton K, Sculpher MJ. Decision modelling for health economic evaluation: Oxford University Press; 2006.

4. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes: Oxford university press; 2015.

5. Blume SS. Insight and industry: on the dynamics of technological change in medicine: Mit Press; 1992.

6. Cutler DM, McClellan M. Is technological change in medicine worth it? Health Aff. 2001;20(5):11-29.

7. Bodenheimer T. High and rising health care costs. Part 2: technologic innovation. Annals of internal medicine. 2005;142(11):932-7.

8. Sorenson C, Drummond M, Khan BB. Medical technology as a key driver of rising health expenditure: disentangling the relationship. ClinicoEconomics and outcomes research: CEOR. 2013;5:223.

9. Keehan SP, Stone DA, Poisal JA, Cuckler GA, Sisko AM, Smith SD, et al. National Health Expenditure Projections, 2016–25: Price Increases, Aging Push Sector To 20 Percent Of Economy. Health Affairs. 2017;36(3):553-63.

10. [Health Pays Off - Between Choice and Solidarity. The Future of Care]. CPB Netherlands Bureau for Economic Policy Analysis; 2013.

11. Saltman RB. Health sector solidarity: a core European value but with broadly varying content. Israel journal of health policy research. 2015;4(1):5.

12. Lippi G, Mattiuzzi C, Cervellin G. No correlation between health care expenditure and mortality in the European Union. European Journal of Internal Medicine. 2016;32:e13-e4.

13. IJzerman MJ, Koffijberg H, Fenwick E, Krahn M. Emerging Use of Early Health Technology Assessment in Medical Product Development: A Scoping Review of the Literature. PharmacoEconomics. 2017;35(7):727-40.

14. McAteer H, Cosh E, Freeman G, Pandit A, Wood P, Lilford R. Cost-effectiveness analysis at the development phase of a potential health technology: examples based on tissue engineering of bladder and urethra. Journal of tissue engineering and regenerative medicine. 2007;1(5):343-9.

15. Buisman LR, Rutten-van Mölken MP, Postmus D, Luime JJ, Uyl-de Groot CA, Redekop WK. The early bird catches the worm: early cost-effectiveness analysis of new medical tests. International journal of technology assessment in health care. 2016;32(1-2):46-53.

16. Markiewicz K, van Til JA, Steuten LM, IJzerman MJ. Commercial viability of medical devices using Headroom and return on investment calculation. Technological forecasting and social change. 2016;112:338-46.

17. Tomaschitz A, Pilz S. Aldosterone to renin ratio–a reliable screening tool for primary aldosteronism? Hormone and Metabolic Research. 2010;42(06):382-91.

18. Funder JW, Carey RM, Mantero F, Murad MH, Reincke M, Shibata H, et al. The Management of Primary Aldosteronism: Case Detection, Diagnosis, and Treatment: An Endocrine Society Clinical Practice Guideline. The Journal of clinical endocrinology and metabolism. 2016;101(5):1889-916.

19. Nishikawa T, Omura M, Satoh F, Shibata H, Takahashi K, Tamura N, et al. Guidelines for the diagnosis and treatment of primary

aldosteronism-The Japan Endocrine Society 2009. Endocrine journal. 2011;58(9):711-21.

20. Perschel FH, Schemer R, Seiler L, Reincke M, Deinum J, Maser-Gluth C, et al. Rapid screening test for primary hyperaldosteronism: ratio of plasma aldosterone to renin concentration determined by fully automated chemiluminescence immunoassays. Clinical chemistry. 2004;50(9):1650-5.

21. Diederich S, Mai K, Bähr V, Helffrich S, Pfeiffer A, Perschel F. The simultaneous measurement of plasma-aldosterone-and-renin-concentration allows rapid classification of all disorders of the renin-aldosterone system. Experimental and clinical endocrinology & diabetes. 2007;115(07):433-8.

22. Tzanela M, Effremidis G, Vassiliadi D, Szabo A, Gavalas N, Valatsou A, et al. The aldosterone to renin ratio in the evaluation of patients with incidentally detected adrenal masses. Endocrine. 2007;32(2):136-42.

23. Willenberg H, Kolentini C, Quinkler M, Cupisti K, Krausch M, Schott M, et al. The serum sodium to urinary sodium to (serum potassium) 2 to urinary potassium (SUSPPUP) ratio in patients with primary aldosteronism. European journal of clinical investigation. 2009;39(1):43-50.

24. Balaş M, Zosin I, Maser-Gluth C, Hermsen D, Cupisti K, Schott M, et al. Indicators of mineralocorticoid excess in the evaluation of primary aldosteronism. Hypertension Research. 2010;33(8):850.

25. Corbin F, Douville P, Lebel M. Active renin mass concentration to determine aldosterone-to-renin ratio in screening for primary aldosteronism. International journal of nephrology and renovascular disease. 2011;4:115.

26. Fischer E, Beuschlein F, Bidlingmaier M, Reincke M. Commentary on the Endocrine Society Practice Guidelines: Consequences of adjustment of antihypertensive medication in screening of primary aldosteronism. Reviews in Endocrine and Metabolic Disorders. 2011;12(1):43-8.

27. Lonati C, Bassani N, Gritti A, Biganzoli E, Morganti A. Measurement of plasma renin concentration instead of plasma renin activity

decreases the positive aldosterone-to-renin ratio tests in treated patients with essential hypertension. Journal of hypertension. 2014;32(3):627-34.

28. Jansen PM, van den Born B-JH, Frenkel WJ, de Bruijne ELE, Deinum J, Kerstens MN, et al. Test characteristics of the aldosterone-to-renin ratio as a screening test for primary aldosteronism. Journal of Hypertension. 2014;32(1):115-26.

29. Li X, Goswami R, Yang S, Li Q. Aldosterone/direct renin concentration ratio as a screening test for primary aldosteronism: A meta-analysis. Journal of the Renin-Angiotensin-Aldosterone System. 2016;17(3).

30. ENSAT-HT trial [Internet]. Bethesda (MD): U.S. National Library of Medicine; 2018 [Available from: https://clinicaltrials.gov/ct2/show/NCT02772315.

31. PRIMAL trial [Internet]. Bethesda (MD): U.S. National Library of Medicine; 2018 [Available from: https://clinicaltrials.gov/ct2/show/NCT03105531.

32. Berge C, Courand P-Y, Harbaoui B, Paget V, Khettab F, Bricca G, et al. Decreased plasma prorenin levels in primary aldosteronism: potential diagnostic implications. Journal of hypertension. 2015;33(1):118-25.

33. Rehan M, Raizman JE, Cavalier E, Don-Wauchope AC, Holmes DT. Laboratory challenges in primary aldosteronism screening and diagnosis. Clinical biochemistry. 2015;48(6):377-87.

34. Monticone S, D'Ascenzo F, Moretti C, Williams TA, Veglio F, Gaita F, et al. Cardiovascular events and target organ damage in primary aldosteronism compared with essential hypertension: a systematic review and meta-analysis. The Lancet Diabetes & Endocrinology. 2018;6(1):41-50.

35. Azizi M, Pereira H, Hamdidouche I, Gosse P, Monge M, Bobrie G, et al. Adherence to Antihypertensive Treatment and the Blood Pressure–Lowering Effects of Renal Denervation in the Renal Denervation for Hypertension (DENERHTN) Trial. Circulation. 2016;134(12):847-57.

36. Raizman JE, Diamandis EP, Holmes D, Stowasser M, Auchus R, Cavalier E. A renin-ssance in primary aldosteronism testing: obstacles and opportunities for screening, diagnosis, and management. Clinical chemistry. 2015;61(8):1022-7.

37. Käyser SC, Dekkers T, Groenewoud HJ, van der Wilt GJ, Carel Bakx J, van der Wel MC, et al. Study Heterogeneity and Estimation of Prevalence of Primary Aldosteronism: A Systematic Review and Meta-Regression Analysis. The Journal of Clinical Endocrinology & Metabolism. 2016;101(7):2826-35.

38. Shah B, Deshpande S. Assessment of Effect of Diabetes on Health-Related Quality of Life in Patients with Coronary Artery Disease Using the EQ-5D Questionnaire. Value in Health Regional Issues. 2014;3:67-72.

39. Williams TA, Lenders JWM, Mulatero P, Burrello J, Rottenkolber M, Adolf C, et al. Outcomes after adrenalectomy for unilateral primary aldosteronism: an international consensus on outcome measures and analysis of remission rates in an international cohort. The Lancet Diabetes & Endocrinology. 2017;5(9):689-99.

40. Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB, Levy D. Impact of Atrial Fibrillation on the Risk of Death. The Framingham Heart Study. 1998;98(10):946-52.

41. Hankey GJ, Jamrozik K, Broadhurst RJ, Forbes S, Burvill PW, Anderson CS, et al. Five-Year Survival After First-Ever Stroke and Related Prognostic Factors in the Perth Community Stroke Study. Stroke. 2000;31(9):2080-6.

42. Goldberg RJ, Ciampa J, Lessard D, Meyer TE, Spencer FA. Long-term survival after heart failure: A contemporary population-based perspective. Archives of Internal Medicine. 2007;167(5):490-6.

43. Höfer S, Benzer W, Oldridge N. Change in health-related quality of life in patients with coronary artery disease predicts 4-year mortality. International Journal of Cardiology. 2014;174(1):7-12.

44. General mortality rates [Internet]. Den Haag: Statistics Netherlands; 2017 [Available from: https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/output/output/life-tables.

45. [Guideline for Conducting Economic Evaluations in Health Care]. National Health Care Institute; 2015.

46. Frempong SN, Sutton AJ, Davenport C, Barton P. Economic evaluation of medical tests at the early phases of development: a systematic review of empirical studies. Expert review of pharmacoeconomics & outcomes research. 2018;18(1):13-23.

47. Postmus D, Graaf G, Hillege HL, Steyerberg EW, Buskens E. A method for the early health technology assessment of novel biomarker measurement in primary prevention programs. Statistics in medicine. 2012;31(23):2733-44.

48. Cornu E, Steichen O, Nogueira-Silva L, Küpers E, Pagny J-Y, Grataloup C, et al. Suppression of Aldosterone Secretion After Recumbent Saline Infusion Does Not Exclude Lateralized Primary AldosteronismNovelty and Significance. Hypertension. 2016;68(4):989-94.

Appendix A

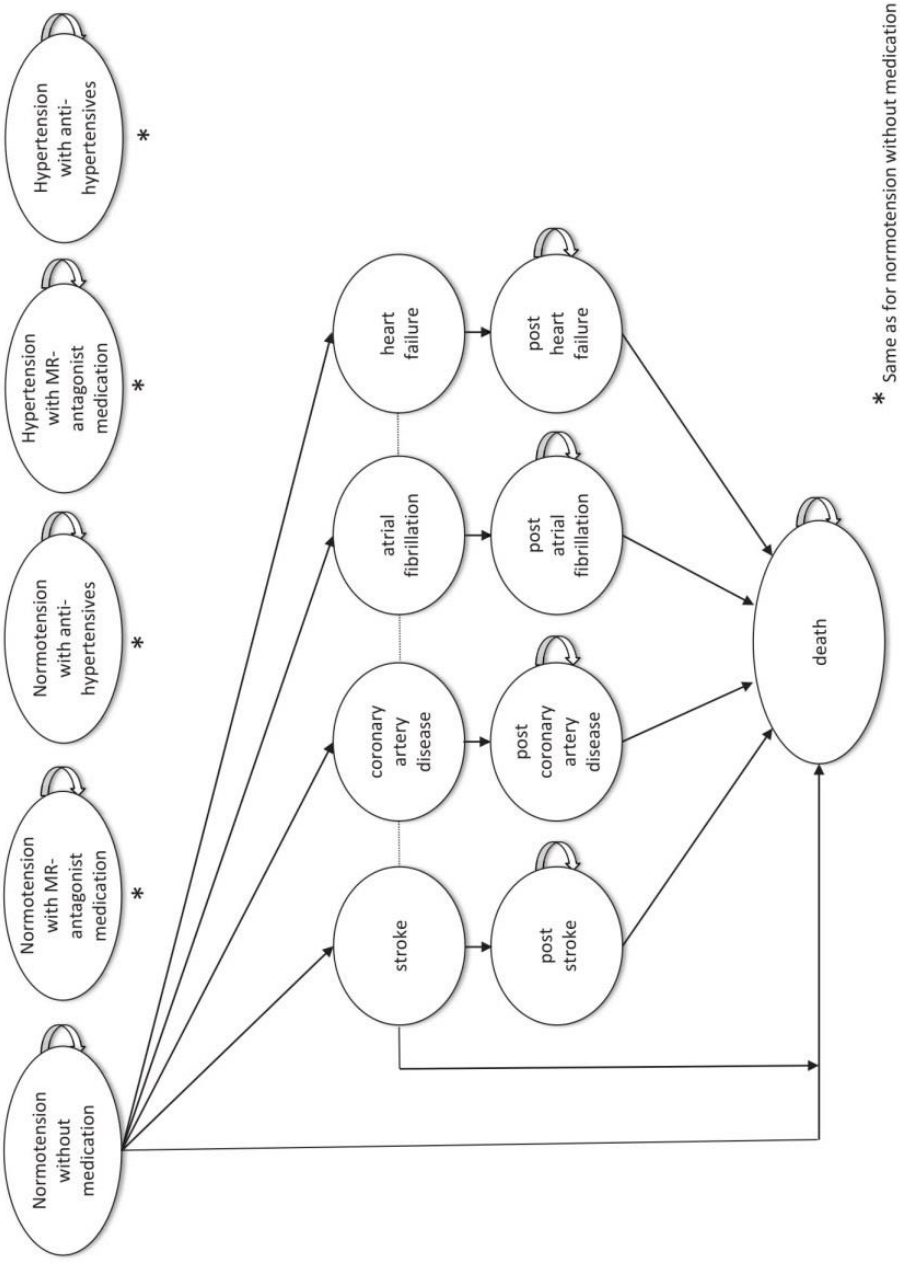
Glossary of terminology surrounding (early) health technology assessment

Term	Meaning
Decision model	A simplified representation of a certain health care setting whose purpose it is to inform health-related decisions by simulating the health effects and costs incurred following a certain health care strategy. (1)
Headroom	The maximum room for improvement within a certain health care strategy. Headroom is calculated by comparing a current care strategy to a hypothetical, perfect care strategy. The difference between the two indicates the room for improvement. Headroom indicates the potential gain in health effects, monetized at a cost-effectiveness threshold, plus or minus any differences in costs between the current and perfect care strategies. Headroom provides an indication of the maximum price at which a perfect innovation could be considered cost-effective.
Cohort state-transition model	A specific type of decision model, suited for simulating periodic outcomes in a hypothetical patient cohort. A cohort state-transition model describes a hypothetical cohort of patients across health states (e.g., healthy, ill) that are associated with a certain health effect and certain costs. At every periodic cycle, patients in the cohort may remain in their original health state or transition to another. The total health effects and costs of the cohort are recorded at each cycle, until the time horizon of the model has been met. The results of different care strategies may then be compared to determine which strategy yields the most favourable average cost-benefit balance. (2)
Probabilistic sensitivity analysis (PSA)	One source of uncertainty in decision modelling stems from the uncertainty regarding the input parameters of a model. The impact of such a parameter uncertainty on the model results can be investigated using probabilistic sensitivity analysis

	(PSA) in which the decision model is re-run with randomly sampled values from distributions of the parameters whose exact value is uncertain. Taken together, all runs of a PSA (e.g., 10,000) reveal the uncertainty in the model results as a consequence of the uncertainty in the input parameter values. PSA can be used to provide simulated 95% confidence intervals of the results (3)
Quality-adjusted life-years	Quality-adjusted life-years (QALYs) are the product of the number of life-years spent in a certain health state and the utility of one's health (or quality of life) during that time. As a result, 10 QALYs may be interpreted as both 10 life-years spent in perfect health and 20 life-years spent with a utility of 0.5 (4).
Utility	Utilities express the societal preference for being in a certain health state. Utilities are typically expressed on a scale from 0 (death) to 1 (perfect health).

Appendix B

Long-term part of the model. The ellipses represent health states. The arrows represent transitions. "MR-antagonist" is short for mineralocorticoid receptor antagonist (spironolactone or eplerenone).



Appendix C

Input parameters for the model.

Parameter	Value	Standard error	Distri-bution	Source
Transition probabilities				
Diagnostic part of the model				
Prevalence of PA in EH patients	0.152	0.01	Beta	(1)
Sensitivity of the current aldosterone-to-renin ratio test	0.890	0.025	Beta	(2)
Specificity of the current aldosterone-to-renin ratio test	0.960	0.006	Beta	(2)
Probability that confirmation test following a positive diagnosis is accurate	1			Assump-tion
Probability that PA patients have an aldosterone-producing adenoma fit for adrenalectomy	0.470	0.01	Beta	(3, 4)
Probability of normotension without medication after adrenalectomy	0.370	0.01	Beta dirichlet	(5)
Probability of normotension with antihypertensive medication after adrenalectomy	0.470	0.01	Beta dirichlet	(5)
Probability of hypertension with antihypertensive medication after adrenalectomy	0.160	0.01	Beta dirichlet	(5)
Probability of normotension with medication for patients with bilateral hyperplasia	0.500	0.01	Beta	(6)
Probability that patient with bilateral hyperplasia uses both MR-antagonist and antihypertensive medication	0.25			Assump-tion
Probability of normotension with antihypertensive medication for non-PA patients	0.500	0.01	Beta	(6)
Long-term part of the model				
Probability of developing a stroke (normotensive patients)	Age-depend ent		Fixed	(7)

Parameter	Value	Standard error	Distri-bution	Source
Probability of developing a stroke (hypertensive patients)	Age-dependent		Fixed	(8)
Probability of developing a stroke (PA patients)	Age-dependent		Fixed	(9)
Probability of developing coronary artery disease (normotensive patients)	Age-dependent		Fixed	(7)
Probability of developing coronary artery disease (hypertensive patients)	Age-dependent		Fixed	(8)
Probability of developing coronary artery disease (PA patients)	Age-dependent		Fixed	(9)
Probability of developing atrial fibrillation (normotensive patients)	Age-dependent		Fixed	(10)
Probability of developing atrial fibrillation (hypertensive patients)	Age-dependent		Fixed	(11)
Probability of developing atrial fibrillation (PA patients)	Age-dependent		Fixed	(9)
Probability of developing heart failure (normotensive patients)	Age-dependent		Fixed	(7)
Probability of developing heart failure (hypertensive patients)	Age-dependent		Fixed	(8)
Probability of developing heart failure (PA patients)	Age-dependent		Fixed	(9)
Probability of dying from stroke (year of incident)	0.365		Fixed	(12)
Probability of dying from stroke (years after)	0.100		Fixed	(12)
Probability of dying from coronary artery disease (year of incident)	0.029		Fixed	(13)

Parameter	Value	Standard error	Distri-bution	Source
Probability of dying from coronary artery disease (years after)	0.026		Fixed	(13)
Probability of dying from atrial fibrillation (year of incident)	0.311		Fixed	(14)
Probability of dying from atrial fibrillation (years after)	0.032		Fixed	(14)
Probability of dying from heart failure (year of incident)	0.373		Fixed	(15)
Probability of dying from heart failure (years after)	0.103		Fixed	(15)
Probability of dying from other causes	Age-dependent		Fixed	(16)
Health effects in the model				
Utility for being normotensive	0.88	0.02	Beta	(17)
Utility for being normotensive with the use of medication	0.86	0.02	Beta	Assump-tion
Utility for being hypertensive with the use of medication	0.83	0.02	Beta	(18)
Utility after having a stroke	0.68	0.02	Beta	(19)
Utility after having coronary artery disease	0.74	0.02	Beta	(20)
Utility after having atrial fibrillation	0.80	0.02	Beta	(21)
Utility after having heart failure	0.60	0.02	Beta	(22)
Utility for being dead (reference value)	0		Fixed	(23)
Discount rate for effects in the model	1.5%		Fixed	(23)
Costs				
Diagnostic part of the model				
Costs of the aldosterone-to-renin ratio test	€47		Fixed	(24)
Costs of a saline infusion test to confirm positive ARR test	€177		Fixed	(25)
Costs of adrenal vein sampling to subtype PA	€1,888		Fixed	(25)

Parameter	Value	Standard error	Distri-bution	Source
Costs of CT scan to subtype PA	€131		Fixed	(26)
Costs of adrenalectomy	€4,850		Fixed	(27)
Long-term part of the model				
Costs of stroke, year of incident	€17,371		Fixed	(28)*
Costs of stroke, years after	€7,427		Fixed	(28)*
Costs of coronary artery disease, year of incident	€16,623		Fixed	(29)
Costs of coronary artery disease, years after	€1,010		Fixed	(29)
Costs of atrial fibrillation, year of incident	€1,647		Fixed	(30, 31)
Costs of atrial fibrillation, years after	€1,558		Fixed	(30, 31)
Costs of heart failure, year of incident	€3,719		Fixed	(29)
Costs of heart failure, years after	€1,555		Fixed	(29)
Costs of cardiovascular mortality	€4,868		Fixed	(32)
Yearly costs of 150mg spironolactone, daily usage	€548		Fixed	(33, 34)
Yearly costs of antihypertensive medication, daily usage	€418		Fixed	(33, 34)
Willingness to pay for a QALY	€20,000		Fixed	(23)
Discount rate for costs in the model	4%		Fixed	(23)

* Assuming 2/3 of strokes is minor

Appendix C: References

1. Jansen PM, van den Born B-JH, Frenkel WJ, de Bruijne ELE, Deinum J, Kerstens MN, et al. Test characteristics of the aldosterone-to-renin ratio as a screening test for primary aldosteronism. Journal of Hypertension. 2014;32(1):115-26.

2. Li X, Goswami R, Yang S, Li Q. Aldosterone/direct renin concentration ratio as a screening test for primary aldosteronism: A meta-analysis. Journal of the Renin-Angiotensin-Aldosterone System. 2016;17(3).

3. Reimel B, Zanocco K, Russo MJ, Zarnegar R, Clark OH, Allendorf JD, et al. The management of aldosterone-producing adrenal adenomas—does adrenalectomy increase costs? Surgery. 2010;148(6):1178-85.

4. Graham U, Ellis P, Hunter S, Leslie H, Mullan K, Atkinson A. 100 cases of primary aldosteronism: careful choice of patients for surgery using adrenal venous sampling and CT imaging results in excellent blood pressure and potassium outcomes. Clinical endocrinology. 2012;76(1):26-32.

5. Williams TA, Lenders JWM, Mulatero P, Burrello J, Rottenkolber M, Adolf C, et al. Outcomes after adrenalectomy for unilateral primary aldosteronism: an international consensus on outcome measures and analysis of remission rates in an international cohort. The Lancet Diabetes & Endocrinology. 2017;5(9):689-99.

6. Azizi M, Pereira H, Hamdidouche I, Gosse P, Monge M, Bobrie G, et al. Adherence to Antihypertensive Treatment and the Blood Pressure–Lowering Effects of Renal Denervation in the Renal Denervation for Hypertension (DENERHTN) Trial. Circulation. 2016;134(12):847-57.

7. IJzerman MJ, Koffijberg H, Fenwick E, Krahn M. Emerging Use of Early Health Technology Assessment in Medical Product Development: A Scoping Review of the Literature. Pharmacoeconomics. 2017;35(7):727-40.

8. Hypertension and cardiovascular disease [Internet]. Bilthoven: National Institute for Public Health and the Environment; 2017 [Available from: https://www.volksgezondheidenzorg.info/onderwerp/bloeddruk/cijfers-context/oorzaken-en-gevolgen#node-verhoogde-bloeddruk-en-hart-en-vaatziekten.

9. Monticone S, D'Ascenzo F, Moretti C, Williams TA, Veglio F, Gaita F, et al. Cardiovascular events and target organ damage in primary aldosteronism compared with essential hypertension: a systematic review and meta-analysis. The Lancet Diabetes & Endocrinology. 2018;6(1):41-50.

10. [Dutch College of General Practitioners' guideline on atrial fibrillation]. Utrecht: Dutch College of General Practitioners, Working Group Atrial Fibrillation; 2017.

11. Benjamin EJ, Levy D, Vaziri SM, D'Agostino RB, Belanger AJ, Wolf PA. Independent risk factors for atrial fibrillation in a population-based cohort: The framingham heart study. JAMA. 1994;271(11):840-4.

12. Hankey GJ, Jamrozik K, Broadhurst RJ, Forbes S, Burvill PW, Anderson CS, et al. Five-Year Survival After First-Ever Stroke and Related Prognostic Factors in the Perth Community Stroke Study. Stroke. 2000;31(9):2080-6.

13. Höfer S, Benzer W, Oldridge N. Change in health-related quality of life in patients with coronary artery disease predicts 4-year mortality. International Journal of Cardiology. 2014;174(1):7-12.

14. Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB, Levy D. Impact of Atrial Fibrillation on the Risk of Death. The Framingham Heart Study. 1998;98(10):946-52.

15. Goldberg RJ, Ciampa J, Lessard D, Meyer TE, Spencer FA. Long-term survival after heart failure: A contemporary population-based perspective. Archives of Internal Medicine. 2007;167(5):490-6.

16. Bodenheimer T. High and rising health care costs. Part 2: technologic innovation. Annals of internal medicine. 2005;142(11):932-7.

17. Hoeymans N, Van Lindert H, Westert G. The health status of the Dutch population as assessed by the EQ-6D. Quality of Life Research. 2005;14(3):655-63.

18. Lawrence WF, Fryback DG, Martin PA, Klein R, Klein BEK. Health status and hypertension: A population-based study. Journal of Clinical Epidemiology. 1996;49(11):1239-45.

19. Luengo-Fernandez R, Leal J, Gray A, Sullivan R. Economic burden of cancer across the European Union: a population-based cost analysis. The Lancet Oncology. 2013;14(12):1165-74.

20. Shah B, Deshpande S. Assessment of Effect of Diabetes on Health-Related Quality of Life in Patients with Coronary Artery Disease Using the EQ-5D Questionnaire. Value in Health Regional Issues. 2014;3:67-72.

21. Kleintjens J, Li X, Simoens S, Thijs V, Goethals M, Rietzschel ER, et al. Cost-effectiveness of rivaroxaban versus warfarin for stroke prevention in atrial fibrillation in the Belgian healthcare setting. Pharmacoeconomics. 2013;31(10):909-18.

22. Calvert MJ, Freemantle N, Cleland JG. The impact of chronic heart failure on health-related quality of life data acquired in the baseline phase of the CARE-HF study. European journal of heart failure. 2005;7(2):243-51.

23. Tan SS, Bouwmans CA, Rutten FF, Hakkaart-van Roijen L. Update of the Dutch manual for costing in economic evaluations. International journal of technology assessment in health care. 2012;28(2):152-8.

24. Costs of laboratory analyses [Internet]. Utrecht: Netherlands Society for Clinical Chemistry and Laboratory Medicine; 2017 [Available from: https://www.nvkc.nl/professional/wie-doet-wat-database.

25. Velasco A, Chung O, Raza F, Pandey A, Brinker S, Arbique D, et al. Cost-Effectiveness of Therapeutic Drug Monitoring in Diagnosing

Primary Aldosteronism in Patients With Resistant Hypertension. The Journal of Clinical Hypertension. 2015;17(9):713-9.

26. Tordrup D, Chouaid C, Cuijpers P, Dab W, van Dongen JM, Espin J, et al. Priorities for health economic methodological research: results of an expert consultation. International journal of technology assessment in health care. 2017;33(6):609-19.

27. Cutler DM, McClellan M. Is technological change in medicine worth it? Health affairs. 2001;20(5):11-29.

28. Greving JP, Buskens E, Koffijberg H, Algra A. Cost-effectiveness of aspirin treatment in the primary prevention of cardiovascular disease events in subgroups based on age, gender, and varying cardiovascular risk. Circulation. 2008;117(22):2875-83.

29. Drost J, Grutters J, van der Wilt G-J, van der Schouw Y, Maas A. Yearly hypertension screening in women with a history of pre-eclampsia: a cost-effectiveness analysis. Netherlands Heart Journal. 2015;23(12):585-91.

30. Stevanović J, Pompen M, Le HH, Rozenbaum MH, Tieleman RG, Postma MJ. Economic evaluation of apixaban for the prevention of stroke in non-valvular atrial fibrillation in the Netherlands. PloS one. 2014;9(8):e103974.

31. Holstenson E, Ringborg A, Lindgren P, Coste F, Diamand F, Nieuwlaat R, et al. Predictors of costs related to cardiovascular disease among patients with atrial fibrillation in five European countries. Europace. 2010;13(1):23-30.

32. Van Exel J, Koopmanschap MA, Van Wijngaarden JD, op Reimer WJS. Costs of stroke and stroke services: Determinants of patient costs and a comparison of costs of regular care and care organised in stroke services. Cost Effectiveness and Resource Allocation. 2003;1(1):2.

33. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes: Oxford university press; 2015.

34. Oortwijn W, Sampietro-Colom L, Habens F. DEVELOPMENTS IN VALUE FRAMEWORKS TO INFORM THE ALLOCATION OF HEALTHCARE RESOURCES. Int J Technol Assess Health Care. 2017;33(2):323-9.

"Defining yourself as against something

says very little about what you are for"

Greg Graffin, professor of Palaeontology and lead singer of
punk rock band Bad Religion

Chapter 8

Data sources and methods used to determine pretest probabilities in a cohort of Cochrane Diagnostic Test Accuracy reviews

M.S. Oerbekke

K. Jenniskens

W.A. van Enst

R.J.P.M. Scholten

L. Hooft

Submitted

Abstract

Introduction: To facilitate the interpretation of diagnostic test accuracy (DTA) parameters it is possible to calculate normalized frequencies. They provide the number of (false) positives and (false) negatives in a tested hypothetical cohort. A pretest probability must be determined to calculate these normalized frequencies. The aim of this study was to assess the data sources and methods used in Cochrane DTA reviews for determining pretest probabilities to facilitate the interpretation of DTA parameters.

Methods: Cochrane DTA reviews published in the Cochrane Database of Systematic Reviews up to and including January 2018 and presenting at least one meta-analytic estimate of the sensitivity and/or specificity as a primary analysis were included in the cohort. Study selection and data extraction were performed by one author and checked by other authors. Observed data sources and methods were categorized.

Results: Fifty-nine DTA reviews were included, comprising of 308 meta-analyses. A pretest probability was used in 148 meta-analyses. Authors used included studies in the DTA review, external sources, and expert opinion as data sources for the pretest probability. When using the included studies in the DTA review, authors used a measure of central tendency whether or not combined with a measure of dispersion to determine the pretest probabilities. Identical pretest probabilities were used for analyses of two or more index tests for the same target conditions. About half (53.6%) of these identical pretest probabilities fell within the prevalence ranges from all analyses within a target condition.

Conclusions: Various methods are used for selecting pretest probabilities and no consensus seems to exists on which data source or method to use. However, there are some considerations to take in to account when presenting DTA results: 1) Consider whether to present normalized frequencies, 2) Consider the influence of the chosen method for selecting a pretest probability on the normalized frequencies, and 3) Consider whether to use identical pretest probabilities that fall within the range of the selected studies when there are multiple meta-analyses for a target condition.

Introduction

Diagnostic tests are essential to clinicians in their daily practice. Test results inform about the preferred healthcare pathway to, ideally, cure a patient from disease. The optimal way to understand a diagnostic test's performance and the downstream consequences for patients is through a test-treatment randomized controlled trial. Such trials provide comparative information on health outcomes (both harms and benefits) of healthcare pathways initiated by the outcome of the diagnostic tests or strategies. However, test-treatment randomized controlled trials are methodologically complex (1). Primary cross-sectional studies are usually an alternative to these complex trials and can be summarized in systematic reviews. Diagnostic test accuracy (DTA) reviews include primary cross-sectional studies using the diagnostic test of interest and aggregate data by meta-analysis so that a pooled sensitivity and specificity is presented. The pooled sensitivity and specificity could help to classify persons correctly as having the target disease or not having the target disease (i.e. true positives and true negatives). Misclassified persons (i.e. false positives and false negatives) in DTA reviews are equally important to report due to the downstream consequences of misclassifications. A false negative classified person may not receive appropriate treatment while a falsely positive classified person may receive abundant treatment.

From literature it seems that clinicians have trouble interpreting accuracy parameters such as sensitivity and specificity (2). To facilitate the interpretation of DTA results absolute numbers of true/false positives and true/false negatives can be presented in a hypothetical cohort of e.g. 1000 persons, which is also known as normalized frequencies (2). However, to calculate normalized frequencies a pretest probability (i.e. the disease prevalence in the hypothetical cohort) needs to be determined. The normalized frequencies are then calculated and reported, after which the diagnostic test's end-user can interpret whether the test performance is acceptable in terms of true or false positives and negatives. Such normalized frequencies are usually presented in summary of findings tables in Cochrane DTA Reviews and in the evidence tables from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) (3).

These normalized frequencies are not only important for clinicians, but also for guideline boards and policy makers. Decisions whether or not to recommend the use of a diagnostic test in a guideline or decisions about health care restitution may be influenced by the presented normalized frequencies, which, in turn, are dependent on the determined pretest probability. While GRADE does not suggest a specific method to determine a pretest probability in its handbook (3, 4) the Cochrane Handbook does propose some methods (e.g. the median disease prevalence or the prevalence from disease registries) although a rationale to use a specific method is not given (5).

Because guidance in determining a pretest probability is minimally described it is unknown what data sources (i.e. the data on which a pretest probability is based) and methods are actually used to determine a pretest probability in DTA reviews. Furthermore, an identical pretest probability is ideally used in analyses of more than one index test for the same target condition in order to enable comparison of the normalized frequencies of the various index tests. Therefore, the aim of this study was to assess the data sources and methods used in a cohort of Cochrane DTA reviews to determine pretest probabilities to facilitate the interpretation of pooled DTA accuracy parameters. A secondary aim was to assess the use of identical pretest probabilities in multiple analyses within a target condition, necessary for the comparison of normalized frequencies.

Methods

Cohort definition

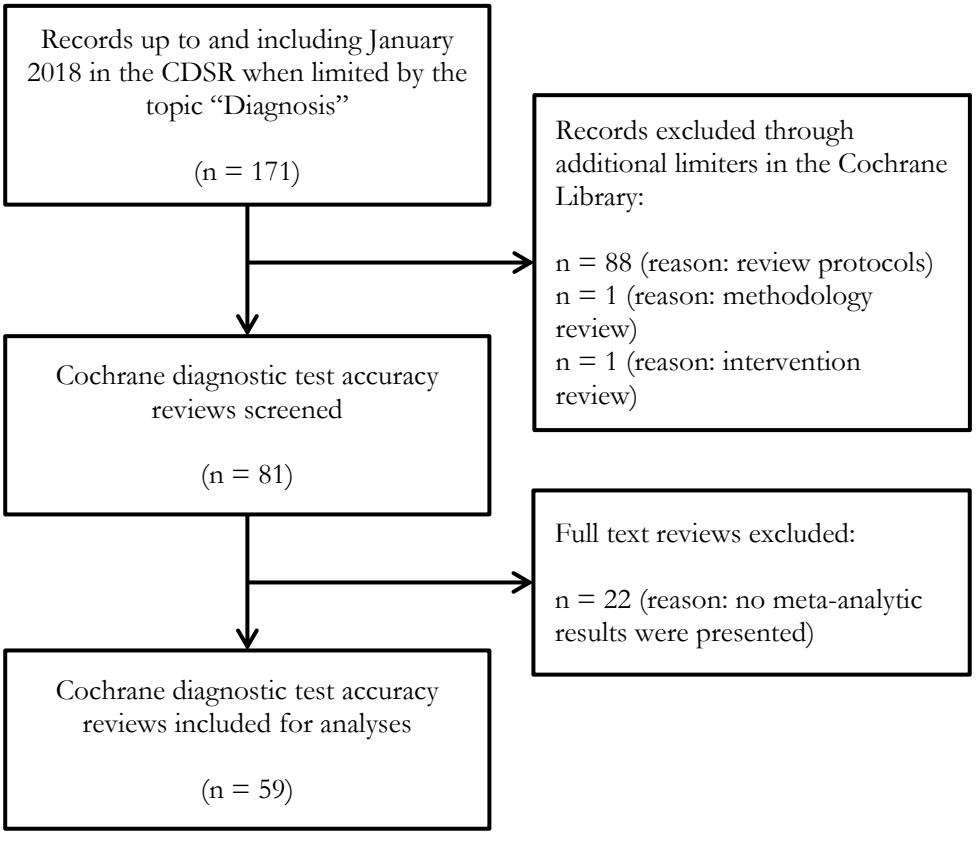
The Cochrane Database of Systematic Reviews (CDSR) was accessed through the Cochrane Library. Cochrane DTA reviews published in the period from inception of the CDSR up to and including January 2018 were potentially eligible to enter the cohort. To obtain DTA reviews the CDSR was browsed by the topic 'Diagnosis', while protocols and intervention or methodology reviews were excluded through limiters in the search engine interface. A DTA review was included in the cohort when it reported at least one meta-analytic estimate of sensitivity and/or specificity (i.e. either retrieved with a bivariate

model or by using a hierarchical summary receiver operating characteristic model) as a primary analysis in the presented tables. The screening and selection of eligible DTA reviews was performed by one author (MSO) and checked by the other authors (KJ, RJPMS, WAvE, LH).

Categorization of sources and methods for determining pretest probabilities

The first step in the categorization of extracted data was to divide the included meta-analyses into two groups; One group of analyses where no pretest probability was used and one group of analyses where a pretest probability was used. Next, for every analysis that used a pretest probability the source of the pretest probability was determined. For example, the pretest probability could be determined based on the studies that were included for one of the target conditions presented in the DTA review. Furthermore, the method for determining the pretest probability itself was recorded. For example, this could be the usage of the mean or median disease prevalence (from studies included for a target condition). General characteristics (e.g. title, publication year), the number of meta-analyses in the review, whether a pretest probability was used, the number of pretest probabilities used (if applicable), the source of data for determining the pretest probability, and the method used for selecting pretest probabilities (if applicable) were extracted by one author (MSO) and checked by the other authors (KJ, RJPMS, WAvE, LH). When a disease prevalence was reported but not used to interpret the sensitivity and/or specificity in some manner, the disease prevalence was not considered as a pretest probability. Descriptive statistics were performed in IBM SPSS Statistics for Windows (Version 21, 2012, Armonk, NY: IBM Corp.).

Figure 1. Flow diagram showing the formation of the cohort and reasons for exclusion. CDSR: Cochrane Database of Systematic Reviews



Pretest probabilities in prevalence ranges of analyses within a target condition

DTA reviews that had analyses for more than one index test for the same target condition were identified from the cohort. The number of analyses for the same target condition, the number of identical pretest probabilities used in those analyses, and whether the pretest probability fell within or outside the range of disease prevalence found in the individual analyses within a target condition were extracted by one author (MSO) and checked by a second

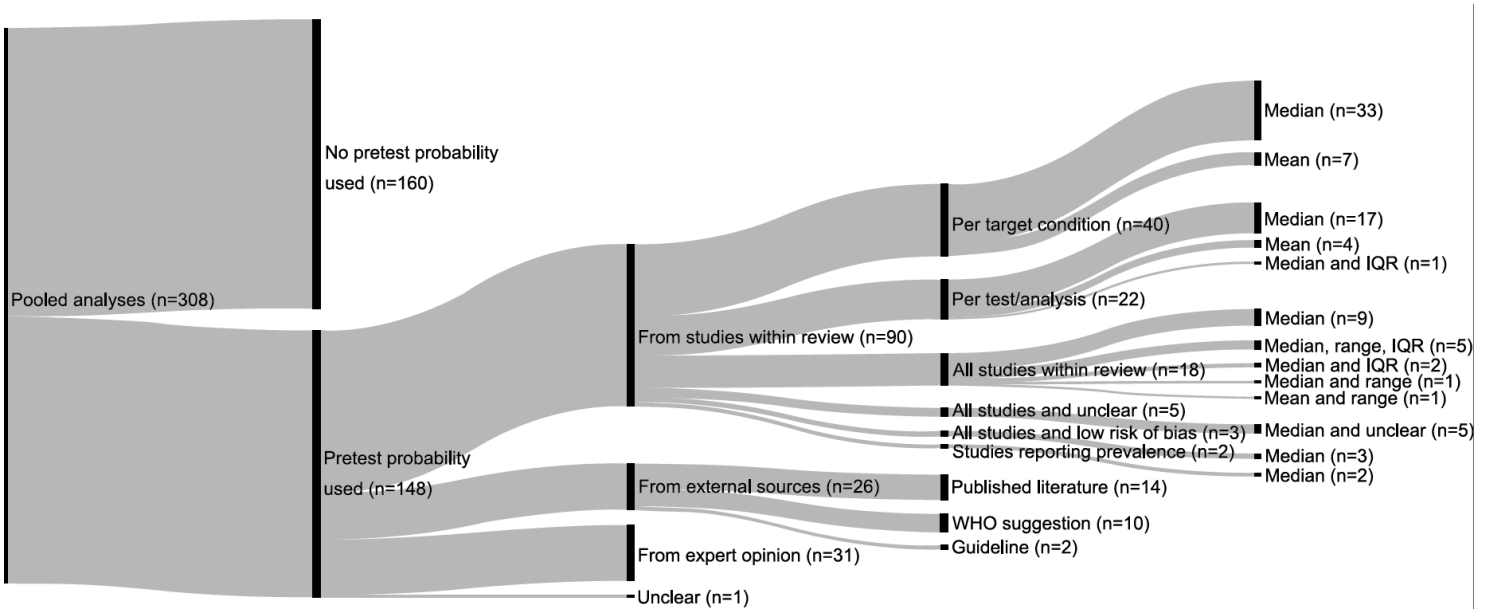
author (KJ). It was recorded whether an identical pretest probability fell inside all prevalence ranges from the analyses within a target condition or whether it fell outside of at least one prevalence range from an analysis within a target condition. When the range of disease prevalence per analysis was not described in the DTA review, it was calculated from the data in the review's appendices.

Results

Cohort description

The CDSR contained 171 documents on the topic 'Diagnosis'. There were 81 reviews left after excluding 88 review protocols and 2 reviews on interventions and methodology. After screening the full text an additional 22 DTA reviews were excluded as no meta-analytic results were presented (referenced in Appendix A). Consequently, 59 Cochrane DTA reviews were included in the cohort (Figure 1 and Appendix A). The 59 DTA reviews in the cohort contained 308 meta-analyses (see Table 1). The number of meta-analyses ranged from 1 to 34 (median: 3) per review. In 16 reviews (27.1%) there were 150 meta-analyses (48.7%) that did not use a pretest probability. Thirty-nine reviews (66.1%) had 143 meta-analyses (46.4%) where a pretest probability was used. Four reviews (6.8%) contained 15 meta-analyses for which a pretest probability was used in 5 analyses. Therefore, a total of 160 analyses were found where no pretest

Figure 2. Sankey plot showing which data sources and methods were used to determine the pretest probability in Cochrane Diagnostic Test Accuracy reviews. IQR: Interquartile Range, WHO: World Health Organization



probability was used and 148 analyses were found where at least one pretest probability was used.

Sources of pretest probabilities

In the 148 analyses in which a pretest probability was used three main categories of data sources were distinguished (Figure 2). The pretest probability was determined from the included studies (90 analyses), from external sources (26 analyses) or based on expert opinion (31 analyses). In one analysis the source was unclear. When the included studies in the review were used to determine one or multiple pretest probabilities the data source could be further differentiated into: all studies included for the target condition (40 analyses), studies used per test/analysis for a target condition (22 analyses), or all studies in the systematic review across all target conditions (18 analyses).

Furthermore, ten other analyses had multiple pretest probabilities determined from all studies in the systematic review and from an unclear method (5 analyses), from all studies in the systematic review and only from studies with a low risk of bias (3 analyses), or only from included studies that reported the disease prevalence (2 analyses). When the pretest probability was determined based on external sources the pretest probability came from a reported disease prevalence in published scientific literature (14 analyses), from a WHO suggestion (10 analyses), or from a guideline (2 analyses). It was considered an expert opinion when authors assumed a pretest probability (31 analyses). See Appendix B for a short description of each category and for examples within each category.

Methods for determining a pretest probability

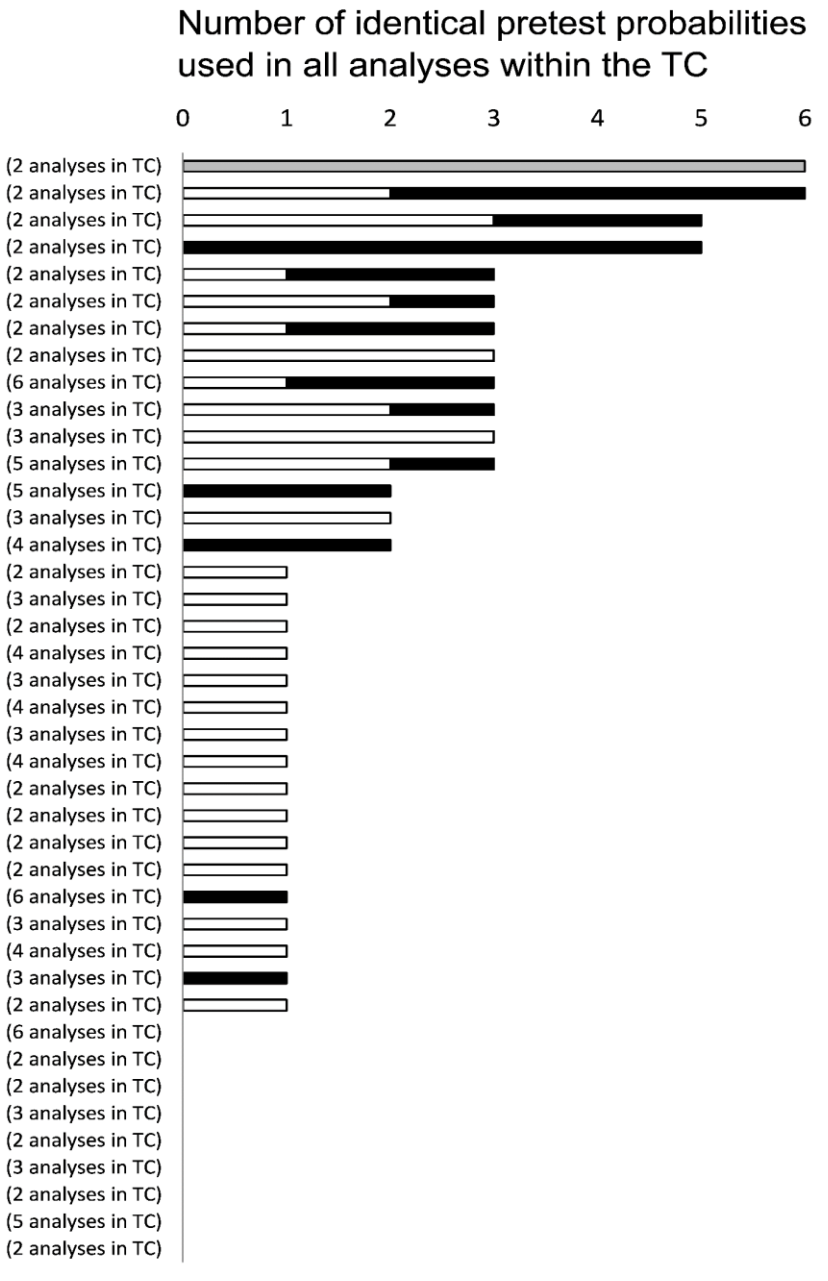
Pretest probabilities based on studies within the review were determined by using measures of central tendency (e.g. median) whether or not combined with measures of dispersion (e.g. range). Using multiple methods resulted in multiple pretest probabilities for a single analysis (e.g. using a median with a range results in three pretest probabilities). The median was used individually (64 analyses) or together with the interquartile range (3 analyses), with the range (1 analysis), with the range and interquartile range

Table 1. General characteristics of Cochrane DTA reviews included in the analysis. DTA: Diagnostic Test Accuracy, IQR: Interquartile Range. [a] Calculated from the analyses that used pretest probabilities. [b] Could not be calculated since there were no analyses using a pretest risk

	Reviews (n)	Meta-analyses[a] n (% using a pretest risk)	Number of meta-analyses per DTA review Median (range)	Number of pretest probabilities used per meta-analysis[a] Median (range)
Total included Cochrane DTA reviews	59	308 (48.1)	3 (1-34)	1 (1-6)
Reviews not using a pretest risk at all	16	150 (0)	4 (2-34)	[b]
Reviews using a pretest risk for all pooled analyses	39	143 (100)	3 (1-16)	1 (1-5)
Reviews reporting analyses with and without pretest risk	4	15 (33.3)	3.5 (3-5)	3 (1-6)

(5 analyses), or with an unclear method (5 analyses). The mean was used individually (7 analyses) or together with the range (1 analysis). Figure 2 shows the methods per data source and the number of analyses in where these methods were used.

Figure 3. Identical pretest probabilities in a target condition and whether they fell inside disease prevalence ranges. TC = Target condition



Pretest probabilities in target conditions with more than one analysis

In 29 reviews there were 41 target conditions with two or more analyses (Figure 3). In 32 target conditions identical pretest probabilities were used in all analyses (range: 1-6 pretest probabilities). In nine target conditions different pretest probabilities were used. From the 69 identical pretest probabilities that were used in the 41 target conditions, 37 pretest probabilities (53.6%) fell inside the disease prevalence ranges of the included studies. However, 26 identical pretest probabilities (37.7%) fell outside at least one prevalence range of an analysis within a target condition, while this remained unclear for six pretest probabilities (8.7%). One target condition had six identical pretest probabilities in both of its analyses, however it was unclear which pretest probability fell inside or outside the disease prevalence ranges.

Discussion

A total of 59 Cochrane DTA reviews were included to assess the data sources and methods used to determine pretest probabilities and to assess the use of identical pretest probabilities in multiple analyses within a target condition. Various sources and methods to determine a pretest probability were found. Sixteen DTA reviews did not use a single pretest probability. Almost half of the observed meta-analyses used at least one pretest probability (range: 1-6 pretest probabilities) to facilitate the interpretability of the results. The median was the most used method to determine a pretest probability. Thirty-nine target conditions contained two or more analyses and used at least one identical pretest probability for all of its analyses (range: 1-6 pretest probabilities). Twenty-six of the identical pretest probabilities (37.7%) fell outside the disease prevalence range of at least one analysis within the target condition.

The Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy proposes to use the median prevalence of the included studies or external sources as methods to select a pretest probability (5). Indeed, the median was used more than any other method in Cochrane DTA reviews when the pretest

probability was determined based on studied included in the review. External sources were also used in 26 analyses. The Cochrane Handbook also suggests the use of disease registries for selecting a representative pretest probability, but no such method was specifically mentioned in the included Cochrane DTA reviews. Observed prevalence from disease surveillance systems and epidemiological surveys by the WHO, however, might be interpreted as disease registries as well. The Cochrane Handbook states that a representative pretest probability may be derived from the included studies only when the studies are representative for the target setting (5), in which case the selected pretest probability will fall within the prevalence range of included studies. However, when determining a pretest probability from external sources or expert opinion a pretest probability may be selected that falls outside the disease prevalence range from the included studies. A pretest probability from an external source could be representative for the target setting, but it might not necessarily be an appropriate pretest probability in the context of the disease prevalence range from the included studies in a meta-analysis. When a representative pretest probability falls outside the disease prevalence range, a more appropriate pretest probability from the range of disease prevalence from the meta-analysis itself might be selected in the context of the data in the meta-analysis. Therefore, a representative and an appropriate pretest probability are not necessarily the same. Appropriateness in this case means to not extrapolate outside of the data in the meta-analysis.

A potential limitation of this study is that only Cochrane DTA reviews were included. Since the Cochrane Handbook proposes to use the median it was beforehand likely to observe that the median is being preferred in Cochrane DTA reviews. Different methods and data sources for determining pretest probabilities in non-Cochrane DTA reviews could have been missed. However, when there were data sources or methods that this study did not address, it adds to the impression that there is no consensus on what data source and method to use for determining a representative or appropriate pretest probability.

From the results of this study no clear guidance can be given on what source or what method should be used for determining a pretest probability. Furthermore, it is unknown if a pretest probability outside the disease

prevalence ranges is problematic in clinical reality. Whether or not it is problematic may also be context dependent, as for some target conditions a certain number of misclassifications are more acceptable than for target conditions where misclassifications have severe downstream consequences. Even if it turns out to be clinically problematic in future research, it presently might still be best practice to facilitate the interpretability of diagnostic accuracy parameters by presenting normalized frequencies. Furthermore, there are some considerations which may be taken in to account when presenting results in DTA reviews.

First to consider is whether to provide a way for end-users to interpret the presented accuracy parameters, as it was observed in this study that about half of all meta-analyses were not accompanied with normalized frequencies from a hypothetical cohort. Literature shows that interpreting diagnostic test accuracy parameters may be troublesome for its users and therefore normalized frequencies may be useful (2). However, choosing pretest probabilities to calculate normalized frequencies is not without difficulties and therefore it is uncertain whether normalized frequencies are trustworthy enough for all decision-making (see the second consideration). The need for interpretability versus the certainty of and need for a truthful representation might determine whether normalized frequencies are calculated.

Secondly, consider giving thought about the influence of the method of selecting the pretest probability on the normalized frequencies from the hypothetical cohort. A guideline board may base their decision about whether or not to recommend a test for clinical practice on the presented normalized frequencies. It is important to understand that different pretest probabilities will result in different normalized frequencies while the sensitivity and specificity remain constant (see Appendix C), potentially influencing the decision-making in practice, policy or guidelines.

Thirdly, when there are multiple meta-analyses for the same target condition, consider whether to use an identical pretest probability in each of those analyses so that the normalized frequencies can be compared. Ideally the selected pretest probabilities fall inside all of the disease prevalence ranges from all individual meta-analyses within the target condition, although this

might not be feasible for every scenario (e.g. when the disease prevalence ranges from the meta-analyses do not overlap). However, from this study no guidance can be provided on whether an identical pretest probability is suitable for all of the disease prevalence ranges in the analyses, even when the pretest probability falls inside all prevalence ranges.

Providing clinicians, policy makers, and guideline boards with methods to facilitate the interpretation of DTA results is not only important for them, but ultimately also for patients who undergo diagnostic tests. Different pretest probabilities will result in different normalized frequencies. However, it is not known whether differences in normalized frequencies caused by the use of different pretest probabilities actually impacts decision-making and whether it will then clinically harm of benefit patients. The future direction of research in this area could focus on whether different pretest probabilities will actually result in a different clinical decision, guideline recommendation, or policy change. Furthermore, future research could focus on developing other strategies for accuracy parameters so that they are both interpretable and helpful when research shows that calculating normalized frequencies may not be beneficial for actual decision-making by clinicians, policy makers, and/or guideline boards.

Various data sources and methods are used to obtain a pretest probability without consensus on which data source or method to use. However, there are three considerations that might be taken in to account when presenting DTA results: 1) Consider whether or not to present normalized frequencies from a hypothetical cohort, 2) Consider the influence of the chosen method for selecting a pretest probability on the normalized frequencies from a hypothetical cohort on which a clinical decision, guideline recommendation or policy change may be based, and 3) Consider to use identical pretest probabilities that fall within the range of the selected studies when there are multiple meta-analyses for a target condition.

References

1. Ferrante di Ruffano L, Dinnes J, Sitch AJ, Hyde C, Deeks JJ. Test-treatment RCTs are susceptible to bias: a review of the methodological quality of randomized trials that evaluate diagnostic tests. BMC medical research methodology 2017, 17(1):35.

2. Whiting PF, Davenport C, Jameson C, Burke M, Sterne JA, Hyde C, Ben-Shlomo Y. How well do health professionals interpret diagnostic information? A systematic review. BMJ open 2015, 5(7):e008155.

3. Mustafa RA, Wiercioch W, Santesso N, Cheung A, Prediger B, Baldeh T, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, Bossuyt P et al. Decision-Making about Healthcare Related Tests and Diagnostic Strategies: User Testing of GRADE Evidence Tables. PloS one 2015, 10(10):e0134553.

4. Schüneman H, Brożek J, Guyatt G, Oxman A. Chapter 7: The GRADE Approach for Diagnostic Tests and Strategies. In: Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach, Updated October 2013. edn. Edited by Schüneman H, Brożek J, Guyatt G, Oxman A: The GRADE Working Group, Available from: https://guidelinedevelopment.org/handbook.

5. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11: Interpreting results and drawing conclusions. In: Cochrane Handbook for Systematic Reviews of Diagostic Test Accuracy, Version 09. edn. Edited by Deeks J, Bossuyt P, Gatsonis C: The Cochrane Collaboration, Available from: http://methods.cochrane.org/sdt/handbook-dta-reviews; 2013.

Appendix A

Reviews excluded from the analysis

Author	Year	Title	Reason for exclusion
Hull et al.	2017	Tests for detecting strabismus in children aged 1 to 6 years in the community	No pooled/summary analysis performed
Davidson et al.	2017	Amylase in drain fluid for the diagnosis of pancreatic leak in post- pancreatic resection	No pooled/summary analysis performed
Mens et al.	2017	Imaging for the exclusion of pulmonary embolism in pregnancy	No pooled/summary analysis performed
Harrison et al.	2016	Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the early diagnosis of dementia across a variety of healthcare settings	No pooled/summary analysis performed
Crawford et al.	2016	Ankle brachial index for the diagnosis of lower limb peripheral arterial disease	No pooled/summary analysis performed
Crawford et al.	2016	D- dimer test for excluding the diagnosis of pulmonary embolism	No pooled/summary analysis performed
Nisenblat et al.	2016	Combination of the non-invasive tests for endometriosis diagnosis	No pooled/summary analysis performed
Liu et al.	2015	Urinary biomarkers for the non-invasive endometriosis diagnosis	No pooled/summary analysis performed
Davis et al.	2015	Montreal Cognitive Assessment for the diagnosis of Alzheimer's disease and other dementias	No pooled/summary analysis performed
Bleeker et al.	2015	123I- MIBG scintigraphy and 18F- FDG- PET imaging for diagnosing neuroblastoma	No pooled/summary analysis performed

Author	Year	Title	Reason for exclusion
Palaniyappan et al.	2015	Voxel- based morphometry for separation of schizophrenia from other types of psychosis in first episode psychosis	No pooled/summary analysis performed
Archer et al.	2015	Regional Cerebral Blood Flow Single Photon Emission Computed Tomography for detection of Frontotemporal dementia in people with suspected dementia	No pooled/summary analysis performed
Arevalo et al.	2015	Mini- Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)	No pooled/summary analysis performed
Hunt et al.	2015	Thromboelastography (TEG) and rotational thromboelastometry (ROTEM) for trauma-induced coagulopathy in adult trauma patients with bleeding	No pooled/summary analysis performed
Fage et al.	2015	Mini- Cog for the diagnosis of Alzheimer's disease dementia and other dementias within a community setting	No pooled/summary analysis performed
McCleery et al.	2015	Dopamine transporter imaging for the diagnosis of dementia with Lewy bodies	No pooled/summary analysis performed
Harrison et al.	2014	Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a general practice (primary care) setting	No pooled/summary analysis performed
Rutten et al.	2014	Laparoscopy for diagnosing resectability of disease in patients with advanced ovarian cancer	No pooled/summary analysis performed

Author	Year	Title	Reason for exclusion
Walsh et al.	2013	Clinical assessment to screen for the detection of oral cavity cancer and potentially malignant disorders in apparently healthy adults	No pooled/summary analysis performed
Hanchard et al.	2013	Physical tests for shoulder impingements and local lesions of bursa, tendon or labrum that may accompany impingement	No pooled/summary analysis performed
Henschke et al.	2013	Red flags to screen for malignancy in patients with low-back pain	No pooled/summary analysis performed
Wiliams et al.	2013	Red flags to screen for vertebral fracture in patients presenting with low- back pain	No pooled/summary analysis performed

Reviews included in the analysis

Author	Year	Title
Koliopoulos et al.	2017	Cytology versus HPV testing for cervical cancer screening in the general population
Abraha et al.	2017	Ultrasonography for endoleak detection after endoluminal abdominal aortic aneurysm repair
Wijedoru et al.	2017	Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever
Nieuwenhuis et al.	2017	Three- dimensional saline infusion sonography compared to two- dimensional saline infusion sonography for the diagnosis of focal intracavitary lesions
Colli et al.	2017	Platelet count, spleen length, and platelet count- to-spleen length ratio for the diagnosis of oesophageal varices in people with chronic liver disease or portal vein thrombosis
Rompianesi et al.	2017	Serum amylase and lipase and urinary trypsinogen and amylase for diagnosis of acute pancreatitis
Best et al.	2017	Imaging modalities for characterising focal pancreatic lesions
Ritchie et al.	2017	CSF tau and the CSF tau/ABeta ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI)
Pammi et al.	2017	Molecular assays for the diagnosis of sepsis in neonates
Tamburrino et al.	2016	Diagnostic accuracy of different imaging modalities following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer
Theron et al.	2016	GenoType® MTBDRsl assay for resistance to second- line anti- tuberculosis drugs
Allen et al.	2016	Diagnostic accuracy of laparoscopy following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer

Author	Year	Title
Shaikh et al.	2016	Dimercaptosuccinic acid scan or ultrasound in screening for vesicoureteral reflux among children with urinary tract infections
Cohen et al.	2016	Rapid antigen detection test for group A streptococcus in children with pharyngitis
Shah et al.	2016	Lateral flow urine lipoarabinomannan assay for detecting active tuberculosis in HIV- positive adults
Nisenblat et al.	2016	Blood biomarkers for the non- invasive diagnosis of endometriosis
Gupta et al.	2016	Endometrial biomarkers for the non- invasive diagnosis of endometriosis
Ratnavelu et al.	2016	Intraoperative frozen section analysis for the diagnosis of early stage ovarian cancer in suspicious pelvic masses
Nisenblat et al.	2016	Imaging modalities for the non- invasive diagnosis of endometriosis
Creavin et al.	2016	Mini- Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations
Leeflang et al.	2015	Galactomannan detection for invasive aspergillosis in immunocompromised patients
Nicholson et al.	2015	Blood CEA levels for detecting recurrent colorectal cancer
Alldred et al.	2015	Urine tests for Down's syndrome screening
Alldred et al.	2015	First trimester serum tests for Down's syndrome screening
Michelessi et al.	2015	Optic nerve head and fibre layer imaging for diagnosing glaucoma
Cruciani et al.	2015	Polymerase chain reaction blood tests for the diagnosis of invasive aspergillosis in immunocompromised people
Mallee et al.	2015	Computed tomography versus magnetic resonance imaging versus bone scintigraphy for clinically suspected scaphoid fractures in patients with negative plain radiographs

Author	Year	Title
Macey et al.	2015	Diagnostic tests for oral cancer and potentially malignant disorders in patients presenting with clinically evident lesions
Hooper et al.	2015	Clinical symptoms, signs and tests for identification of impending and current water- loss dehydration in older people
Ochodo et al.	2015	Circulating antigen tests and urine reagent strips for diagnosis of active schistosomiasis in endemic areas
Harrison et al.	2015	Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a secondary care setting
Gurusamy et al.	2015	Endoscopic retrograde cholangiopancreatography versus intraoperative cholangiography for diagnosis of common bile duct stones
Gurusamy et al.	2015	Ultrasound versus liver function tests for diagnosis of common bile duct stones
Giljaca et al.	2015	Endoscopic ultrasound versus magnetic resonance cholangiopancreatography for common bile duct stones
Mocellin et al.	2015	Diagnostic accuracy of endoscopic ultrasonography (EUS) for the preoperative locoregional staging of primary gastric cancer
Smailagic et al.	2015	18F- FDG PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI)
Soares et al.	2015	First rank symptoms for schizophrenia
Pavlov et al.	2015	Transient elastography for diagnosis of stages of hepatic fibrosis and cirrhosis in people with alcoholic liver disease
Shaikh et al.	2015	Procalcitonin, C- reactive protein, and erythrocyte sedimentation rate for the diagnosis of acute pyelonephritis in children
Virgili et al.	2015	Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy

Author	Year	Title
Abba et al.	2014	Rapid diagnostic tests for diagnosing uncomplicated non- falciparum or Plasmodium vivax malaria in endemic countries
Schmidt et al.	2014	PET- CT for assessing mediastinal lymph node involvement in patients with suspected resectable non- small cell lung cancer
Colli et al.	2014	Capsule endoscopy for the diagnosis of oesophageal varices in people with chronic liver disease or portal vein thrombosis
Josephson et al.	2014	Computed tomography angiography or magnetic resonance angiography for detection of intracranial vascular malformations in patients with intracerebral haemorrhage
Zhang et al.	2014	11C- PIB- PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI)
Lawrie et al.	2014	Sentinel node assessment for diagnosis of groin lymph node involvement in vulval cancer
Boelaert et al.	2014	Rapid tests for the diagnosis of visceral leishmaniasis in patients with suspected disease
Ritchie et al.	2014	Plasma and cerebrospinal fluid amyloid beta for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI)
Quinn et al.	2014	Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within community dwelling populations
Taylor et al.	2014	Computed tomography (CT) angiography for confirmation of the clinical diagnosis of brain death
Steingart et al.	2014	Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults
Lenza et al.	2013	Magnetic resonance imaging, magnetic resonance arthrography and ultrasonography for assessing rotator cuff tears in people with shoulder pain for whom surgery is being considered

Author	Year	Title
Arbyn et al.	2013	Human papillomavirus testing versus repeat cytology for triage of minor cytological cervical lesions
Wang et al.	2012	Clinical symptoms and signs for the diagnosis of Mycoplasma pneumoniae in children and adolescents with community- acquired pneumonia
Alldred et al.	2012	Second trimester serum test for Down Syndrome screening
Wang et al.	2011	Cardiac testing for coronary artery disease in potential kidney transplant recipients
Abba et al.	2011	Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries
Windt et al.	2010	Physical examination for lumbar radiculopathy due to disc herniation in patients with low- back pain
Brazzelli et al.	2009	Magnetic resonance imaging versus computed tomography for detection of acute vascular lesions in patients presenting with stroke symptoms

Appendix B

Examples of data source categories

From included studies

From all studies included for the target condition

There are multiple target conditions in the systematic review and a target condition has one or multiple index tests. All studies from all analyses for a single target condition were used to determine a pretest probability. See Lenza et al. (2013) for an example (1).

From studies used per test/analysis for a target condition

A target condition has multiple index tests. For each analysis of the index test the included studies for that index test were used to determine a pretest probability. See Colli et al. (2017) for an example (2).

From all studies in the systematic review across all target conditions

A review used all of the included studies for analyses across all of the target conditions to determine a pretest probability. Analyses were also placed in this category if there was only one target condition defined in the systematic review and a pretest risk was determined from all included studies for that target condition, unless specifically stated otherwise by the authors. See Leeflang et al. (2015) for an example (3).

From all studies in the systematic review and from an unclear method

An analysis used two pretest probabilities to calculate normalized frequencies. One of the pretest probabilities was determined from all of the included studies for analyses across all target conditions, while the other pretest probability had an unclear data source. See Abba et al. (2011) for an example (4).

From all studies in the systematic review and only from studies with a low risk of bias

An analysis used two pretest probabilities to calculate normalized frequencies. One pretest probability was determined by all of the included studies for analyses across all target conditions, while the other pretest probability was calculated solely from studies with a low risk of bias. See Colli at al. (2014) for an example (5).

Only from included studies that reported the disease prevalence

Only studies that reported their sample's disease prevalence were used to determine the pretest probability. See Ritchie et al. (2017) for an example (6).

From external sources

From published scientific literature

The pretest probability used in the systematic review was based on or informed by the disease prevalence as reported in published scientific literature. See Wijedoru et al. (2017) for an example (7).

From a WHO suggestion

The pretest probability used in the systematic review was based on or informed by the disease prevalence as suggested by the WHO. See Steingart et al. (2014) for an example (8).

From a guideline

The pretest probability used in the systematic review was based on or informed by the disease prevalence as reported in a guideline. See Wang et al. (2011) for an example (9).

Expert opinion

It was considered expert opinion when authors determined a pretest probability based on an assumption. See Shaikj et al. (2016) for an example (10).

Appendix B: References

1. Lenza M, Buchbinder R, Takwoingi Y, Johnston RV, Hanchard NC, Faloppa F. Magnetic resonance imaging, magnetic resonance arthrography and ultrasonography for assessing rotator cuff tears in people with shoulder pain for whom surgery is being considered. *The Cochrane database of systematic reviews* 2013(9):Cd009020. doi: 10.1002/14651858.CD009020.pub2 [published Online First: 2013/09/26]

2. Colli A, Gana JC, Yap J, Adams-Webber T, Rashkovan N, Ling SC, Casazza G. Platelet count, spleen length, and platelet count-to-spleen length ratio for the diagnosis of oesophageal varices in people with chronic liver disease or portal vein thrombosis. *The Cochrane database of systematic reviews* 2017;4:Cd008759. doi: 10.1002/14651858.CD008759.pub2 [published Online First: 2017/04/27]

3. Leeflang MM, Debets-Ossenkopp YJ, Wang J, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, Reitsma JB, Zhang M, Bossuyt PM, Vandenbroucke-Grauls CM. Galactomannan detection for invasive aspergillosis in immunocompromised patients. *The Cochrane database of systematic reviews* 2015(12):Cd007394. doi: 10.1002/14651858.CD007394.pub2 [published Online First: 2015/12/31]

4. Abba K, Deeks JJ, Olliaro P, Naing CM, Jackson SM, Takwoingi Y, Donegan S, Garner P. Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries. *The Cochrane database of systematic reviews* 2011(7):Cd008122. doi: 10.1002/14651858.CD008122.pub2 [published Online First: 2011/07/08]

5. Colli A, Gana JC, Turner D, Yap J, Adams-Webber T, Ling SC, Casazza G. Capsule endoscopy for the diagnosis of oesophageal varices in people with chronic liver disease or portal vein thrombosis. *The Cochrane database of systematic reviews* 2014(10):Cd008760. doi: 10.1002/14651858.CD008760.pub2 [published Online First: 2014/10/02]

6. Ritchie C, Smailagic N, Noel-Storr AH, Ukoumunne O, Ladds EC, Martin S. CSF tau and the CSF tau/ABeta ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild

cognitive impairment (MCI). *The Cochrane database of systematic reviews* 2017;3:Cd010803. doi: 10.1002/14651858.CD010803.pub2 [published Online First: 2017/03/23]

7. Wijedoru L, Mallett S, Parry CM. Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever. *The Cochrane database of systematic reviews* 2017;5:Cd008892. doi: 10.1002/14651858.CD008892.pub2 [published Online First: 2017/05/26]

8. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *The Cochrane database of systematic reviews* 2014(1):Cd009593. doi: 10.1002/14651858.CD009593.pub3 [published Online First: 2014/01/23]

9. Wang LW, Fahim MA, Hayen A, Mitchell RL, Baines L, Lord S, Craig JC, Webster AC. Cardiac testing for coronary artery disease in potential kidney transplant recipients. *The Cochrane database of systematic reviews* 2011(12):Cd008691. doi: 10.1002/14651858.CD008691.pub2 [published Online First: 2011/12/14]

10. Shaikh N, Spingarn RB, Hum SW. Dimercaptosuccinic acid scan or ultrasound in screening for vesicoureteral reflux among children with urinary tract infections. *The Cochrane database of systematic reviews* 2016;7:Cd010657. doi: 10.1002/14651858.CD010657.pub2 [published Online First: 2016/07/06]

Appendix C

A worked example for calculating normalized frequencies:
High grade vesicoureteral reflux in children with urinary tract infection

In the Cochrane DTA review of Shaikh et al. (2016) summary sensitivity and specificity were obtained for both ultrasound and dimercaptosuccinic acid renal scans used for detecting vesicoureteral reflux. The authors assumed a pretest probability of 13% in a hypothetical cohort of n=1000 for high grade vesicoureteral reflux. We use the data of the ultrasound analysis for high grade vesicoureteral reflux to calculate the mean, median, interquartile range (IQR), and the range of the disease prevalence and calculate the accompanying normalized frequencies. The authors were able to include 11 studies (2498) participants in the ultrasound analysis for high grade vesicoureteral reflux. This resulted in a summary sensitivity and specificity of 59% and 79%, respectively. From the data of the analysis (test 2 on page 115 of the DTA review), the disease prevalence of the included studies for the ultrasound analysis were calculated: 8%, 9.1%, 9.3%, 15.6%, 16.6%, 17.7%, 19.5%, 20,1%, 21,2%, 23.9%, 31%

Therefore:
Range lower limit = 8%
IQR lower limit = 9.3%
Assumed = 13%
Mean = 17.5%
Median = 17.7%
IQR upper limit = 21.2%
Range upper limit = 31%

Calculating the normalized frequencies

The mean, median, interquartile range, range and the assumed pretest probability were used as pretest probabilities when calculating the normalized frequencies. The sensitivity and specificity used in these calculations will remain unchanged. With this we show that normalized frequencies vary when using different pretest probabilities determined by various methods (see Table).

Table – Normalized frequencies for various pretest probabilities while the sensitivity and specificity remain constant				
Ultrasound detecting high grade vesicoureteral reflux in a hypothetical cohort of n = 1,000 **Summary sensitivity: 59%** **Summary specificity: 79%**				
Pretest probability	True positive	False positive	False negative	True negative
8% (range lower limit)	47	193	33	727
9.3% (IQR lower limit)	55	190	38	717
13% (assumed)	77	183	53	687
17.5% (mean)	103	173	72	652
17.7% (median)	104	173	73	650
21.2% (IQR upper limit)	125	165	87	623
31% (range upper limit)	183	145	127	545
IQR: interquartile range				

Appendix C: Reference

Shaikh N, Spingarn RB, Hum SW. Dimercaptosuccinic acid scan or ultrasound in screening for vesicoureteral reflux among children with urinary tract infections. Cochrane Database Syst Rev. 2016 Jul 5;7:CD010657. doi: 10.1002/14651858.CD010657.pub2. Review. PubMed PMID: 27378557; PubMed Central PMCID: PMC6457894

"I advocate passionate dedication
to the pursuit of short-term goals.
Be micro-ambitious. If you focus too far in front of you,
you won't see the shiny thing out the corner of your eye"

Tim Minchin, Australian comedian

Chapter 9

General discussion

Lessons learned

Diagnostic research may seem as a straightforward practice at first sight, however, there are still many methodological hurdles that require the attention of diagnostic researchers. In this thesis I have outlined some of these hurdles and provided directions towards solving them. The lessons learned in this thesis are:

- Dichotomisation of the presence or absence of the target condition by an imperfect reference standard leads to biased estimates of index test sensitivity and specificity estimates.
- Approaches using probabilistic estimates for presence of the target condition elicited from expert panels yield different diagnostic accuracy estimates of the index test(s) under study compared to the traditional dichotomous target condition classification approach.
- The framework of diagnostic accuracy becomes less appropriate if the final diagnosis is uncertain in a substantial proportion of study participants.
- Overdiagnosis is a complex issue, occurring not only in oncology, but across clinical fields and in different contexts. The lack of consensus on the definition is leading to linguistic confusion and inefficient communication between researchers.
- Our framework explicating the mechanisms leading to overdiagnosis (and other concepts related to 'too much medicine') in context of stages of the clinical pathway, allows researchers to better understand, describe, and investigate overdiagnosis for their specific situation.
- Strategies for reducing 'too much medicine' are dependent on the mechanism(s) through which it occurs and should therefore be tailored to their specific situation.
- Early impact assessment of prediction models using decision analytic modelling has great potential, as it provides professionals with:
    - Insight in the mechanism through which a prediction model should provide health and/or monetary benefit;
    - The likelihood of cost-effectiveness of a prediction model in a given scenario;

    - The ability to adapt and optimize the design of subsequent model impact trials.
- Performing a decision-analytical based impact assessment of a prediction model prior to conducting a randomised impact trial is feasible, provided that there is sufficient data on diagnostic performance, consumption of healthcare resources, and expected effectiveness of subsequent treatments.
- Decision analytical based impact and headroom assessment of a diagnostic test or model at a very early stage of development, gives insight in the potential value of the concept idea of a novel diagnostic test or model.
- There is great heterogeneity in the methods used for selecting pretest probabilities in summary of findings tables in Cochrane diagnostic test accuracy reviews.

In the sections below I will elaborate on two aspects of this thesis in particular, namely the use of expert panels as reference standard in diagnostic research and studying the impact of diagnostic and prognostic prediction models based on prior knowledge before conducting a randomised impact trial.

Expert panels as reference standard

When a single preferred reference test ('gold standard') is lacking, or incorporating multiple tests into a composite reference standard is not feasible (e.g. tests that require qualitative assessment such as MRI scans), expert panels can be used as a reference standard in scientific diagnostic research. However, forcing these experts to reach a final diagnosis (target condition is either absent or present) for each individual is likely to introduce bias in diagnostic accuracy estimates of the index test under evaluation. (Chapter 2) A solution for reducing this bias is providing probabilistic estimates for the presence of the target condition, allowing for more valid estimation of the diagnostic accuracy estimates of the test(s) under study. (Chapter 3) Currently the number of studies asking for non-dichotomous target condition is limited, prohibiting these types of assessments.

Although Chapters 2 and 3 provide insight in the bias introduced by forcing dichotomisation of the target condition (absent or present) by expert panels, and probabilistic estimates of target condition presence are suggested as a potential solution, several methodological challenges regarding expert panel diagnosis remain. In the case study described in Chapter 3, four experts were approached to participate, providing a dichotomous target condition classification and a probabilistic estimate of presence of the target condition for each study participant. However, not every study participant was reviewed by all four experts, resulting in missing values for the outcome (final diagnosis) in some patients. If these values are missing at random, the mean probability of presence of the target condition would not be biased. (1-3) However, it is plausible that these values are not missing at random, thereby biasing the probability of presence of the target condition, and consequently accuracy estimates of the index test(s) or model under evaluation.

One of the key assumptions when using expert panels as a reference standard is that this panel is well calibrated: i.e. the mean of the probabilities of target condition presence provided by each of the panel members for a particular patient, is an accurate estimate of the probability of the presence of the target condition in that patient. This is a strong assumption, as can be demonstrated by the following example: if two experts in a panel provide probabilities for presence of the target condition of 0.1 and 0.9 for the same study individual, then the mean of 0.5 is unlikely to reflect the true probability of disease presence for the individual. Another issue is that experts from varying clinical specialties may also look at diagnostic information differently, hence the composition of the expert panel may affect the mean probability of target condition presence in an individual.

Another key issue is how to view the probabilities of presence of the target condition given by the expert panel. One possibility is that the results of the index test are incorporated in these probabilities; then the weighting method is then appropriate approach for calculating accuracy. Incorporated here means that, either these results are formally available to the experts, or the information from the index test is captured by information from other tests or follow-up. The drawback of making index test results available to the expert is the fear of incorporation bias. This means that the experts base

their final diagnosis to strongly on the index test results, because the index test is new and sounds promising, overestimating its value. An alternative for an expert panel would be a latent class model. Latent class models acknowledge that a gold standard does not exist and that the available pieces of information are all related to an unknown (latent) true status: target condition present or absent. Diagnostic accuracy measures for the test(s) under study can be estimated by the latent class model. The benefit is that the statistical latent class model is based on the observed patterns in the data and is not influenced by expectations about the new index test. Drawbacks come from the statistical definition of the target condition which may deviate from the clinical relevant interpretation of what the condition entails. Furthermore, results of the latent class model are directly subject to assumptions made during modelling.

Directions for future research

The results of our studies suggest that probabilistic estimates of target condition presence provide more insight in the certainty surrounding target condition classification by the expert panel, and may help provide more accurate estimates for the index test(s) under evaluation. Acquiring these estimates in all studies involving expert panels is of key importance, as this information cannot be obtained after the study has been performed. Ideally, probabilistic estimates become part of formal guidance on methodology and reporting of the use of expert panels in future studies.

On a methodological level there are still significant challenges with regard to analysis of probabilistic estimates of target condition presence obtained from expert panels. As described above, non-random missing values of probabilistic estimates of target condition presence could lead to biased accuracy estimates of the index test under study. Furthermore, probabilistic estimates can be missing at random both within study participants (e.g. experts might be more reluctant in giving probability of target disease presence in more complex patients) or within experts (e.g. some experts might be more committed to providing a probability of target disease presence for each patient). Future (e.g. simulation) studies could provide insight in the effects of different patterns of missingness on the mean probability of target condition presence within study participants, as well as

the impact on accuracy estimates of the diagnostic index test(s) or model under evaluation.

Another topic for future research is how to take uncertainty between experts in a panel into account in the analysis. Alternatives to solely taking the mean value of probabilities of target condition presence within a study participant might be sought after. For example, sensitivity analyses can be performed, in which the outliers of expert panel estimates within each study participant are excluded, or that these outliers may receive less weight when calculating the mean probability of target condition presence. Other methods for incorporating uncertainty of the assigned expert probabilities include using a beta-binomial distribution or bootstrapping to reflect uncertainty in the target condition presence estimates, and by extension the index test or prediction model accuracy estimates. Simulation studies, reflecting a series of hypothetical scenarios, may provide more insight in the validity of each of these methods.

A more fundamental issue arises when there is still considerable uncertainty about target condition presence in a substantial proportion of patients. In this situation, the traditional diagnostic accuracy framework may no longer be appropriate when evaluating new diagnostic tests. In such scenarios, abandoning the diagnostic accuracy framework may generate more insight. Future research may focus on exploring alternatives to this framework, such as relating index test results to the risk of future clinical events, or assessing whether test results can discriminate who will have the most benefit from an intervention in a randomised trial.

Decision analytical modelling for studying impact of diagnostic and prognostic prediction models

Impact assessment of diagnostic and prognostic prediction models is vital to ensure that they have a significant benefit on patient health outcomes and/or healthcare costs when they are implemented in daily practice. There is no straightforward association between a model's predictive accuracy and its clinical utility. Hence there is no single threshold value that determines what is considered "accurate enough", warranting use of a model in patient care.

(4) Trials are often considered to establish the impact of implementing a model in clinical practice on patient relevant outcomes and/or healthcare costs. However, as stated in Chapter 6, clinical impact trials of prediction models are yet rare due to the fact that they are complex in their organization, time-consuming, and costly. Even when these trials are performed, the impact of a prediction model on health outcomes and / or healthcare costs observed in such trials is often below expectations, contributing to research waste. (5, 6)

Thousands of prediction models have been developed throughout the years, with for example already 363 developed and another 473 validated for cardiovascular disease alone (7), in which the main aim is to assess discrimination and calibration of the prediction model under study. Only a small proportion of these models are used in clinical daily practice, most of which have never undergone impact assessment.

A systematic overview of impact studies of diagnostic or prognostic prediction models on clinical decision making and patient outcome in randomised trials is, to our knowledge, still lacking. However, there are some case examples worth mentioning. A trial on a prediction model aimed at providing prophylactic anti-emetics for post-operative nausea and vomiting showed that the impact of using the prediction model in clinical practice did not reduce event rate below what was beforehand expected. (6) Another trial in the UK looked at a prediction model for reducing emergency hospital admissions in an elderly population. (8) Their conclusion was unambiguous: "Introduction of PRISM increased emergency episodes, hospitalisation and costs across, and within, risk levels without clear evidence of benefits to patients." Finally, in a recent trial, a prediction model was used to identify people at high risk of not returning to work after orthopaedic surgery, and provide them with extra coaching or evaluation. (9) They found no significant clinical improvement, attributing it to either insufficient directive guidelines for management or fear of disadvantaging patients.

Decision analytic modelling (DAM) prior to conducting a randomised prediction model impact trial (as described in Chapter 6) is an important alternative, if not prerequisite, before performing a costly (cluster)

randomised trial. (5, 10) DAMs can be performed in a fraction of the time of an actual impact trial, resulting in minimal research time and costs to be invested, leading to less research waste in prediction model research.

Exactly how many randomised trials have been performed for all the thousands of prediction models is unknown. There is, however, a systematic review of all risk prediction models for which a modelling assessment has been performed. (11) A total of 60 prediction models were identified (only a small proportion of the thousands of developed prediction models) for which a modelling assessment had been performed. Quality of analyses and reporting varied between these studies. Whether the authors of these modelling assessments reported positive, negative, or neutral impact of the prediction models under study on health outcomes or costs, is not stated in the systematic review. Neither can we be sure that this is a comprehensive overview, as the search was restricted to HTA related outcomes.

As stated before, disappointing results of a randomised trial on a prediction model's impact on health outcomes and healthcare costs, contributes to research waste. Worldwide numbers related to research waste are scarce, but estimates state that approximately 85% of money spent in research is being wasted. (12, 13) It is has been stated that study design flaws and incomplete or inaccurate reporting account for two thirds of that number.

Efforts to reduce research waste in medicine have been described in several key publications. (14, 15) Among the features for clinical useful research are the use of other sources of evidence besides randomised trials, value for money, and patient centeredness. DAMs, used to quantify the impact of a diagnostic or prognostic prediction model without having to do a large scale randomised trials, tick the first box, as they are a clear alternative to costly impact trials. Second, DAMs provide value for money, as they can be performed by a single person, don't require lengthy regulatory approval by medical ethical committees, and take months rather than years to complete. Lastly, patient centeredness can be assured by involving patients during model construction phase and by using patient relevant outcomes such as QALYs.

Directions for future research

DAMs are a valuable asset to evaluate impact of prediction models at an early stage of their development. There are many opportunities for expanding this relatively novel field of prediction model research.

A comprehensive overview of both model-based and trial-based impact assessments of prediction models, looking at the methods used, study quality, and proportion of studies in which the impact of a prediction model was on par with expectations, will serve the field. Such a systematic review could also provide the starting point of how quality assessment tools, such as the Cochrane risk of bias tool (for randomised trials), the Drummond checklist (for health economic evaluations) and PROBAST (for development or validation of prediction model studies), can be used for quality assessment of in the mixed field of prediction model impact research. (16-19)

Currently there is also still a lack of clear guidance on how to use DAMs specifically for assessing the impact of prediction models on health outcomes and healthcare costs without empirical evidence from a randomised trial. Though we have provided a short piece of guidance in Chapter 6, the exact methodological challenges and choices that researchers are faced with at each stage have not been fully elaborated on. Furthermore, there are inherently greater limitations regarding data availability, as resources tend to be limited at an early stage of prediction model development. An overview of where such resources can be located, could prove to be a practical and valuable tool for researchers aiming to perform DAMs. Other guidance may include considering implications of the intended use of a prediction model (diagnostic or prognostic) and choice of management recommendations associated with predicted risk categories of the model on DAM structure. Information on methodologies particularly useful at an early stage of development, such as value of information and headroom analysis, may also provide useful to researchers. (20-22)

Finally, it would be worthwhile to look beyond the early impact assessment of prediction models: what does it provide us, and in what way can we use its outcomes? Its results could be used for a hard go/no-go decision, as

randomised impact trials for prediction models with either a very high or very low likelihood of success, may not be deemed necessary. However, a DAM? could also provide a starting point for additional research before deciding whether or not a trial is warranted. An example of this is a qualitative pilot study, looking at clinicians' view on and support of prediction model management recommendations. Furthermore, sample size calculations could be performed based on the results of a DAM analysis, and multiple-criteria decision-making (MCDM) could be used to incorporate results on cost-effectiveness with additional items (e.g. appropriateness, urgency) in a broader context. (23) Providing researchers with more guidance on the use and interpretation of DAMs before conducting a costly randomised impact trial, could reduce uncertainty on the impact of a prediction model on health and monetary outcomes, allowing more appropriate decision-making on which prediction models require randomised trials, and ultimately increase the efficiency and value of such trials.

References

1. Royston P. Multiple Imputation of Missing Values. The Stata Journal. 2004;4(3):227-41.

2. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine. 2010;29(28):2920-31. Epub 2010/09/16.

3. Rubin D. Inference and missing data. Biometrika 1976;63(3):581–92.

4. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. The American statistician. 2008;62(4):314-20. Epub 2009/01/10.

5. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagnostic and Prognostic Research. 2018;2(1):11.

6. Kappen TH, Moons KG, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, van Klei WA. Impact of risk assessments on prophylactic antiemetic prescription and the incidence of postoperative nausea and vomiting: a cluster-randomized trial. Anesthesiology. 2014;120(2):343-54. Epub 2013/10/10.

7. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416. Epub 2016/05/18.

8. Snooks H, Bailey-Jones K, Burge-Jones D, Dale J, Davies J, Evans B, et al. Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC). Southampton (UK)2018.

9. Plomb-Holmes C, Hilfiker R, Leger B, Luthi F. Impact of a non-return-to-work prognostic model (WORRK) on allocation to rehabilitation clinical pathways: A single centre parallel group randomised trial. PLoS One. 2018;13(8):e0201687. Epub 2018/08/03.

10. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8. Epub 2012/03/09.

11. van Giessen A, Peters J, Wilcher B, Hyde C, Moons C, de Wit A, et al. Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2017;20(4):718-26. Epub 2017/04/15.

12. Glasziou P, Chalmers I. Is 85% of health research really "wasted". The BMJ. 2016.

13. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009;374(9683):86-9. Epub 2009/06/16.

14. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383(9912):166-75. Epub 2014/01/15.

15. Ioannidis JP. Why Most Clinical Research Is Not Useful. PLoS Med. 2016;13(6):e1002049. Epub 2016/06/22.

16. Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the economic evaluation of health care programme. 3rd ed. Oxford: Oxford University Press; 2005.

17. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. Cochrane Database of Systematic Reviews. 2016(10).

18. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019;170(1):W1-W33. Epub 2019/01/01.

19. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51-8. Epub 2019/01/01.

20. McAteer H, Cosh E, Freeman G, Pandit A, Wood P, Lilford R. Cost-effectiveness analysis at the development phase of a potential health technology: examples based on tissue engineering of bladder and urethra. Journal of tissue engineering and regenerative medicine. 2007;1(5):343-9. Epub 2007/11/27.

21. Cosh E, Girling A, Lilford R, McAteer H, Young T. Investing in new medical technologies: A decision framework. Journal of commercial biotechnology. 2007;13(4):263-71.

22. Tuffaha HW, Gordon LG, Scuffham PA. Value of information analysis in healthcare: a review of principles and applications. Journal of medical economics. 2014;17(6):377-83. Epub 2014/03/22.

23. Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making--An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2016;19(1):1-13. Epub 2016/01/23.

"I would if I could, but I can't, so I won't"

Poster from my English classroom in high school

Summary

Diagnostic research may seem straightforward at first sight. We provide diagnostic labels to patients to indicate in whom a target condition is present or absent. Those labels respectively distinguish who will benefit from intervention and who won't. Implementing a perfect accurate test or (diagnostic or prognostic) prediction model will result in better health outcomes for patients and/or a reduction in healthcare costs. Unfortunately, reality is not so simplistic, warranting the use of more complex methodology to evaluate these tests and prediction models. The aim of this thesis is to address methodological issues in evaluation and impact assessment of diagnostic tests and models, and propose alternative approaches to reduce bias, research waste and improve communication by providing methodological guidance.

The objective in **Chapter 2** is to study the impact of ignoring uncertainty by forcing dichotomous classification (presence or absence) of the target disease on estimates of diagnostic accuracy of an index test. We evaluated the bias in estimated index test accuracy when forcing an expert panel to make a dichotomous target disease classification for each individual. Data for various scenarios with expert panels were simulated by varying the number and accuracy of "component reference tests" available to the expert panel, index test sensitivity and specificity, and target disease prevalence. The results showed that index test accuracy estimates are likely to be biased when there is uncertainty surrounding the presence or absence of the target disease. Direction and amount of bias depend on the number and accuracy of component reference tests, target disease prevalence, and the true values of index test sensitivity and specificity. In this simulation, forcing expert panels to make a dichotomous decision on target disease classification in the presence of uncertainty leads to biased estimates of index test accuracy. Empirical studies are needed to demonstrate whether this bias can be reduced by assigning a probability of target disease presence for each individual, or using advanced statistical methods to account for uncertainty in target disease classification.

Expert panels, used as reference standard in diagnostic accuracy studies, typically classify each patient as having or not having the target condition,

even when they have remaining uncertainty about that classification. This has been shown to lead to biased diagnostic accuracy estimates of the index test. **Chapter 3** aims to show how probabilistic estimates of presence of a target condition elicited from an expert panel can be used in diagnostic accuracy research. The SPACE (SePsis in Acutely ill patients in the Emergency department) study, aimed at investigating the diagnostic value of clinical decision rules (SIRS, qSOFA, CBJ) for diagnosis of sepsis in the emergency room, was used as a case study. Both dichotomous (i.e. present or absent) and probabilistic estimates of sepsis status were obtained from the expert panels. Measures of diagnostic accuracy were calculated using three approaches: (1) (traditional) dichotomous sepsis classification; (2) an approach using probabilistic estimates for presence of sepsis as weights; (3) an approach using these probabilistic estimates in combination with the diagnostic odds ratio (DOR). A total of 306 patients were included in the analysis. A skewed distribution of probabilistic estimates for the presence of sepsis by the panel was observed (median=0.30). The panel expressed considerable uncertainty whether sepsis was present or not (probabilities between 0.2 and 0.8) in 57% of patients. Estimates of diagnostic accuracy varied considerably between the dichotomous and two probabilistic approaches, but also between the two probabilistic approaches. For example, sensitivity of SIRS was 91% for the dichotomous approach, 74% for the probabilistic weighting approach, and 99% for probabilistic DOR approach. Specificity was 46%, 47% and 60% for these approaches respectively. Eliciting probabilistic estimates of target condition presence from expert panels can provide valuable insight in the uncertainty that is normally ignored in dichotomous target disease classification. Different approaches exist on how to incorporate this uncertainty when estimating diagnostic accuracy measures and results can vary substantially depending on the assumptions made. When substantial uncertainty about the final diagnosis is present in a considerable proportion of patients, it may be questioned whether the diagnostic accuracy framework is still useful.

Chapter 4 focuses on overdiagnosis, aiming to provide insight into how and in what clinical fields overdiagnosis is studied and give directions for further applied and methodological research. A scoping review was performed in

which Medline was searched. All English studies on humans published up to August 2017 in which overdiagnosis was discussed as a dominant theme were included. Studies were assessed on clinical field, study aim (i.e. methodological or non-methodological), article type (e.g. primary study, review), the type and role of diagnostic test(s) studied and the context in which these studies discussed overdiagnosis. From 4896 studies, 1851 were included for analysis. Half of all studies on overdiagnosis were performed in the field of oncology (50%). Other prevalent clinical fields included mental disorders, infectious diseases and cardiovascular diseases accounting for 9%, 8% and 6% of studies, respectively. Overdiagnosis was addressed from a methodological perspective in 20% of studies. Primary studies were the most common article type (58%). The type of diagnostic tests most commonly studied were imaging tests (32%), although these were predominantly seen in oncology and cardiovascular disease (84%). Diagnostic tests were studied in a screening setting in 43% of all studies, but as high as 75% of all oncological studies. The context in which studies addressed overdiagnosis related most frequently to its estimation, accounting for 53%. Methodology on overdiagnosis estimation and definition provided a source for extensive discussion. Other contexts of discussion included definition of disease, overdiagnosis communication, trends in increasing disease prevalence, drivers and consequences of overdiagnosis, incidental findings and genomics. Overdiagnosis is discussed across virtually all clinical fields and in different contexts. The variability in characteristics between studies and lack of consensus on overdiagnosis definition indicate the need for a uniform typology to improve coherence and comparability of studies on overdiagnosis.

Concepts related to 'too much medicine' (such as overdiagnosis) remain a complex multifaceted issue, difficult to grasp and dissect. Although valuable descriptive frameworks have been proposed, these have not tackled the issues related to too much medicine across clinical domains, nor have they provided actionable strategies for reducing them. In **Chapter 5** we provide a conceptual framework aimed at distinguishing uncertainty over thresholds and errors, two key mechanisms leading to 'too much medicine', and placing these in the clinical pathway of screening, diagnosis, prognosis and treatment

of individuals. This allows researchers to evaluate concepts related to 'too much medicine' in the context of their own specific research, and facilitates communication between researchers, healthcare providers and patients. Based on the mechanism(s) at play, we provide strategies for reducing too much medicine.

Chapter 6 continues on the topic of evaluating impact of prediction models. The main goal in this chapter was to demonstrate how decision analytic models (DAM) can be used to quantify impact of using a (diagnostic or prognostic) prediction model in clinical practice, and provide general guidance on how to perform such assessments. A DAM was developed to assess the impact of using the HEART score for predicting major adverse cardiac events (MACE). Impact on patient health outcomes and healthcare costs was assessed in scenarios by varying compliance with and informed deviation (ID) (using additional clinical knowledge) from HEART score management recommendations. Probabilistic sensitivity analysis was used to assess estimated impact robustness. Impact of using the HEART score on health outcomes and healthcare costs was influenced by interplay of compliance with and ID from HEART score management recommendations. Compliance of 50% (with 0% ID) resulted in increased missed MACE and costs compared to usual care. Any compliance combined with at least 50% ID, reduced both costs and missed MACE. Other scenarios yielded a reduction in missed MACE at higher costs. DAM is a useful approach to assess impact of using a prediction model in practice on health outcomes and healthcare costs. This approach is recommended before conducting an impact trial to improve its design and conduct.

With seemingly unlimited technological possibilities yet limited budgets, clinicians face the challenge of which novel ideas to pursue and which to lay aside. Although early health economic modelling methods may support innovation decisions, they are not yet widely known or used in the evaluation of diagnostic tests and prediction models. The aim in **Chapter 7** is to illustrate early health economic modelling to clinicians by applying its methods to the case of diagnosing primary aldosteronism (PA) in patients with hypertension. We developed a cohort state-transition model to simulate diagnosis, treatment, and long-term health outcomes for patients aged ≥40

years with resistant hypertension suspected of PA. We included relevant literature and Dutch costing data and took a lifetime, societal perspective on costs and health effects (quality-adjusted life-years, QALYs). In our model we compared the current aldosterone-to-renin ratio test for diagnosing PA to a hypothetical new test. During a patient's lifetime, a perfect diagnostic test would yield 0.027 QALYs and increase costs by €43. At a cost-effectiveness threshold of €20,000 per QALY, the maximum price for this perfect test to be cost-effective is €498 (95% CI: €275 - €808). The value of the perfect test was most strongly influenced by the sensitivity of the current biomarker test. Threshold analysis showed the novel test needs a sensitivity of at least 0.9 and a specificity of at least 0.7 to be cost-effective. Applying a model-based approach to determine the added value of a clinical innovation in PA diagnostics, we demonstrated there was room for improvement while indicating a maximum price per test, supporting the conclusion that early health economic modelling is useful and feasible in clinical practice to determine the cost-effectiveness of novel ideas prior to extensive development activities and clinical implementation. More applications of early modelling through collaborations between health economists and clinical experts will illustrate the benefits and help further the accessibility of early health economic modelling in dealing with innovation.

To facilitate the interpretation of diagnostic test accuracy (DTA) parameters it is possible to calculate normalized frequencies. They provide the number of (false) positives and (false) negatives in a tested hypothetical cohort. A pretest probability must be determined to calculate these normalized frequencies. The aim of **Chapter 8** is to assess the data sources and methods used in Cochrane DTA reviews for determining pretest probabilities to facilitate the interpretation of DTA parameters. Cochrane DTA reviews published in the Cochrane Database of Systematic Reviews up to and including January 2018 and presenting at least one meta-analytic estimate of the sensitivity and/or specificity as a primary analysis were included in the cohort. Study selection and data extraction were performed by one author and checked by other authors. Observed data sources and methods were categorized. Fifty-nine DTA reviews were included, comprising of 308 meta-analyses. A pretest probability was used in 148 meta-analyses. Authors used

included studies in the DTA review, external sources, and expert opinion as data sources for the pretest probability. When using the included studies in the DTA review, authors used a measure of central tendency whether or not combined with a measure of dispersion to determine the pretest probabilities. Identical pretest probabilities were used for analyses of two or more index tests for the same target conditions. About half (53.6%) of these identical pretest probabilities fell within the prevalence ranges from all analyses within a target condition. Various methods are used for selecting pretest probabilities and no consensus seems to exist on which data source or method to use. However, there are some considerations to take in to account when presenting DTA results: 1) Consider whether to present normalized frequencies, 2) Consider the influence of the chosen method for selecting a pretest probability on the normalized frequencies, and 3) Consider whether to use identical pretest probabilities that fall within the range of the selected studies when there are multiple meta-analyses for a target condition.

The general discussion of this thesis (**Chapter 9**) directs attention to two topics: methods for using expert panels as a reference standard in diagnostic research, and methods for early assessment of the impact of prediction models on health outcomes and healthcare costs. Remaining challenges are described for both topics and suggestions for future research are given. These suggestions include dealing with missing estimates of probability of target disease presence by the expert panel, taking into account the uncertainty between experts within a panel, and providing more guidance on the use and interpretation of DAMs before conducting a randomised impact trial. Focus on these methodological topics should ultimately improve efficiency and value of diagnostic research, and reduce research waste.

"God might play dice with the universe

*but they are **THE BEST** dice in the universe"*

Michael "Vsauce" Stevens, science educator
(sic. of quantum mechanics)

Samenvatting

Diagnostisch onderzoek lijkt op het eerste gezicht misschien eenvoudig. We geven diagnostische labels aan patiënten om aan te duiden bij wie een aandoening aanwezig of afwezig is. Die labels onderscheiden respectievelijk wie van interventie zou profiteren en wie niet. Implementatie van een perfect accurate test of (diagnostisch of prognostisch) voorspellingsmodel zal leiden tot gezondheidswinst voor patiënten en / of een verlaging van de ziektekosten. Helaas is de realiteit niet zo simplistisch, waardoor complexere methodologie nodig is om deze testen en voorspellingsmodellen te evalueren. Het doel van dit proefschrift is om methodologische kwesties rondom de evaluatie van (de impact van) diagnostische testen en predictie modellen onder de loep te nemen, alternatieve aanpakken voor te stellen om zo bias en verspilling van onderzoek te verminderen, en communicatie te verbeteren door het aanreiken van methodologische handvatten.

Het doel in hoofdstuk 2 is om de impact te bestuderen van het negeren van onzekerheid door het forceren van dichotome classificatie (aanwezigheid of afwezigheid) van de aandoening op schattingen van de diagnostische accuratesse van een indextest (de test waarin we geïnteresseerd zijn). We evalueerden de bias van de geschatte indextest accuratesse door het expertpanel te forceren om een dichotome ziekteclassificatie te maken voor elk individu. Data voor verschillende scenario's met expertpanels werden gesimuleerd door het aantal en de accuratesse van 'componentreferentietesten' beschikbaar voor het expertpanel, de sensitiviteit en specificiteit van de indextest, en de ziekteprevalentie te variëren. De resultaten lieten zien dat schattingen van de accuratesse van indextesten waarschijnlijk vertekend zijn wanneer er onzekerheid bestaat over de aanwezigheid of afwezigheid van aandoening. De richting en mate van bias hangen af van het aantal en de nauwkeurigheid van componentreferentietesten, de ziekteprevalentie en de werkelijke waarden van de sensitiviteit en specificiteit van de indextest. In deze simulatiestudie leidde het forceren van expertpanels om een dichotome beslissing te nemen over ziekteclassificatie in de aanwezigheid van onzekerheid tot vertekende schattingen van de accuratesse van de indextest. Empirische studies zijn nodig om aan te tonen of deze bias kan worden verminderd door kans op aanwezigheid van de aandoening toe te laten wijzen voor elk individu, of

door geavanceerde statistische methoden te gebruiken om rekening te houden met onzekerheid in de classificatie van aandoening.

Expertpanels, die als referentiestandaard worden gebruikt in diagnostische accuratesse studies, classificeren doorgaans elke patiënt met het wel of niet hebben van de desbetreffende aandoening, zelfs als ze onzeker zijn over die classificatie. Er is aangetoond dat dit leidt tot vertekende schattingen van de diagnostische accuratesse van de indextest. Hoofdstuk 3 beoogt te laten zien hoe probabilistische schattingen van de aanwezigheid van aandoening verkregen van een panel van experts kunnen worden gebruikt in diagnostisch accuratesse onderzoek. De SPACE studie, gericht op het onderzoeken van de diagnostische waarde van klinische beslisregels (SIRS, qSOFA, CBJ) voor de diagnose van sepsis op de spoedeisende hulp, werd gebruikt als casus. Zowel dichotome (d.w.z. aanwezig of afwezig) en probabilistische schattingen van de sepsis-status werden verkregen van het expertpanel. Uitkomstmaten van diagnostische accuratesse werden berekend met behulp van drie methoden: (1) (traditionele) dichotome sepsis-classificatie; (2) een benadering die probabilistische schattingen voor de aanwezigheid van sepsis gebruikt als weging; (3) een benadering die probabilistische schattingen gebruikt in combinatie met de diagnostische odds ratio (DOR). Een totaal van 306 patiënten werden geïncludeerd in de analyse. Er was een scheve verdeling van de van het expertpanel verkregen probabilistische schattingen op aanwezigheid van sepsis (mediaan = 0,30). Het panel toonde aanzienlijke onzekerheid of sepsis al dan niet aanwezig was (kans tussen 0,2 en 0,8) bij 57% van de patiënten. Schattingen van diagnostische accuratesse varieerden aanzienlijk tussen de dichotome en twee probabilistische methoden, maar ook tussen de twee probabilistische methoden. De sensitiviteit van SIRS was bijvoorbeeld 91% voor de dichotome methode, 74% voor de probabilistische wegingsmethode en 99% voor de probabilistische DOR-methode. De specificiteit was respectievelijk 46%, 47% en 60% voor deze methoden. Het verkrijgen van probabilistische schattingen op de aanwezigheid van aandoening van expertpanels kan waardevol inzicht verschaffen in de onzekerheid rondom dichotome classificatie die normaal wordt genegeerd. Er bestaan verschillende methoden om deze onzekerheid mee te nemen bij het schatten van diagnostische accuratesse van een

indextest, en de resultaten kunnen aanzienlijk variëren afhankelijk van de gemaakte aannames. Wanneer er substantiële onzekerheid aanwezig is bij de diagnoses van een aanzienlijk deel van de patiënten, kan het de vraag zijn of het diagnostische accuratesse raamwerk nog steeds bruikbaar is.

Hoofdstuk 4 richt zich op overdiagnose, met als doel inzicht te geven in hoe en in welke klinische gebieden overdiagnose wordt bestudeerd, en richtingen voor te stellen voor verder toegepast en methodologisch onderzoek. Er is een verkennend literatuuronderzoek uitgevoerd waarbij Medline is doorzocht. Alle Engelse studies die onderzoek deden in mensen, gepubliceerd tot augustus 2017, waarbij overdiagnose als een dominant thema werd besproken, werden geïncludeerd. Studies werden geclassificeerd op klinisch gebied, onderzoeksdoel (d.w.z. methodologisch of niet-methodologisch), artikeltype (bijv. primair of overzichtsonderzoek), het type en de rol van diagnostische test(en) die werden onderzocht, en de context waarin deze studies overdiagnose bespraken. Uit 4896 studies werden 1851 verder geanalyseerd. De helft van alle onderzoeken naar overdiagnose werd uitgevoerd op het gebied van oncologie (50%). Andere veelvoorkomende klinische velden waren mentale stoornissen, infectieziekten, en hart- en vaatziekten, die respectievelijk 9%, 8%, en 6% van de studies vertegenwoordigden. Overdiagnose werd bekeken vanuit een methodologisch perspectief in 20% van de onderzoeken. Primaire studies waren het meest voorkomende artikeltype (58%). Het type diagnostische tests dat het meest werd bestudeerd waren beeldvormende tests (32%), hoewel deze voornamelijk werden gezien in oncologie en cardiovasculaire aandoeningen (84%). Diagnostische testen werden in een screening situatie bestudeerd in 43% van alle onderzoeken, en maar liefst in 75% van alle oncologische onderzoeken. De context waarin studies overdiagnose bespraken was het meest frequent gerelateerd aan de schatting ervan (53%). Methodologie van de schatting en definitie van overdiagnose vormden een bron voor uitgebreide discussie. Andere contexten van discussie waren de definitie van ziekte, communicatie van overdiagnose, trends in toenemende prevalentie van aandoening, drijfveren en de gevolgen van overdiagnose, incidentele bevindingen, en genetica. Overdiagnose wordt besproken op vrijwel alle klinische gebieden en in verschillende contexten. De diversiteit in

kenmerken tussen studies en het ontbreken van consensus over de definitie van overdiagnose geven de noodzaak aan van een uniforme typologie om de coherentie en vergelijkbaarheid van studies die overdiagnose bespreken te verbeteren.

Concepten met betrekking tot Too much medicine (zoals overdiagnose) blijven complex, veelzijdig, en moeilijk te begrijpen en ontleden kwesties. Hoewel er waardevolle descriptieve raamwerken zijn voorgesteld, hebben deze niet de problemen aangepakt die verband houden met Too much medicine in verschillende klinische domeinen, en hebben ze geen bruikbare strategieën geboden om Too much medicine te reduceren. In hoofdstuk 5 bieden we een conceptueel raamwerk gericht op het onderscheiden van onzekerheid over afkapwaarden en afwijkingen, twee belangrijke mechanismen die leiden tot Too much medicine, en plaatsen we deze in het klinische pad van screening, diagnose, prognose en behandeling van individuen. Dit stelt onderzoekers in staat om concepten gerelateerd aan Too much medicine te evalueren in de context van hun eigen specifieke onderzoek, en vergemakkelijkt het de communicatie tussen onderzoekers, zorgverleners en patiënten. Gebaseerd op de mechanismes die er spelen, stellen wij strategieën voor ten behoeve van het verminderen van Too much medicine.

Hoofdstuk 6 gaat verder over het evalueren van de impact van voorspellingsmodellen. Het belangrijkste doel van dit hoofdstuk was om aan te tonen hoe besliskundige modellen, ofwel *decision analytic models* (DAM's), kunnen worden gebruikt om de impact van het gebruik van een (diagnostisch of prognostisch) voorspellingsmodel in de klinische praktijk te kwantificeren en algemene richtlijnen te geven voor het uitvoeren van dergelijke evaluaties. Er is een DAM ontwikkeld om de impact te onderzoeken van het gebruik van de HEART-score voor het voorspellen van belangrijke ongunstige cardiovasculaire uitkomsten, ofwel *major adverse cardiac events* (MACE). De impact op gezondheidsuitkomsten en zorgkosten van patiënten werd in verschillende scenario's onderzocht door naleving van aanbevelingen voor behandeling op basis van de HEART-score, en geïnformeerde afwijking hiervan (door aanvullende klinische kennis), ofwel *informed deviation* (ID), te variëren. Probabilistische sensitiviteitsanalyse werd gebruikt om de

robuustheid van de gevonden resultaten te beoordelen. Het effect van het gebruik van de HEART-score op gezondheidsuitkomsten en kosten voor gezondheidszorg werd beïnvloed door een samenspel van naleving en ID van aanbevelingen voor behandeling op basis van de HEART-score. Naleving van 50% (met 0% ID) resulteerde in een toename van gemiste MACE en kosten in vergelijking met de gebruikelijke zorg. Voor elk scenario van naleving gecombineerd met minimaal 50% ID, werden zowel de kosten als het gemiste MACE gereduceerd. Andere scenario's leidden tot een afname van het aantal gemiste MACE tegenover hogere kosten. DAM is een nuttige methode om de impact van een voorspellingsmodel op gezondheidsuitkomsten en ziektekosten te evalueren wanneer deze in de dagelijkse praktijk toegepast zou worden. Deze aanpak wordt aanbevolen voor het uitvoeren van een klinische impactstudie, zodat tijdig het ontwerp en de uitvoering kunnen worden verbeterd.

Met schijnbaar onbeperkte technologische mogelijkheden en beperkte budgetten staan clinici voor de uitdaging welke nieuwe ideeën moeten worden nagestreefd en welke terzijde moeten worden geschoven. Het vroegtijdig uitvoeren van gezondheidseconomische evaluaties biedt mogelijk ondersteuning rondom beslissingen omtrent innovaties, maar het is nog niet algemeen bekend of ze tevens gebruikt zouden kunnen worden bij de evaluatie van diagnostische testen en voorspellingsmodellen. Het doel in hoofdstuk 7 is om vroege gezondheidseconomische evaluaties aan clinici te illustreren door deze methodiek toe te passen bij het diagnosticeren van primair aldosteronisme (PA) bij patiënten met hypertensie. We hebben een cohort toestand transitie model ontwikkeld om de diagnose, behandeling en gezondheidsuitkomsten op de lange termijn te simuleren voor patiënten van 40 jaar of ouder met resistente hypertensie die van PA worden verdacht. We hebben relevante literatuur en Nederlandse kostendata gebruikt, in combinatie met een levenslang, maatschappelijk perspectief op kosten en gezondheidseffecten (op kwaliteit gecorrigeerde levensjaren, *quality adjusted life-years*, QALY's). In ons model hebben we de huidige test, het aldosteron-renine-ratio, voor het diagnosticeren van PA vergeleken met een hypothetische nieuwe test. Tijdens de levensduur van een patiënt zou een perfecte diagnostische test 0,027 QALY's opleveren en de kosten verhogen

met € 43. Bij een kosteneffectiviteitsdrempel van € 20.000 per QALY is de maximale prijs waarbij deze perfecte test nog kosteneffectief is € 498 (95% BI: € 275 - € 808). De waarde van de perfecte test werd het sterkst beïnvloed door de sensitiviteit van de huidige biomarkertest. *Threshold* analyse liet zien dat de nieuwe test een sensitiviteit van ten minste 0,9 en een specificiteit van ten minste 0,7 nodig heeft om kosteneffectief te zijn. Door een modelmatige aanpak toe te passen voor het bepalen van de toegevoegde waarde van een klinische innovatie in PA-diagnostiek, hebben we aangetoond dat er ruimte was voor verbetering onder de assumptie van een maximale prijs per test. Dit ondersteunt de conclusie dat vroegtijdige economische evaluatie bruikbaar en haalbaar is om te bepalen of nieuwe innovaties of ideeën kosteneffectief kunnen zijn voorafgaand aan uitgebreide doorontwikkeling en klinische implementatie. Door samenwerkingen tussen gezondheidseconomen en klinische deskundigen zullen meer toepassingen en voordelen van vroegtijdig modellering geïllustreerd kunnen worden, en zal de toegankelijkheid van vroege economische modellen ten behoeve van evaluatie van nieuwe innovaties bevorderen.

Om de interpretatie van diagnostische test accuratesse (DTA) parameters te vergemakkelijken, is het mogelijk om genormaliseerde frequenties te berekenen. Deze geven het aantal (fout) positieve en (fout) negatieve testresultaten in een hypothetisch cohort. Een vooraf-kans moet worden bepaald om deze genormaliseerde frequenties te berekenen. Het doel van hoofdstuk 8 is om te onderzoeken welke gegevensbronnen en methoden worden gebruikt in Cochrane DTA literatuuroverzichten ten behoeve van het bepalen van de vooraf-kans om zo de interpretatie van DTA parameters te vergemakkelijken. Cochrane DTA literatuuroverzichten gepubliceerd in de Cochrane database tot en met januari 2018, met ten minste één meta-analytische schatting van de sensitiviteit en / of specificiteit als primaire analyse, werden opgenomen in het cohort. Studie selectie en data-extractie werden uitgevoerd door één auteur en gecontroleerd door andere auteurs. De gebruikte gegevensbronnen en methoden werden gecategoriseerd. Vijfenvijftig DTA evaluaties werden geïncludeerd, bestaande uit 308 meta-analyses. Een vooraf-kans werd gebruikt in 148 meta-analyses. Auteurs gebruikten de onderzoeken in hun DTA literatuuronderzoek, externe

bronnen, en expert opinie als gegevensbronnen bij het bepalen van de vooraf-kans. Bij gebruik van studies uit de DTA literatuuroverzichten, gebruikten auteurs een mate van centrale neiging al dan niet gecombineerd met een spreidingsmaat om de vooraf-kans te bepalen. Identieke vooraf-kansen werden gebruikt voor analyses van twee of meer indextesten voor diagnostiek omtrent dezelfde aandoening. Ongeveer de helft (53,6%) van deze identieke vooraf-kansen viel binnen het bereik van prevalenties geobserveerd in studies over dezelfde aandoening. Verschillende methoden worden gebruikt voor het selecteren van de vooraf-kansen en er lijkt geen consensus te bestaan over welke gegevensbron of methode te gebruiken. Er zijn echter enkele overwegingen waarmee rekening mee gehouden moet worden bij het presenteren van DTA resultaten: 1) Overweeg of genormaliseerde frequenties gepresenteerd dienen te worden, 2) Overweeg de invloed van de gekozen methode voor het selecteren van een vooraf-kans op de genormaliseerde frequenties, en 3) Overweeg het gebruik van identieke vooraf-kansen die vallen binnen het bereik van de geselecteerde studies wanneer er meerdere meta-analyses zijn die zich richten op een aandoening.

De algemene discussie van dit proefschrift (hoofdstuk 9) richt de aandacht op twee onderwerpen: methoden voor het gebruik van expertpanels als referentiestandaard in diagnostisch onderzoek, en methoden voor vroege evaluatie van de impact van voorspellingsmodellen op gezondheidsuitkomsten en kosten voor de gezondheidszorg. Resterende uitdagingen worden beschreven voor beide onderwerpen en suggesties voor toekomstig onderzoek worden gegeven. Deze suggesties omvatten het omgaan met ontbrekende schattingen van de kans op aanwezigheid van de aandoening van interesse door het expert panel, rekening houden met de onzekerheid tussen experts binnen een panel, en het bieden van meer advies over het gebruik en de interpretatie van DAM's voordat een gerandomiseerde impactstudie wordt uitgevoerd. Focus op deze methodologische onderwerpen zou uiteindelijk de efficiëntie en waarde van diagnostisch onderzoek moeten verbeteren en onderzoekverspilling moeten verminderen.

"Always be relentlessly positive"

Sean "Day9" Plott, former pro-gamer

Dankwoord

En dan nu het belangrijkste gedeelte van dit proefschrift: de bedankjes. Het heeft een aantal jaar geduurd, met ups en downs, maar dan is toch dit proefschrift er als resultaat uitgerold. Ondanks dat enkel mijn naam op de kaft staat, had ik dit niet kunnen doen zonder de mensen om mij heen. Dus bij dezen, een dankbetuiging aan jullie.

Geachte Prof. Dr. K.G.M. Moons, beste Carl. Als promovendus loop je soms wel eens een overleg binnen als je met je handen in het haar zit, en het even niet meer zo goed weet. Als dat eens onverhoopt gebeurde, dan wist je tijdens ons overleg er binnen een uur voor te zorgen dat er een positieve draai aan werd gegeven, en ik met frisse energie uit de deur uitliep. Dat is echt een gave. Daarnaast heb je ook altijd aandacht voor de persoon, en ben je niet te beroerd om gewoon te beginnen met de vraag 'hoe gaat het met je?'. Bedankt dat je mijn promotor was, en ik hoop dat onze samenwerking de aankomende jaren alleen maar meer en beter gaat worden.

Geachte Dr. C.A. Naaktgeboren, beste Christiana. Vanaf dag 1 was jij de persoon bij wie ik binnen lopen als ik vragen had of ergens niet uit kwam. Ik heb veel van je geleerd, niet alleen inhoudelijk over het opzetten en uitvoeren van verschillende soorten onderzoek, maar ook de praktische kant van het reilen en zeilen binnen het Julius Centrum. Jij was altijd gene die snel en uitgebreid feedback gaf op mijn werk, en dat heeft dit proefschrift gemaakt tot wat het is geworden. Bedankt voor de afgelopen jaren en ik hoop dat we elkaar in de toekomst nog eens tegenkomen.

Geachte Dr. L. Hooft, beste Lotty (ook wel bekend als Looty). Na het vertrek van Joris was er een plek die vrijkwam binnen mijn promotieteam. Jouw naam viel al snel, en ondanks je toen al overvolle schema, was jij bereid om deze positie in te vullen als mijn copromotor. En wat een waardevolle en leuke toevoeging was dat. We hebben niet alleen goed samengewerkt binnen de projecten die al liepen, maar we zijn eveneens nog een week naar de WHO in Geneve afgereisd. Daar heb ik je leren kennen als een aimabel persoon en trotse moeder van je dochtertje Mikki. Nu met jou als mijn direct leidinggevende kijk ik al met plezier uit naar de aankomende jaren als postdoc verder te gaan in Utrecht.

Geachte Dr. J.A.H. de Groot, beste Joris. Ook jou ben ik zeker niet vergeten. Het eerste jaar van mijn promotietraject heb je als copromotor een blijvende indruk achtergelaten. In positieve zin, want de manier waarop ik in alle rust en kalmte met jou en Christiana mijn eerste projecten kon opzetten, bespreken, en evalueren was een ontzettend fijne kickstart voor mijn promotie. Na ongeveer een jaar kreeg ik een berichtje van je: "ik moet je spreken, heb je straks even tijd?". Dat klonk ernstig, en je vertelde dat je had besloten om het Julius Centrum te verlaten en bij Philips aan de slag te gaan. Ondanks dat ik teleurgesteld was dat je ons ging verlaten, vond ik het ook ontzettend leuk voor dat je deze nieuwe uitdaging aanging. Hopelijk heb je het daar nog steeds naar je zin, en we gaan elkaar ongetwijfeld nog wel eens tegenkomen op congressen of borrels.

Geachte Dr. J.B.R. Reitsma, beste Hans. Formeel geen onderdeel van mijn promotieteam, maar informeel overal bij betrokken. Dat is steevast hoe ik mensen uitleg wat jouw rol is geweest binnen mijn promotie. Je bent bij alle projecten betrokken geweest die in deze thesis terugkomen, en wat ontzettend fijn is het geweest om jou erbij gehad te hebben. Altijd zo hartelijk, open, en vrolijk gestemd. Daarnaast heb ik altijd heel goed met je kunnen sparren, en was en ben je nog steeds een bron van inspiratie voor mij. Ik hoop dat we ook de aankomende jaren nog veelvuldig kunnen blijven samenwerken!

Geachte Prof. Dr. R.J.P. Scholten, beste Rob. Tijdens mijn promotie hebben we elkaar vooral gezien en gesproken tijdens meetings en borrels, en hebben we niet echt projecten samen opgepakt. Maar des te meer heb je mij in de maanden na het inleveren van dit proefschrift in no time geïntroduceerd en wegwijs gemaakt binnen het onderwijs dat we geven in het Julius Centrum. Dank daarvoor, en ik hoop dat we samen het onderwijs de aankomende tijd weer naar een hoger niveau kunnen tillen!

Geachte Prof. Dr. C.J. Kalkman, Prof. Dr. M.M. Roovers, Prof. Dr. C.H. van Gils, Prof. Dr. E. Buskes, Dr. G.W.J. Frederix, beste leden van de beoordelingscommissie. Dank voor de bereidheid van het lezen en beoordelen van mijn proefschrift.

Aan Valentijn en Saskia, mijn roomies van Matthias van Geuns 5.18. We hebben toch een slordige paar jaar bij elkaar op de kamer gezeten, en wat fijn was het om jullie in de buurt te hebben. Niet alleen om te sparren als ik met een probleem zat, of om te ventileren als het een keer tegenzat, maar ook om gewoon een stukje te wandelen en een bakje koffie te halen. @Saskia, bedankt voor je tomeloze enthousiasme en positieve energie. @Valentijn, ik kijk er nu al weer naar uit als je weer terugverhuisd naar het Stratenum, zodat we weer ongegeneerd de nerd uit kunnen hangen.

Loan, Anne-Karien, Pauline, Giske, Chris, Nicole, Carline, Timo, Faas, Suzanne, Romin, en Anna-Maria. Ik heb met elk van jullie op de gezelligste kamer in het Stratenum mogen vertoeven. Van serieuze gesprekken, tot tafelvoetbal (ik sta officieel nog steeds aan kop trouwens), tot boulderen en biertjes drinken, mijn promotietijd is er door jullie alleen maar leuker op geworden. Dank jullie wel!

Het epi-methoden team, dank voor alle wijze woorden. Mede dankzij jullie input heeft dit proefschrift de kwaliteit behaald die het nu heeft. In het bijzonder wil ik Anneke nog bedanken voor de vele (lunch)wandelingen, en hulp en wijze raad rondom het afronden van dit proefschrift.

Promovendi van de JOB, dank voor alle informatie, hulp, en feedback tijdens mijn promotie-tijd. In het bijzonder wil ik nog bedanken Marian, Eveline, & Anouk: wat gaaf om met jullie de Promovenski 2018 te hebben mogen organiseren. Ik heb het hele weekend in Winterberg ontzettend genoten.

Michiel, Nick, René, Sander, Simon, Stef, en Toon, beste heren van de ASV Pap in de Beane. In September 2007 werd de eerste grondslag gelegd voor dit illustere gezelschap, en ruim 12 jaar later mag ik nog steeds genieten van jullie grappen en grollen. Of het nu met hotrods door Düsseldorf racen is, een potje Secret Hitler spelen tijdens ons vriendenweekend, of gewoon biertjes drinken in de Aesulaaf, wij vermaken ons wel. Het blijft mooi om te zien dat we na al deze tijd nog steeds zo goed bevriend zijn, en ik hoop dat er nog vele jaren volgen.

Karst, Rudy, en Jasmijn. Het is toch ongelofelijk hoe wij naar al die jaren nog steeds contact hebben. Zelfs voor een reünie met het team van vroeger bleek

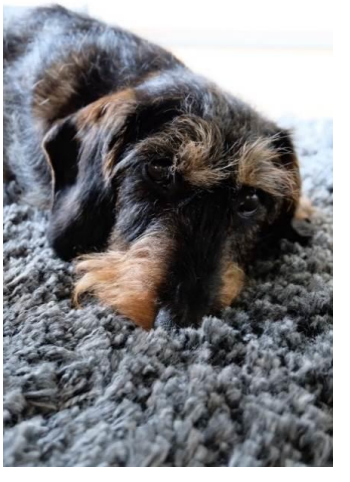
nog veel animo. Ik hoop dat we deze lijn voor de rest van ons leven doortrekken, en ondanks dat de afstand misschien wat groter wordt, we nooit te beroerd zijn om met elkaar een hapje te gaan eten of een biertje/whisky te gaan drinken.

Ruud, Tristan, Roel, en Thijmen, dank voor alle heerlijke borreltjes in de prachtige stad Nijmegen. Dat er nog velen mogen volgen!

Beste familie en kennissen, en in het bijzonder oma, de interesse die jullie in mijn promotietraject door de jaren heen hebben getoond heeft me enthousiast gehouden en me geholpen om telkens weer te reflecteren op 'the bigger picture' rondom mijn onderzoek. Dank!

Theo en Marie-José, lieve pap en mam. Ooit had ik op de middelbare school het geweldige idee om maar iets economisch te gaan doen, waarop jullie zeiden: "Zou je dat wel doen? Is iets met menselijk lichaam niet meer iets voor jou?". Toen kon ik nooit weten dat die beslissing verstrekkende gevolgen zou hebben, en zonder jullie had ik nooit dit geweldige carrièrepad gekozen. Jullie hebben me altijd gesteund en gestimuleerd op alle mogelijke manieren die jullie maar konden, zonder jullie had ik dit niet gekund.

Lieve Daan, in wat een sneltreinvaart hebben we de afgelopen jaren alles doorgemaakt. Af en toe wat downs, ook heel veel ups, maar wat mooi dat ik ze allemaal met jou mee te mogen maken. Je houdt het gek genoeg nog steeds met mij vol, en ik hoop dat daar de rest van ons leven ook geen verandering in gaat komen. Naast jou kreeg ik ook gelijk je zoontje Toine erbij. Dat was wel even wennen, maar eigenlijk weet ik nu al niet meer beter, en ben ik blij dat we het zo goed met elkaar kunnen vinden. Heel veel liefde voor jullie.



En dan blijft er nog maar een over…

Een knuffel voor die lieve, gekke, territoriale, kuilen-gravende, speeltjes vernielende, aanhankelijke, luierende, stronteigenwijze levensgenieter: onze teckel Trijntje.

"You need equality of opportunity

Not equality of outcome"

Jordan Peterson, clinical psychologist

Curriculum Vitae

Kevin Jenniskens was born on the 31st of July, 1988, in Venlo, the Netherlands. He obtained his master's degree in Biomedical Sciences, with specialisations in Health Technology Assessment (HTA) and Toxicology combined with a consultancy profile, from the Radboud University Nijmegen in 2014. After his studies he worked as a junior researcher at the department of Health Evidence of the Radboudumc, and as a consultant at KALCIO healthcare. Kevin developed an interest in methodological and applied research, as well as teaching. He found those three aspects in a PhD position at the Julius Center for Health Sciences and Primary Care for which he applied and got accepted to in late 2015.

From January 2016 onward he worked on his PhD, focusing on evaluation of methodology surrounding diagnostic tests and prediction models. During the next 3.5 years he performed research, which he presented to his peers during various national and international conferences, from Quebec to Barcelona. Alongside his scientific research, Kevin also assisted in teaching general epidemiologic courses to (bio)medical students, as well as advanced courses for the master of Epidemiology. He completed the first two projects for THINC (The Healthcare Innovation Center) after its founding. During his PhD, he also obtained his Basiscursus Regelgeving en Organisatie voor Klinisch onderzoekers (BROK), Basiskwalificatie Onderwijs (BKO), and post-graduate master in Epidemiology.

After obtaining his PhD, Kevin will continue working as a postdoctoral researcher at the Julius Center for Health Sciences and Primary Care, both as a university teacher, as a researcher on topic of measuring impact of prediction models on health outcomes and costs.

"The meaning of life, in my humble opinion,

is maximising QALYs:

Living as long as you can, as happy as you can"

Self-quoted

List of publications

Jenniskens K, Lagerweij GR, Naaktgeboren CA, Hooft L, Moons KGM, Poldervaart JM, Koffijberg H, Reitsma JB. *Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial.* J Clin Epidemiol. 2019 Jul 19;115:106-115.
Available on: https://www.ncbi.nlm.nih.gov/pubmed/31330250

Jenniskens K, Naaktgeboren CA, Reitsma JB, Hooft L, Moons KGM, van Smeden M. *Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: A simulation study*. J Clin Epidemiol. 2019;111:1-10. Epub 2019/03/25.
Available on: https://www.ncbi.nlm.nih.gov/pubmed/30904568

Jenniskens K, de Groot JAH, Reitsma JB, Moons KGM, Hooft L, Naaktgeboren CA. *Overdiagnosis across medical disciplines: a scoping review*. BMJ Open. 2017 Dec 27;7(12):e018448.
Available on: https://www.ncbi.nlm.nih.gov/pubmed/29284720

Jenniskens K, Janssen PA. *Newborn outcomes in British Columbia after caesarean section for nonreassuring fetal status*. J Obstet Gynaecol Can. 2015 Mar;37(3):207-13.
Available on: http://www.ncbi.nlm.nih.gov/pubmed/2601867

Makai P, IntHout J, Deinum J, **Jenniskens K**, Wilt GJV. *A Network Meta-Analysis of Clinical Management Strategies for Treatment-Resistant Hypertension: Making Optimal Use of the Evidence*. Journal of general internal medicine. 2017;32(8):921-30.
Available on: https://www.ncbi.nlm.nih.gov/pubmed/28275946

De Vries HS, te Morsche RH, **Jenniskens K**, Peters WH, de Jong DJ. *A functional polymorphism in UGT1A1 related to hyperbilirubinemia is associated with a decreased risk for Crohn's disease*. J Crohns Colitis. 2012 Jun;6(5):597-602.
Available on: http://www.ncbi.nlm.nih.gov/pubmed/22398043

Kluytmans A, Deinum J, **Jenniskens K**, van Herwaarden AE, Gloerich J, van Gool AJ, van der Wilt GJ, Grutters JP. *Clinical biomarker innovation: when is it worthwhile?* Clinical Chemistry and Laboratory Medicine (CCLM). 2019 Jul 9.
Available on: https://www.ncbi.nlm.nih.gov/pubmed/31287794