Methodologic approaches to evaluating new highly sensitive diagnostic tests: avoiding overdiagnosis

Joris A.H. de Groot PhD, Christiana A. Naaktgeboren PhD, Johannes B. Reitsma MD PhD, Karel G.M. Moons PhD

■ Cite as: *CMAJ* 2017 January 16;189:E64-8. doi: 10.1503/cmaj.150999

A major contributor to the rising problem of overdiagnosis, with the subsequent risk of overtreatment, is the development of highly sensitive diagnostic technologies that challenge and sometimes expand existing disease definitions. Whereas such technologies might help to identify new or milder disease, make diagnoses earlier or identify previously undetected abnormalities, the impact on prognosis and treatment often remains uncertain. It is often unclear what proportion of additional abnormalities found is benign or should prompt no intervention, and what proportion is indeed clinically relevant and worthy of treatment. Therefore, it is unclear how many people might receive a diagnosis and unnecessary treatment as a result of widespread introduction of the new test. As such, new diagnostic technologies may lead to overdiagnosis and overtreatment.

Traditional cross-sectional studies of diagnostic accuracy that only employ the current reference standard are insufficient to evaluate the clinical relevance of many new highly sensitive tests. Appropriate methods for assessing the performance of these tests should be developed to overcome the challenges induced by present-day technologic developments. Improved data analysis and presentation of current studies of diagnostic accuracy, and better use of existing data are required. Test–treatment trials may be needed to measure the effectiveness or impact of new highly sensitive diagnostic tests on improving patient-relevant outcomes, to avoid the harms of overdiagnosis and overtreatment.

KEY POINTS

- The development of ever-more-sensitive diagnostic technologies that challenge existing disease definitions is a major contributor to the rising problem of overdiagnosis and the subsequent risk of overtreatment.
- Standard studies of diagnostic accuracy are insufficient to evaluate the true clinical benefit of many new highly sensitive tests.
- Improved data analysis and presentation of current diagnostic studies as well as better use of existing data are necessary.
- Test–treatment trials may still be needed to determine whether the new test will improve patient-relevant outcomes, or rather induce overdiagnosis and overtreatment.

What is overdiagnosis and how might new diagnostic tests contribute?

Overdiagnosis due to the introduction of new tests occurs when the tests identify abnormalities that are indolent, nonprogressive or regressive, and that, if left untreated, will not cause symptoms or shorten an individual's life.[1] A well-known example of a test detecting ever-smaller abnormalities of uncertain prognosis and therefore leading to overdiagnosis is enhanced mammography for detecting ductal carcinoma in situ.[2–4]

The problem of overdiagnosis, in general, arises and persists due to multiple complex forces such as financial gain, legal concerns and media hype.[5,6] Some have argued that millions of people may be receiving unwarranted and potentially harmful treatments and have advocated for fairer, more rational and less wasteful health care systems.[7,8] A major driver of overdiagnosis and overtreatment is the development of ever-more-sensitive diagnostic tests, such as biomarkers and high-resolution imaging techniques.[9,10] More sensitive tests might identify milder or subclinical disease that may enable patients to benefit from early treatment; these tests can also increase our understanding of diseases and stimulate the search for new and specific treatments. However, it is not always clear whether patients with newly detected abnormalities have the same poor prognosis or similarly respond to usual treatment protocols as patients with established disease. In such cases, a new test could potentially lead to overdiagnosis and overtreatment. For example, spiral computed tomography is capable of detecting very small pulmonary emboli (i.e., subsegmental embolisms), but do these subclinical emboli have the same clinical consequences for the patient as larger emboli, do patients have a similar prognosis and will therapy have similar effects? Or will treatment do more harm than good?

Another example of a test that detects "new abnormalities" is 7 tesla magnetic resonance imaging of the brain.[11] This technology is sensitive to detection of a range of abnormalities, but it is not clear if detected lesions are early, mild or even different patterns of existing diseases. As such, questions about patient prognosis and response to any treatment are again difficult to answer.

What is the problem?

The standard method of ascertaining the accuracy of a new diagnostic test uses a cross-sectional design (Figure 1) in which the new test is compared with the current best reference stan-

dard (i.e., the method or combination of methods that in current practice is the best way to establish whether the target disease is present or absent).[9] This approach is not well suited to evaluate whether patients actually benefit from use of the new test. The new test might well be superior to the prevailing reference standard in a study of diagnostic accuracy, in which case the study may provide too little or even biased information about the true value of the new test.

Furthermore, in standard studies of diagnostic accuracy, disease is assumed to be a dichotomous state. Either the disease is present or it is not, and either you benefit from treatment or you do not (Figure 2A). However, disease is now understood, in most cases, to be rather a spectrum of abnormalities on a more continuous scale. New highly sensitive tests might reveal to us more of the full spectrum of disease, but we are also confronted with a need to decide on a disease threshold. When a new highly sensitive test identifies new (milder or less advanced) cases, the question is, "What clinical action, if any, should be taken on these new cases?" (Figure 2B).[10]

Another assumption of standard studies of diagnostic accuracy is that the reference standard test is the best way to classify disease status. Under this assumption, all new abnormalities detected by the new test that are diagnosed negative (i.e., disease is absent) by the reference standard will be considered false-positive results (Figure 3A). The question then arises of whether to stick to the prevailing disease definition or to lower the threshold for presence of disease (light-shaded rounded squares, Figure 3B).[12] However, it cannot be assumed that all of these new cases benefit from identification in the same way as the cases detected with the prevailing reference standard (dark-shaded squares, Figure 3B). These patients may have milder or earlier stages of disease (Figure 2B), and therefore their signs and

	Disease	No disease
Index test positive	TP	FP
Index test negative	FN	TN

Sensitivity: TP/(TP+FN)
Specificity: TN/(TN+FP)
Positive predictive value: TP/(TP+FP)
Negative predictive value: TN/(TN+FN)

Figure 1: Standard cross-sectional design for determining diagnostic accuracy. Note: TP = true positive, FP = false positive, FN = false negative, TN = true negative.

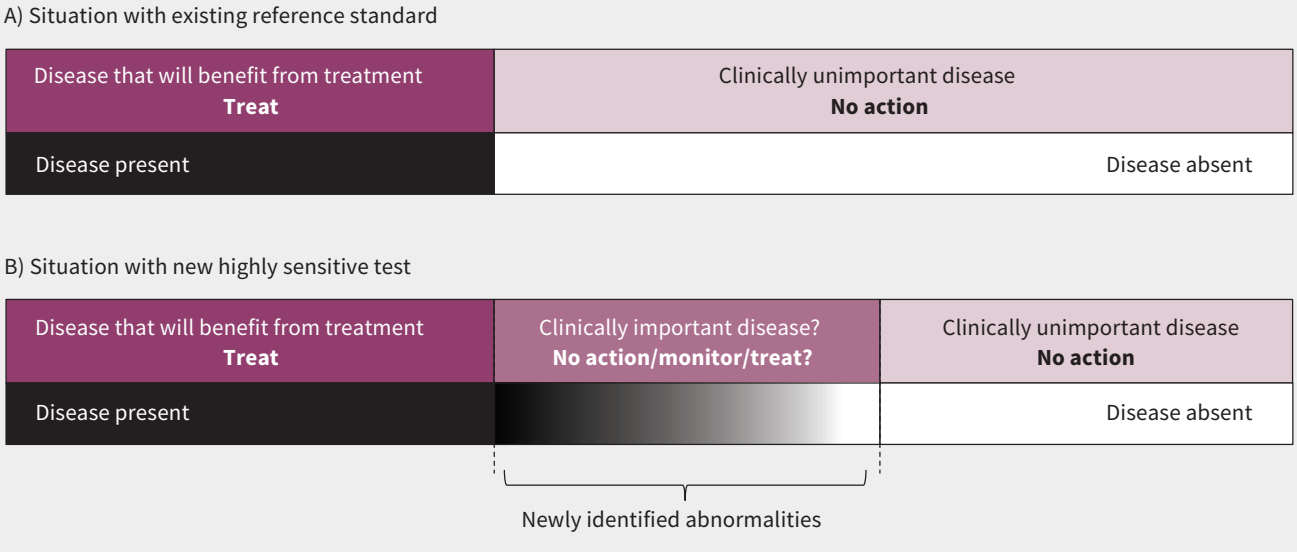

Figure 2: How disease definitions based on results of the existing reference standard and accompanying treatment decisions (A) are challenged by new highly sensitive tests detecting a broader spectrum of disease (B).[10]

symptoms, prognosis and response to treatment may differ from those of patients whose diseases are detected with the prevailing reference standard.

However, before lowering the disease threshold, one should first consider the consequences of such a change by studying the prognosis of these newly identified abnormalities or disease states, whether they would benefit from closer monitoring, and what their outcomes are after being treated. Treating and even monitoring these new cases may be at least unnecessary and wasteful in some cases, and harmful in others (Figure 3C).

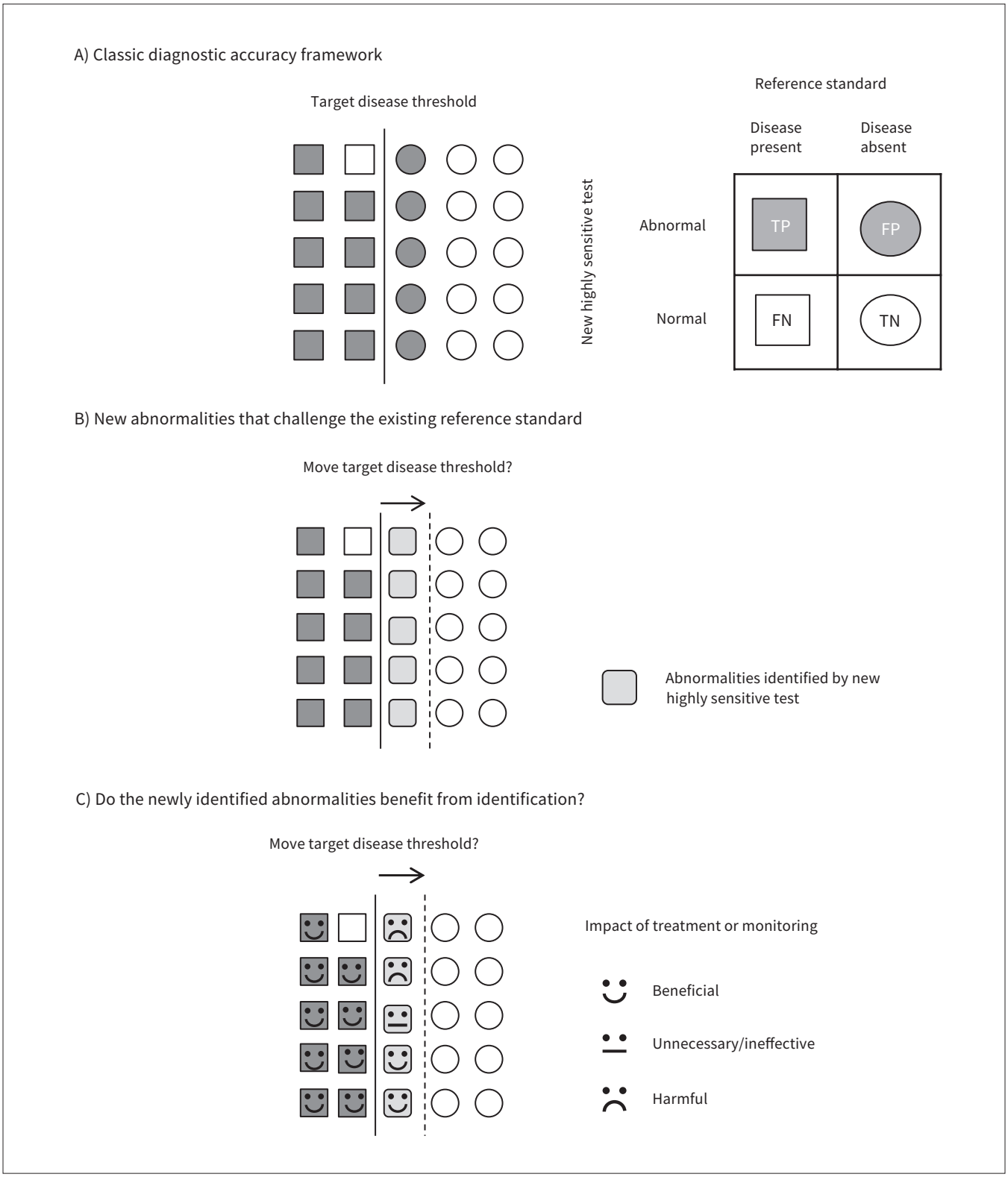

Figure 3: Schematic representation of considering the consequences of allowing a new highly sensitive test to broaden a disease definition.

What is the solution?

We are in urgent need of a more extensive methodologic framework to evaluate whether a new diagnostic test that challenges prevailing disease definitions and reference standards leads to real clinical benefit or rather to overdiagnosis and overtreatment. Allowing advancing technologies to drive the process of lowering disease thresholds instead of first investigating their true clinical relevance would be unwise. Of course, when new diagnostic technologies emerge, there may be a need to carefully re-examine which subgroups of patients need which treatment or whether we should refrain from treatment at all in some patients.

Enhancing traditional cross-sectional studies of diagnostic accuracy

Traditional cross-sectional studies of diagnostic accuracy may be adapted to provide valuable information if analyzed and presented in a different way and by additional use of readily available sources of information. The key concept is to assume that there is no reference standard and to compare the tests on an equal footing by focusing on all the patients for whom the results of the new test and the existing reference standard disagree.[12] By focusing on data from these patients, we are able to examine the clinical consequences of choosing one test or the other as the reference standard. Also in standard cross-sectional studies of diagnostic accuracy, researchers often collect detailed information on relevant signs, symptoms, other concomitant diagnostic test results and even prognostic factors. By use of such data from the patients with discordant results, the new highly sensitive test and the existing reference standard can be compared according to this additional information, and thus it can be determined whether they are different and perhaps reflect different disease stages or entities. It may even be possible to ascertain which test is better at distinguishing the clinically relevant disease stages.[12] We suggest that such information should always be provided by researchers when reporting findings from a cross-sectional study of diagnostic accuracy.

Moreover, information gleaned from such methods can further be used by researchers to determine whether patients comparable to the patients with discordant results in their cross-sectional study are already enrolled in reported prognostic or randomized treatment studies. By establishing the reported prognosis and response to treatment of such patients, diagnostic researchers may actually infer the clinical relevance of the patients with discordant results and thus of diagnosing previously undetected abnormalities or earlier disease.

Since it is not currently necessary to show proven or estimated diagnostic accuracy to be allowed market access in the European CE (conformity marking) approval system, new tests, including highly sensitive tests, are often introduced in clinical practice before formal comparisons of accuracy against the prevailing reference standard are made. A formal cross-sectional study on the diagnostic accuracy of a new test that is already on the market may be enhanced by analysis of routine care data that is already available (e.g., from hospital registries or postmarket surveillance). Such data may include results of the new test actually observed in practice, the relation of these new test results with other diagnostic test results or prognostic factors, administered treatments based on the new test's results, and patient outcomes associated with use of the new test.

Despite use of additional sources of information through the approaches described above, it may still not be possible to estimate accurately the true benefit or impact of a new test and subsequently administered treatments. However, such approaches are invaluable in determining which new tests require further examination by gathering direct evidence on benefits of new highly sensitive tests.

Gathering direct evidence of clinically relevant outcomes: randomized test–treatment trials

Prospective test–treatment trials compare relevant health outcomes of patients randomly assigned to a test–treatment pathway that includes the new test with patients assigned to an existing test–treatment pathway. Such trials are lengthy and expensive, and therefore not routinely undertaken. One of the few examples of a test evaluated extensively in this manner (with respect to both benefits and harms) is B-type natriuretic peptide (BNP) as a diagnostic test to detect heart failure.[13]

Currently, there are various designs of such test–treatment trials.[14,15] In randomize-all designs, all patients meeting the trial eligibility criteria, irrespective of their result on the new test, are randomly allocated to either experimental or control treatment. Thereafter, associations between test status and treatment response are evaluated. A more efficient design is the so-called enrichment design, in which all potentially eligible patients are first tested using the new diagnostic test, and only patients with positive test results are randomly assigned to the experimental or control intervention. Other patients are in principle excluded from further investigation in the study.

In so-called test-strategy designs, patients with positive test results would receive experimental therapy, and all patients with negative test results would get standard care. More subtle modifications to these three designs are possible.

Conclusion

Ongoing technologic advances in medicine will continue to lead to the development of highly sensitive diagnostic tests that challenge existing reference standards and sometimes expand current disease definitions, which may contribute to overdiagnosis and overtreatment.[10] Traditional cross-sectional studies of diagnostic accuracy are insufficient to evaluate new tests that outperform the current reference standard.[14]

There is uncertainty and debate about which is the most valid, efficient and informative design for evaluation of a new highly sensitive test. Although test–treatment trials provide the most direct evidence, they may be inefficient and require long follow-up periods, which means that it is not feasible to perform such studies for every new test entering the market.[16,17] Linked evidence approaches, which extrapolate short-term intermediate outcomes to long-term patient-relevant outcomes, can improve the efficiency of such trials by shortening the follow-up period and selecting the patient subgroups that are most likely

benefitting from the new test.[18] Additionally, focusing on the patients with discordant results from traditional cross-sectional studies on diagnostic accuracy, by using a design in which only these patients are randomly assigned to treatment, can make test–treatment trials more efficient, although they may still require substantial effort and expense.

Therefore, we recommend first carefully examining discordant pairs from the traditional cross-sectional studies of diagnostic accuracy using routinely collected observational data, as outlined in the "What is the solution?" section, to help determine whether a test–treatment trial is necessary. Decision-analytic and cost-effectiveness modelling studies may also be useful gatekeepers or even potential replacements for test–treatment trials.[19,20] These approaches integrate best-available information on test accuracy along with the treatment effectiveness and costs to provide insight into potential short- and long-term effects of tests. In clinical practice we focus on identifying which patients require treatment. It is important to distinguish between the extension and refinement of classifications of a disease for the purpose of clinical research and clinical practice.

References

1. Moynihan R, Henry D, Moons KG. Using evidence to combat overdiagnosis and overtreatment: evaluating treatments, tests, and disease definitions in the time of too much. *PLoS Med* 2014;11:e1001655.
2. Welch HG. Overdiagnosis and mammography screening. *BMJ* 2009;339:b1425.
3. Gøtzsche PC. Ramifications of screening for breast cancer: overdiagnosis in the Malmö trial was considerably underestimated. *BMJ* 2006;332:727.
4. Jørgensen KJ, Gotzsche PC. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *BMJ* 2009;339:b2587.
5. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. *BMJ* 2012;344:e3502.
6. Hoffman JR, Cooper RJ. Overdiagnosis of disease: a modern epidemic. *Arch Intern Med* 2012;172:1123-4.
7. Welch G, Schwartz L, Woloshin S. *Overdiagnosed: making people sick in pursuit of health*. Boston: Beacon Press; 2011.
8. Macdonald H, Loder E. Too much medicine: the challenge of finding common ground. *BMJ* 2015;350:h1163.
9. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the crosssectional study. In: *The evidence base of clinical diagnosis*. London (UK): BMJ Books; 2002:39-60.
10. Lord SJ, Staub LP, Bossuyt PM, et al. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ* 2011;343:d4684.
11. van der Kolk AG, Hendrikse J, Zwanenburg JJ, et al. Clinical applications of 7 T MRI in the brain. *Eur J Radiol* 2013;82:708-18.
12. Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? *Ann Intern Med* 2008;149:816-22.
13. Troughton RW, Frampton CM, Yandle TG, et al. Treatment of heart failure guided by plasma aminoterminal brain natriuretic peptide (N-BNP) concentrations. *Lancet* 2000;355:1126-30.
14. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006;144:850-5.
15. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. *J Clin Epidemiol* 2009;62:364-73.
16. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356:1844-7.
17. Ferrante di Ruffano L, Davenport C, Eisinga A, et al. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *J Clin Epidemiol* 2012;65:282-7.
18. Staub LP, Dyer S, Lord SJ, et al. Linking the evidence: intermediate outcomes in medical test assessments. *Int J Technol Assess Health Care* 2012;28:52-8.
19. Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. *BMC Med Res Methodol* 2013;13:12.
20. Schaafsma JD, van der Graaf Y, Rinkel GJ, et al. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. *J Clin Epidemiol* 2009;62:1248-52.

Competing interests: None declared.

This article has been peer reviewed.

Affiliation: Julius Center for Health Sciences and Primary Care, Utrecht, the Netherlands

Contributors: All of the authors contributed substantially to the conception and design of this article. Joris de Groot and Christiana Naaktgeboren drafted the manuscript, which Johannes Reitsma and Karel Moons revised. All of the authors gave final approval of the version to be published and agreed to act as guarantors of the work.

Funding: The authors gratefully acknowledge the support by The Netherlands Organisation for Scientific Research (projects 9120.8004 and 918.10.615).

Correspondence to: Joris de Groot, j.degroot-17@umcutrecht.nl