Anticipating missing reference standard data when planning diagnostic accuracy studies

Christiana A Naaktgeboren,[1] Joris A H de Groot,[1] Anne W S Rutjes,[2,3] Patrick M M Bossuyt,[4] Johannes B Reitsma,[1] Karel G M Moons[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3584 CG Utrecht, Netherlands

[2]CTU Bern, Department of Clinical Research, University of Bern, Switzerland

[3]Institute of Social and Preventive Medicine, University of Bern, Switzerland

[4]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands

Correspondence to: J B Reitsma
j.b.reitsma-2@umcutrecht.nl

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2016;352:i402
http://dx.doi.org/10.1136/bmj.i402

Accepted: 30 December 2015

Results obtained using a reference standard may be missing for some participants in diagnostic accuracy studies. This paper looks at methods for dealing with such missing data when designing or conducting a prospective diagnostic accuracy study

The problem: missing reference standard data

Diagnostic studies typically evaluate the accuracy of one or more tests, markers, or models by comparing their results with those of, ideally, a "gold" reference test or standard.[1 2] In such studies, the outcome—that is, the presence or absence of the target disease as determined by the chosen reference standard—is often missing in some of the study participants. This is known as partial verification.[3 4] When only the participants who received the reference standard are included in the analysis (complete case analysis), estimates of the accuracy of the diagnostic test(s), marker(s), or model(s) under study, such as the sensitivity, specificity, predictive values, likelihood ratios, or C index, can be biased.[5-8]

There are many reasons why missing reference standard results may occur in diagnostic studies, as well as various approaches to deal with these missing outcomes in the statistical analysis.[3 4 8-16] Ideally, how missing outcome data will eventually be dealt with is determined during the design phase of the study as opposed to later during the data analysis phase.

Here, we build on previous research on methods for dealing with missing outcomes during the data analysis phase to look at specific measures that can be taken when designing or conducting a diagnostic accuracy study. This paper focuses on prospective studies in which all included patients suspected of having the disease of interest receive all tests under study as well as the reference standard. It does not cover alternative designs, such as separate sampling of diseased participants and healthy controls, or retrospective studies in which patients who have received both the index test and reference standard are identified in hospital databases.[17] Firstly, we discuss the various reasons for missing reference standard results, then we consider the proposed solutions to handle patterns of missing data, and we end with an overview of specific measures that can be considered in the design phase of a prospective diagnostic study to improve the proposed solutions.

SUMMARY POINTS

- Missing reference standard results—that is, missing data on the target disease status—are common in diagnostic accuracy studies
- Analyses that include only the study participants for whom the target disease status is actually measured may produce biased estimates of accuracy
- Several statistical methods to reduce this bias are available; however, they all rely on assumptions about the pattern of missing outcomes, which are sometimes unverifiable
- This paper provides an overview of the different patterns of missing data on the reference standard, the recommended corresponding solutions, and the specific measures that can be taken before and during a prospective diagnostic study to enhance the validity and interpretation of these solutions

How the problem arises

In clinical practice, the diagnostic process begins when signs, symptoms, or test results signal a possible target disease. Patients go through a diagnostic pathway, typically starting with inexpensive, non-invasive tests to rule out the presence of the disease.[18] For those in whom the presence of the disease is still suspected, additional tests may follow that are increasingly costly, burdensome, and even risky. For safety and efficiency, not all patients originally suspected of having a disease eventually go on to receive the complete battery of tests.

In prospective diagnostic studies—that is studies that do not use routine care data such as hospital or primary care records—all study participants ideally receive all tests, markers, or models under study (from now on referred to as index tests) and then the reference standard to assign their final diagnosis. Nevertheless, even in predesigned prospective studies, missing outcomes on the reference standard are likely to occur and in some situations may even be unavoidable. These missing outcomes may occur haphazardly, in a more or less predictable way, or even by design (see table 1 for examples). As in any clinical study, haphazardly missing data may result from, for example, lost blood samples, technical failures, or accidental deviations from the study protocol. For example, in a study on the accuracy of rapid diagnostic tests for malaria, a few blood samples were lost before they could be examined under the microscope.[19] We refer to this as "incidental missing data." Although this type of missing data leads to a loss of precision, it does not necessarily lead to biased estimates of test accuracy owing to the complete randomness of the missing outcomes.

Commonly, though, clear reasons exist why some participants in a study do not undergo the reference standard. It may be specified in the protocol of a prospective accuracy study, for instance, that to reduce

Table 1 | Examples of different mechanisms for missing outcomes in diagnostic accuracy studies

Pattern of missing outcome	Target condition	Test(s) under evaluation	Reference standard	Reason reference standard result (diagnostic outcome) is missing in some participants
Incidental missing data[19]	Malaria	Rapid diagnostic test	Microscopy	Blood samples were lost, so data are probably missing completely at random
Data missing by study design[21]	Cervical cancer	Visual inspection with acetic acid	Colposcopy with biopsy	Screening large population is expensive, and reference test is burdensome, so to reduce study costs and burden to patients only a random sample of those with normal screening tests received reference standard
Data missing due to clinical practice[22]	Inflammatory bowel disease	Faecal calprotectin	Endoscopy with biopsy	Endoscopy with biopsy is invasive, so it was applied only to patients at high risk (those with at least one "red flag" symptom)
Data missing due to infeasibility[23]	Breast cancer	Ultrasonography	Biopsy	Biopsy is impossible to perform when no lesion is detected during mammography, so it was only done in participants with abnormal ultrasound results

study costs or burden to patients only a randomly selected subset of patients in a specific subgroup are to be verified by the preferred reference standard. We refer to this pattern as "data missing by study design").[20] For example, in a study on the diagnostic accuracy of visual inspection with acetic acid for detecting cervical cancer, in which the reference standard was colposcopy with biopsy, only a random subset of participants in whom no abnormalities were seen during visual inspection underwent colposcopy with a series of randomly located biopsies.[21]

In many diagnostic studies, the intention is to perform the reference standard in all patients, but for a variety of reasons missing outcomes occur. Typically, this is not a completely random process. Missingness may depend on several factors, such as severity of symptoms and other preceding test results, resulting in complicated patterns of missing outcomes that are also related to the results of index test. We refer to this pattern as "data missing due to clinical practice." Selective missing data are likely to cause biased estimates of accuracy of the index test in a complete case analysis. An example is a study on the diagnostic accuracy of faecal calprotectin for irritable bowel disease; endoscopy combined with biopsy, the invasive reference standard, was limited to patients at high risk, defined as those with at least one predefined red flag symptom.[22]

In some clinical scenarios, it may be technically impossible to perform the reference standard in a well defined subgroup of participants. We refer to this as "data missing due to infeasibility." This is common in cancer screening studies in which the reference standard is invasive. A specific example is a study on the diagnostic accuracy of ultrasonography for detecting breast cancer, in which one could not do a biopsy when no lesion was observed.[23]

How to deal with missing reference standard results

Understanding why missing outcomes occur is necessary for judging whether estimates of diagnostic accuracy are at risk of being biased, as well as whether and how this bias can be corrected for (table 2). In addition to keeping careful track of the reasons for missing reference standards, analytical methods are available to help to distinguish between "incidental missing data" and "data missing due to clinical practice." A method commonly used to identify the risk of bias due to missing data is to compare the distribution of the patients' characteristics and results of the index test(s) among the study participants with and without a missing outcome.[24] If differences exist, the estimates based only on the participants with observed reference standard results (complete case analysis) are assumed to be at risk of bias, as those participants are not a completely random subset of the initial study population. Another method to judge the potential for bias is to do a sensitivity analysis to explore whether the range of values for the accuracy estimates of the index test are consistent with the data. Such a sensitivity analysis quantifies the possible range of sensitivities, specificities, predictive values, or C indices if all participants with a missing outcome were considered as either diseased or non-diseased. A web tool has been developed that plots a so-called test ignorance region (available at uwmsk.org/gsa).[25] If the accuracy of the index test(s) from the complete case analysis falls outside this test ignorance

Table 2 | Analytical approaches to reduce bias in estimated accuracy of diagnostic test(s), marker(s), or model(s) under study, introduced when preferred reference standard is not performed (that is, outcome is missing) in some study participants

Method	Description
Sensitivity analysis[25]	Quantify possible range of accuracy if participants with missing preferred reference standard result were classified as either diseased or non-diseased
Complete case analysis	Include only participants in whom preferred reference standard is performed in analysis
Inverse probability weighting ("Begg and Greenes method")[1126]	Inflates number of participants by multiplying each cell or category (in which not all participants underwent preferred reference standard) by inverse probability of having outcome verified
Multiple imputation[1516]	Multiple complete datasets are created by using available data to predict plausible values for missing outcomes. Analyses are performed on these imputed datasets, and accuracy estimates of diagnostic index test(s), marker(s), or model(s) are pooled
Differential verification[32]	Perform a different (usually less accurate) reference standard in participants in whom preferred reference standard is missing. Subsequently, one may use one or both of following options:
Report results separately by reference standard used[32]	When index test, marker, or model results determine which subsequent test is used to verify outcome and outcome of alternative reference standard is clinically interpretable
Bayesian correction method for differential verification[14]	Such analysis adjusts accuracy estimates of index tests, on basis of assumptions about accuracy of reference standards used and verification pattern

region, the assumption that the data are missing haphazardly (completely at random) is not reasonable, so accuracy estimates are likely to be biased and should therefore be adjusted.

When outcomes are missing haphazardly (the pattern "incidental missing data")—that is, unrelated to any observed or unobserved patients' characteristics or test results—and the study is large enough, a complete case analysis that includes only participants who underwent the reference standard will produce estimates similar to those obtained if all original study participants had been included, except that these accuracy estimates will be less precise. In that case, participants with the outcome can be seen as a completely random sample of the original study group, still representing a random sample from the study population defined by the eligibility criteria.

When outcomes are missing selectively (as is the case for all patterns except "incidental missing data"), a complete case analysis will probably produce biased estimates of accuracy. Analytical approaches for reducing the bias introduced by missing outcomes essentially use the available data to reconstruct the missing outcome (see table 2 for an overview of these methods).[11 14-16 26] These methods either require knowledge of or make assumptions about the pattern of the missing outcomes.

A straightforward correction method was developed by Begg and Greenes, who used inverse probability weighting, a technique also often used in causal research.[11] Their approach can provide unbiased accuracy estimates of the index test(s) when the missingness is actually random given the result of the index test(s). For a dichotomous index test, this method is equivalent to inflating the two-by-two table by multiplying each cell by the inverse probability of having undergone the reference standard. The assumption then is that patients with a negative (or positive) index test result who have not been verified would have shown comparable results to those with a negative (or positive) index test result who were verified. This method can be extended to incorporate additional factors that may have led to the missing outcomes. However, when the mechanism of the missing outcome data is not so straightforward and is based on multiple variables rather than only the index test(s), a more advanced method of reconstructing the data, such as multiple imputation, may be recommended instead.[15 27]

Imputation is the substitution of missing data with plausible values to allow for analysis of the entire dataset. Multiple imputation is a statistical procedure that uses all available patients' data to predict the missing data, in this case the missing outcome.[28] These missing data are predicted multiple times, resulting in several complete datasets, often 10 or more, on which standard analyses are then performed.[29] The accuracy estimates of the index tests from these datasets are then averaged to provide an overall estimate, with adjusted confidence intervals that reflect the uncertainty resulting from the missing data. The more accurately the available data predict the missing outcomes, the less biased and more precise the accuracy estimates after multiple imputation will be. Even if some of the variables that influenced missingness are not available in the data, multiple imputation will probably still result in less biased results of the accuracy of the index tests than will complete case analysis.[30] The challenge to multiple imputation is that it depends on the ability of additional patients' data to accurately predict the missing reference standard results. Other, less straightforward, analytical methods for complex missing patterns exist, for which we refer to an overview of the literature.[12]

Instead of approaching the bias introduced by missing outcomes by using purely analytical correction methods, an alternative approach is to rely on results from a second reference standard to determine the outcome in participants missing the preferred reference standard. The use of different reference standards in different participants is known as differential verification.[3 9 31 32] If the alternative reference standard classifies disease status with less accuracy than does the preferred standard, this approach essentially results in misclassification of the outcome.[33] As such, it may increase, rather than reduce, the bias in the estimated accuracy of the index test(s). When differential verification is present, one might consider using an empirical bayesian correction method that takes into account the verification pattern as well as bias due to imperfections in the reference standards.[14] This model requires specification of the pattern by which participants receive one reference standard or the other. It allows the researchers to incorporate their beliefs about the accuracy of the reference standards with respect to the true disease of interest in the form of previous distributions. Challenges to the bayesian correction method are understanding and specifying a potentially complex verification pattern and the availability of evidence on which to base beliefs about the accuracy of the reference standard. In the particular situation in which the type of reference standard a participant receives is completely dependent on the result of the index test, marker, or model, the predictive values are clinically interpretable. This would happen, for example, if all participants whose (dichotomous) index test result is abnormal receive the preferred reference standard and all others receive an alternative. In that case, one may simply choose to report results stratified by the index test results—that is, predictive values.[32]

Considerations for study design, analysis, reporting, and interpretation of results

Obviously, missing outcomes in diagnostic accuracy studies should ideally be avoided, as in any clinical study. All solutions for correcting bias introduced by missing outcomes are suboptimal. However, we argue that when missing outcomes are anticipated before the start of a diagnostic study, timely actions can be planned to optimise the validity of the study results. The protocols of prospective diagnostic accuracy studies can be enhanced by including information on the expected pattern of missing outcomes, as well as the chosen design and analytical solutions for reducing

the impact of these missing outcomes. In addition to presenting results that have been adjusted for missing outcomes, transparent reporting of the pattern of missing outcomes is important; this can be represented in a flowchart as recommended in the STARD guidelines.[34] Such reporting facilitates readers' judgment of the risk of bias introduced by the missing outcomes and the appropriateness of the analytical solutions used to correct for this bias.

Table 3 contains an overview of the patterns for missing outcomes and the relation of these patterns to possible design, analytical, and reporting considerations. The appendix contains a worked out example for each of these patterns, using the clinical examples in table 1 as inspiration.

Incidental missing data

A small amount of completely random missing data is almost inevitable in any study for reasons unrelated to any patients' characteristics or index test results, such as data entry errors or dropping a blood sample. In an adequately sized study, excluding from the analysis participants for whom the reference standard result is missing completely at random will not bias the results—it will only decrease precision. The percentage of missing outcomes should be reported, as well as the distribution of patients' characteristics and index test results among those without and with missing outcomes, to allow the reader to judge whether they were missing completely at random and their exclusion thus would not lead to bias.[34] Additionally, a sensitivity analysis as described above may provide further insight into the potential impact of the missing outcomes.

Data missing by study design

For efficiency, technical, or ethical reasons, it may be desirable not to perform the reference standard by design in all participants but only in a random sample of, for example, those with "normal" index test results and to adjust for this partial verification in the analysis ("data missing by study design"). This may be an efficient approach in situations in which the prevalence of disease is low—for instance, in screening. Unfortunately, no a priori sample size calculations are available to determine how large such random samples need to be. One must ensure that the random sample that will be verified by the reference standard will contain a sufficient number of participants with and without the target condition.[35] Therefore, researchers choosing such a design should provide a rationale for the number and type of participants who will randomly be verified in specific subgroups.

Data missing due to clinical practice

When the outcome is missing more often in participants with specific characteristics or index test results, such

Table 3 | Anticipating missing results on best available or preferred reference standard (missing outcomes): considerations for design, conduct, analysis, reporting, and interpretation

Characteristic	Incidental missing data	Data missing by research design	Data missing due to clinical practice	Data missing due to infeasibility
Description	Missingness likely to be completely random	Planned verification in only random sample of pre-specified subgroup(s) of patients	Missingness more likely to occur in certain patients	Preferred reference standard not performed in any patient within pre-specified subgroup
Examples of mechanisms	Technical failures; accidental loss of blood samples	Costs or logistics hinder performing reference standard in all patients	Patient/physician's decision not to perform preferred reference standard (for example, in patients with low probability of target disease)	Technically/ethically impossible to perform reference standard in certain patients (for example, histology in patients with normal imaging results)
Proposed analytical solutions*	Complete case analysis may suffice	A: Inverse probability weighting may suffice (reweight patients in random sample on basis of sampling fraction).	A: Multiple imputation (impute missing reference standard result)	Perform alternative reference standard in non-verified patients (differential verification) and report results per reference standard
		B: Multiple imputation of missing reference standard result may also be used if other factors may also have influenced eventual decision to perform reference standard	B: Bayesian correction method for differential verification (perform secondary reference standard in non-verified patients and adjust for its imperfection)	
Design and conduct				
General	Take measures to prevent missing results on preferred reference standard and, if applicable, on any other reference standard used. Document reasons for missing results on preferred and, if applicable, any other reference standard used			
Specific	–	Consider number of patients in subgroup that will be verified	A: Perform additional tests and record additional information to improve imputation.	–
			B: Apply secondary reference standard and obtain and incorporate external data on its imperfection	
Analysis, reporting, and interpretation				
General	Report reasons for missing results on preferred reference standard and, if applicable, on any other reference standard used. Report flow of patients through study according to STARD guidelines flowchart. Consider sensitivity analysis			
Specific	–	Provide rationale for subgroup and number of patients in subgroup that will be verified	–	Provide rationale for chosen alternative reference standard and discuss its clinical meaning. Report accuracy results of index test(s) stratified by type of reference standard used

*See table 2 for details of these analytical approaches.

doi: 10.1136/bmj.i402 | BMJ 2016;352:i402 | the bmj

as those with less severe symptoms or normal index test results, a complete case analysis will probably result in biased estimates of the accuracy of the index test. Whether this is the case can be inferred from a comparison of the distributions in participants with and without missing outcomes. If investigators plan to use an analytical method to correct for this bias, such as inverse probability weighting or multiple imputation, they should take appropriate actions for collecting additional information on study participants, such as signs, symptoms, and perhaps even additional test results, that will improve the performance of these methods. When the pattern by which patients receive one or another reference standard is more complex, as is often the case, multiple imputation is preferable to inverse probability weighting, as it makes accounting for more than one factor easier.

Sometimes a secondary reference standard—that is, a test that provides information about the outcome—is available but is less accurate than the preferred reference standard. Instead of using analytical correction methods to correct for partial verification bias, one can use this secondary reference standard to assess the outcome in participants who did not receive the preferred reference standard. A bayesian correction method can then be used to calculate the proper index test accuracy estimates.[14] Here, it is important to report the assumptions made, such as the accuracy of the secondary reference standard with respect to the preferred reference standard. We stress that all of these methods to correct for bias due to "data missing due to clinical practice" assume that the pattern of missing reference standard results either is known or can be predicted by observed information.

Data missing due to infeasibility

When performing the reference standard in any of the participants in specific subgroups is explicitly decided against or even impossible—for example, no biopsy of the breast in women without any abnormality on mammography (table 1)—some alternative measure of the target disease should be obtained. In the design phase of a study, the decision can be made to use an alternative reference standard in these participants, a common choice being clinical follow-up. Rather than focusing on how well the index test results correspond to the preferred reference standard, it may be more relevant to focus on whether the index test provides information about clinically relevant outcomes. If so, the clinical relevance of this alternative reference standard should be discussed. One should then focus on the accuracy estimates of the index tests across strata of the index test results—that is, presenting predictive values.[3 32]

Considerations for study design

Although we have provided guidance for how to handle missing data on the reference standard, we stress that situations exist in which these approaches to deal with missing reference standard data may not be possible or cannot remove the bias, even when researchers anticipate the missing reference standards before the study starts. Additionally, although unbiased estimates of diagnostic test accuracy help to evaluate potential clinical value, cross sectional accuracy studies do not always provide the information needed when forming a conclusion about whether a test improves the care of patients. Hence, in some situations, it may be necessary to go beyond accuracy studies and opt for alternative designs that focus on estimating or comparing the clinical value of tests in terms of their ability to improve actual outcomes for patients.[36-40] This may often be the case when missing outcomes are unavoidable or the new index test is hypothesised to outperform the reference standard.

Conclusion

Despite efforts to assess the outcome in all participants in a diagnostic accuracy study, missing reference standard results (that is, missing outcomes) are often inevitable and should be anticipated in any prospective diagnostic accuracy study. Analyses that include only the participants in whom the reference standard was performed are likely to produce biased estimates of the accuracy of the index tests. Several analytical solutions for dealing with missing outcomes are available; however, these solutions require knowledge about the pattern of missing data, and they are no substitute for complete data. Researchers should anticipate the mechanisms that generate missing reference standard results before the start of a study, so that measures and actions can explicitly be taken to reduce the potential for biased estimates of the accuracy of the tests, markers, or models under study, as well as to facilitate correction in the analysis phase. In all cases, researchers should include in their study report how missing data on the index test and reference standard were handled, as invited by the STARD reporting guideline.[34]

Contributors: All authors participated in the conception and design of the article, worked on the drafting of the article and revising it critically for important intellectual content, and have approved the final version to be published.

Funding: Netherlands Organization for Scientific Research (project 918.10.615).

Competing interests: All authors have read and understood the BMJ policy on declaration of interests and declare the following interests: none.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA, ed. *The evidence base of clinical diagnosis*.BMJ Books, 2002: 1-18.
2 Grobbee DE, Hoes AW. *Clinical epidemiology*.Jones and Bartlett Publishers, 2009.
3 Rutjes AW. *Sources of bias and variation in diagnostic studies*. Febodruk BV, 2005 (available at http://dare.uva.nl/record/1/242222).
4 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;62:797-806. doi:10.1016/j.jclinepi.2009.02.005.
5 Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202. doi:10.7326/0003-4819-140-3-200402030-00010.

6 Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76. doi:10.1503/cmaj.050090.
7 Ransohoff DF, Muir WA. Diagnostic workup bias in the evaluation of a test. Serum ferritin and hereditary hemochromatosis. *Med Decis Making* 1982;2:139-45. doi:10.1177/0272989X8200200205.
8 Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992;45:581-6. doi:10.1016/0895-4356(92)90129-B.
9 de Groot JA, Bossuyt PM, Reitsma JB, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ* 2011;343:d4770. doi:10.1136/bmj.d4770.
10 Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. *Med Decis Making* 1992;12:22-31. doi:10.1177/0272989X9201200105.
11 Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;39:207-15. doi:10.2307/2530820.
12 Alonzo TA. Verification bias - impact and methods for correction when assessing accuracy of diagnostic tests. *Rev Stat* 2014;12:67-83.
13 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 Technol Assess* 2007;11(50):iii, ix-51.
14 de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. *Epidemiology* 2011;22:234-41. doi:10.1097/EDE.0b013e318207fc5c.
15 de Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, Moons KG. Correcting for partial verification bias: a comparison of methods. *Ann Epidemiol* 2011;21:139-48. doi:10.1016/j.annepidem.2010.10.004.
16 de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. *Stat Med* 2008;27:5880-9. doi:10.1002/sim.3410.
17 Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335-41. doi:10.1373/clinchem.2005.048595.
18 Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. *J Epidemiol Community Health* 2002;56:337-8. doi:10.1136/jech.56.5.337.
19 Ahmed R, Levy EI, Maratina SS, et al. Performance of four HRP-2/pLDH combination rapid diagnostic tests and field microscopy as screening tests for malaria in pregnancy in Indonesia: a cross-sectional study. *Malar J* 2015;14:420. doi:10.1186/s12936-015-0943-5.
20 Katki HA, Li Y, Edelstein DW, Castle PE. Estimating the agreement and diagnostic accuracy of two diagnostic tests when one test is conducted on only a subsample of specimens. *Stat Med* 2012;31:436-48. doi:10.1002/sim.4422.
21 University of Zimbabwe/JHPIEGO Cervical Cancer Project. Visual inspection with acetic acid for cervical-cancer screening: test qualities in a primary-care setting *Lancet* 1999;353:869-73. doi:10.1016/S0140-6736(98)07033-0.
22 Holtman GA, Lisman-van Leeuwen Y, Kollen BJ, et al. Challenges in diagnostic accuracy studies in primary care: the fecal calprotectin example. *BMC Fam Pract* 2013;14:179. doi:10.1186/1471-2296-14-179.
23 Lehman CD, Lee CI, Loving VA, Portillo MS, Peacock S, DeMartini WB. Accuracy and value of breast ultrasound for primary imaging evaluation of symptomatic women 30-39 years of age. *AJR Am J Roentgenol* 2012;199:1169-77. doi:10.2214/AJR.12.8842.
24 Groenwold RH, Moons KG, Vandenbroucke JP. Randomized trials with missing outcome data: how to analyze and what to report. *CMAJ* 2014;186:1153-7. doi:10.1503/cmaj.131353.
25 Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Stat Med* 2003;22:2711-21. doi:10.1002/sim.1517.
26 Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013;22:278-95. doi:10.1177/0962280210395740.
27 Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393. doi:10.1136/bmj.b2393.
28 Little RJA, Rubin DR. *Statistical analysis with missing data.* Wiley, 2002doi:10.1002/9781119013563.
29 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30:377-99. doi:10.1002/sim.4067.
30 Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;59:1092-101. doi:10.1016/j.jclinepi.2006.01.009.
31 Alonzo TA, Brinton JT, Ringham BM, Glueck DH. Bias in estimating accuracy of a binary screening test with differential disease verification. *Stat Med* 2011;30:1852-64. doi:10.1002/sim.4232.
32 Naaktgeboren CA, de Groot JA, van Smeden M, Moons KG, Reitsma JB. Evaluating diagnostic accuracy in the face of multiple reference standards. *Ann Intern Med* 2013;159:195-202. doi:10.7326/0003-4819-159-3-201308060-00009.
33 Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* Lippincott Williams & Wilkins, 2008.
34 Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527. doi:10.1136/bmj.h5527.
35 Cronin AM, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: a simulation study. *BMC Med Res Methodol* 2008;8:75. doi:10.1186/1471-2288-8-75.
36 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. doi:10.7326/0003-4819-144-11-200606060-00011.
37 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92. doi:10.1136/bmj.332.7549.1089.
38 Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? *Ann Intern Med* 2008;149:816-22. doi:10.7326/0003-4819-149-11-200812020-00009.
39 Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. *Ann Epidemiol* 2006;16:540-4. doi:10.1016/j.annepidem.2005.10.004.
40 Lijmer JG, Leeflang M, Bossuyt PM. *Proposals for a phased evaluation of medical tests.* Agency for Healthcare Research and Quality, 2009.

Supplementary figure and tables