RESEARCH METHODS & REPORTING

Verification problems in diagnostic accuracy studies: consequences and solutions

Joris A H de Groot,[1] Patrick M M Bossuyt,[2] Johannes B Reitsma,[2] Anne W S Rutjes,[3] Nandini Dendukuri,[4] Kristel J M Janssen,[1] Karel G M Moons[1]

[1]Julius Center for Health Sciences and Primary care, UMC Utrecht, PO Box 85500, 3508GA Utrecht, Netherlands

[2]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center Amsterdam, 1100 DE Amsterdam, Netherlands

[3]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine-University of Bern, 3012 Bern, Switzerland

[4]Royal Victoria Hospital, Quebec, Canada H3A 1A1

Correspondence to: J A H de Groot j.degroot-17@umcutrecht.nl

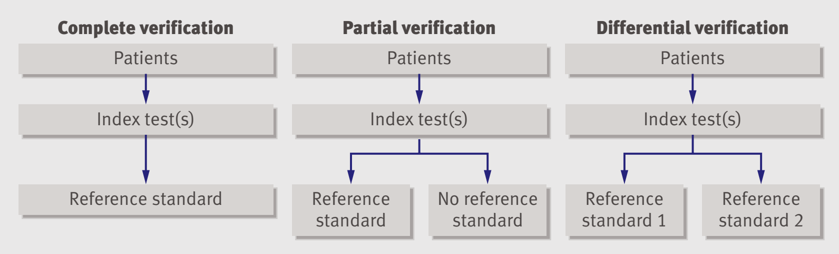
Accepted: 2 May 2011

Cite this as: *BMJ* 2011;343:d4770

doi: 10.1136/bmj.d4770

In diagnostic accuracy studies the ability of a test or combination of tests to correctly identify patients with or without the target condition is verified by applying a reference standard in all patients who have undergone the index test. Incomplete or improper disease verification is one of the major sources of bias in diagnostic accuracy studies.

The accuracy of a diagnostic test or combination of tests (such as in a diagnostic model) is the ability to correctly identify patients with or without the target disease. In studies of diagnostic accuracy, the results of the test or model under study are verified by comparing them with results of a reference standard, applied to the same patients, to verify disease status (see first panel in figure).[1] Measures such as predictive values, post-test probabilities, ROC (receiver operating characteristics) curves, sensitivity, specificity, likelihood ratios, and odds ratios express how well the results of an index test agree with the outcome of the reference standard.[2] Biased and exaggerated estimates of diagnostic accuracy can lead to inefficiencies in diagnostic testing in practice, unnecessary costs, and physicians making incorrect treatment decisions.



Diagnostic accuracy studies with (*a*) complete verification by the same reference standard, (*b*) partial verification, or (*c*) differential verification

SUMMARY POINTS

In studies of diagnostic accuracy studies, ideally all patients undergoing the index test are verified by the reference standard

This is not always possible, and incomplete or improper disease verification is one of the major sources of bias in diagnostic accuracy studies

Partial verification bias occurs when not all patients are verified by the reference standard; instead, disease verification is related to other, previous (index) test results or patient characteristics. Multiple imputation methods can be used to correct for the partial verification bias

An alternative reference test may be used for those cases where verification with the preferred reference test is not possible. This can result in differential verification bias if the results of both reference tests are treated as equal and interchangeable, when they are really of different quality or define the target condition differently. Instead, the estimated accuracy of the diagnostic index test should be reported separately for each reference test

The reference standard ideally provides error-free classification of the disease outcome presence or absence. In some cases, it is not possible to verify the definitive presence or absence of disease in all patients with the (single) reference standard, which may result in bias. In this paper, we describe the most important types of disease verification problems using examples from published diagnostic accuracy studies. We also propose solutions to alleviate the associated biases.

Partial verification

Often not all study subjects who undergo the index test receive the reference standard, leading to missing data on disease outcome (see middle panel in figure). The bias associated with such situations of partial verification is known as partial verification bias, work-up bias, or referral bias.[3-5]

Clinical examples of partial verification

Various mechanisms can lead to partial verification (see examples in table 1).

When the condition of interest produces lesions that need biopsy and subsequent histological verification (as in many cancers), it is impossible to verify negative index test results (“where to biopsy?”). An example is F-18 fluorodeoxyglucose positron emission tomography (FDG-PET) to detect possible distant metastases before planning curative surgery in patients with carcinoma of the oesophagus: only the hotspots detected by PET can be sampled by biopsy and verified histologically.[6]

Ethical reasons can also play a role in withholding a reference standard. Angiography is still considered the best method for detecting pulmonary embolisms, but, because of its invasiveness and risk of serious complications, it is now considered unethical to perform this reference standard in low risk patients, such as those with a low clinical probability and negative D-dimer result.[10]

Sometimes the reference standard may be temporarily unavailable, or patients and doctors may decide to refrain from disease verification. In a study evaluating the accuracy of digital rectal examination and prostate specific

Table 1 | Examples of diagnostic accuracy studies with problems in disease verification

Index test(s)	Target condition	Reference standard	Problem	Study
Partial verification				
Positron emission tomography (PET)	Distant metastases	Histology of biopsy	Only PET hotspots were (can be) biopsied	Lee 2001[6]
Digital rectal examination and prostate specific antigen	Prostate cancer	Combination of transrectal ultrasound plus biopsy	54/145 men not verified for unknown reasons	Pode 1995[7]
Dobutamine-atropine stress echocardiography	Coronary artery disease	Coronary angiography	Only a small sample of patients verified because of clinicians' decision	Elhendy 1998[8]
Hepatic scintigraphy	Liver cancer	Liver biopsy with pathology	39% of index test positives and 63% of test negatives not verified for unspecified reasons	Kline 2001[9]
D-dimer and alveolar dead space measurement	Pulmonary embolism	Pulmonary angiography	Not all patients verified for unspecified reasons	Drum 1972[10]
Differential verification				
Elbow extension test	Elbow fracture	Radiography or follow-up	Index test positives received radiography, index test negatives received follow-up	Appelboam 2008[11]
D-dimer test	Deep vein thrombosis (DVT)	Ultrasonography of the legs	Patients with negative D-dimer test or clinically low risk of DVT were verified by follow-up at 3 months	Buller 2009[12]
Patient history, physical examination, and laboratory tests	Serious bacterial infection	Cultures of blood, spinal fluid, urine, stools, or a panel diagnosis	Mixture of reference standards, as used in clinical practice	Bleeker 2001[13]
Ventilation/perfusion lung scans	Acute pulmonary embolism	Scintigraphy, pulmonary angiography, or follow-up	Mixture of reference standards, as used in clinical practice	PIOPED Investigators 1990[14]

antigen (PSA) for the early detection of prostate cancer, 145 out of 1000 men fulfilled the criterion for verification by the reference standard (transrectal ultrasound combined with biopsy). However, 54 of these men did not undergo the reference standard, for unknown reasons.[7] In another study the accuracy of dobutamine-atropine stress echocardiography for the diagnosis of coronary artery disease was assessed, with coronary angiography as the reference.[8] Only a small proportion of patients received this reference standard because the clinicians' decision to refer to angiography depended on the patient's history and test results.

Potential for bias

The above examples show that partial disease verification, and thus missing disease outcome status in some of the patients, is often not completely at random or non-selective. It is usually based on results of the index test under study or other observed patient variables or test results. If so, the missing outcome status is selectively missing, as the reason for disease verification is associated with other information. For example, patients with a positive index test result or with a high clinical suspicion based on other variables (that is, high probability before the index test) are often more likely to be verified by the reference test than patients with negative test results or a low probability before the index test. Simply leaving such selectively unverified patients out of the analysis will leave a non-random (selective) part of the original group for analysis and thus generate biased estimates of the accuracy of the index test under study.

The direction and size of this bias will depend on how selective the reason for non-verification is, the number of patients whose results are not verified, and the ratio between the number of patients with positive and negative index test results that remain unverified.[5] The bias always occurs in the estimates of the sensitivity and specificity of the diagnostic index test or model under study, and often also in the predictive value. When the reason for partially missing outcomes is based only on the results of the index test, the predictive values of this index test will indeed be unbiased (see below). If, however, the reason for referral for reference testing is not only due to the index test results but also to other patient information, the predictive values of the index test will be affected.[15]

Corrections for partial verification bias

One of the early methods to correct for partial verification bias was developed by Begg and Greenes.[16] Briefly, this method uses only the pattern of disease and non-disease verified by the reference standard among the patients with a positive or negative result of the (single) index test under study. This pattern is then used to calculate the expected number of diseased and non-diseased among the non-verified patients with a positive or negative index test result to obtain an inflated 2×2 table as if all patients were verified by the reference standard. This correction method assumes that the reason for referral to the reference test is only due to the result of the index test under study. Hence, conditional on these index test results, the decision to verify is in fact a random process. The method can also be extended to more than one test result, but this requires exact knowledge of the reasons and patterns behind the partial verification.[16 17]

More recently, multiple imputation methods have been proposed to correct for partial verification problems.[18 19] Multiple imputation can be viewed as a "statistical" workout of the intuitive "diagnostic reasoning" of the clinician. Just as a clinician in practice decides whether to refer a patient for disease verification by a (more invasive, burdensome, or costly) reference standard based on all available patient information, multiple imputation techniques also use all available information of a patient—and that of similar patients—to estimate the most likely value of the missing reference test result in non-verified patients.

Imputation methods comprise two phases—an imputation phase where each missing reference test result is estimated and imputed from all available patient information, and an analysis phase where accuracy estimates of the diagnostic index test or model are computed by standard procedures based on the now completed dataset. Several imputation variants are available, ranging from single imputation of missing reference test values to multiple imputation.[20 21] Instead of filling in a single value for each missing value, as with single imputation, multiple imputation procedures replace each missing value with a set of plausible values to represent the uncertainty about the imputed value. These multiple imputed datasets are then analysed, one by one, again by standard procedures. The results from these analyses are combined to produce accuracy estimates of the diagnostic index test(s) or model and confidence intervals that properly reflect the uncertainty due to missing values.[20 21]

For optimal application of multiple imputation techniques to address partial verification, it is important for researchers to collect as much detailed data as possible on study subjects that could potentially drive the (selective) referral for reference testing. The performance of the multiple imputation or other correction methods will improve with more and better information that may be involved in disease verification decisions. The flexibility of the multiple imputation method enables the incorporation of multiple pieces of observed patient information, not only the results of the index test under study, thereby increasing the likelihood of correctly imputing missing reference test values in patients in whom the disease status was selectively not verified.[17-19]

The discussed mathematical methods to correct for selectively missing verification, and thus partial verification bias, make use of observed (patient) information or variables. They assume that the reasons for missing verification depend on the observed information only. Clearly, this assumption cannot be tested with the data at hand, since non-observed information is, by definition, not available. If one expects selectively missing reference test results as a result of unobserved information, there are methods to perform additional (sensitivity) analysis to quantify to what extent the diagnostic accuracy estimates of the index test change under these situations.[22 23]

Differential verification

Another common approach in diagnostic accuracy studies is to use an alternative, second best, reference test in those subjects for whom the first, preferred reference test cannot or will not be used (see third panel in figure). Although this seems a clinically appealing and ethical approach, bias arises when the results of the two reference tests are treated as interchangeable. Both reference tests are, almost by definition, of different quality in terms of classification of the target disease or may even define the target disease differently.[24 25] Hence, simply combining all disease outcome data in a single analysis (table 2), as if both reference tests are yielding the same disease outcomes, does not reflect the "true" pattern of disease presence and absence. Such an estimation of disease prevalence differs from what one would have obtained if all subjects had undergone the preferred reference standard. Consequently, all estimated measures of the accuracy of the diagnostic index test or model will be biased. This is called differential verification bias.[3 4]

When evaluating a new marker for acute appendicitis, histopathology of the appendix is the preferred reference test, but clinical follow-up is sometimes used as an alternative (for example, if histopathology is considered too invasive).

Table 2 | Effect of differential disease verification in a diagnostic accuracy study. If the preferred reference test (R) is used only to verify positive index test results while an alternative reference test (S) is used to validate index test negatives, simply combining the results ignores the fact that both reference tests have different abilities to determine disease presence or absence, and so the disease status is ambiguously defined

	Reference test results							
	Verification with test R			Verification with test S			Differential verification with either	
Index test result	+ve	−ve		+ve	−ve		+ve?	−ve?
+ve	a	b	+	—	—	≠	a	b
−ve	—	—		c	d		c	d

Compared with histopathology, clinical follow-up is likely to have a higher implicit threshold to detect appendicitis, so it will label more patients as non-diseased (that is, no appendicitis). Thus, these two reference tests define the target condition in a different way. Histopathology might seem the preferred reference test because it reveals even the smallest number of inflamed cells, but one could argue that the more relevant information for clinical practice is not whether the patient has inflamed cells but whether the patient will recover without intervention. This would make clinical follow-up the preferred reference, even though it would be unethical to adopt for all subjects and to withhold surgery. This does mean that accuracy estimates from a combination of histopathology and follow-up will differ systematically from what one would have obtained if all index test results had been verified by either clinical follow-up or histology.

Because accuracy estimates of the new index test ignore the use of different reference tests, they are also difficult to interpret. In situations of differential verifications such as this, the results should be corrected and reported separately for each reference standard to provide informative and unbiased measures of accuracy of the diagnostic index test or model. We illustrate this with a clinical example from the recent literature.

Clinical example

In a recent study[11] the elbow extension test (EET) was examined for its accuracy in ruling out elbow fractures. The preferred reference test was radiography. For unstated reasons (costs, efficiency, or minimising radiation exposure), radiography was planned in patients with a positive EET result whereas the patients with a negative EET received a structured follow-up assessment by telephone after 7–10 days to verify whether elbow fracture was absent (the alternative reference test). Only patients who met any of the pre-specified recall criteria were asked to return to the emergency department for radiography. The rest were considered not to have a clinically significant elbow fracture. The resulting data are shown in table 3.

The authors reported overall estimates of accuracy of the EET, ignoring the use of different reference standards (table 4, first row). Though both radiography and structured follow-up are useful verification methods, their results are not necessarily interchangeable.

The availability of 181 patients with a negative EET who were, after all, evaluated by radiography ("protocol violations" in table 3) enables us to apply the above mentioned

Table 3 | **Distribution of patients in study of diagnostic accuracy of elbow extension test (EET) verified against radiography (for positive EET) or clinical follow-up (for negative EET)***

	Radiography		Follow-up	
	Fracture	No fracture	Fracture	No fracture
Positive EET	521	617	NA	NA
Negative EET	14†	167†	3	414

*Data from Appelboam et al, 2008.[11]

†Data available as a result of protocol violations.

correction methods for partial verification, under the assumption that, conditional on the index test result, the decision to verify is a random process.

The corrected values of sensitivity and specificity clearly show the consequences of differential verification (table 4, second row). We found differences in the estimates of EET accuracy when verification bias is simply ignored and when it is adjusted for. The negative predictive value (the item of primary interest, to rule out elbow fractures), with respect to radiography alone was lower than the value reported by the authors and fell below the desired value of ≥97%. This clearly shows that two reference tests should not be viewed as one.

(For a more detailed discussion of this example and the possibilities to correct for differential verification, see de Groot et al, 2011.[26])

Further corrections for differential verification bias

Recently, a Bayesian method was proposed for simultaneously adjusting for differential verification bias and for the fact that these multiple reference tests were imperfect.[26] The method produces accuracy measures both with respect to the latent disease status and with respect to the use of different reference tests. The former can be considered as a more general measure of performance of the index test with respect to a theoretically defined target condition or disease status since none of the reference tests used is considered perfect. However, the index test's accuracy measures for each of the reference standards may be considered of greater clinical relevance, as these reflect the accuracy against the reference tests that are commonly also performed in daily practice, and on which patient management decisions will often be based.

Conclusion

In diagnostic accuracy studies, all efforts should be made to verify as many test results as possible, preferably all, with the optimal reference test to avoid bias. In practice, the burden on patients, costs, or other reasons often prevent this from happening (table 1).[27]

If test outcome is verified by the reference test for only some of the patients, which is usually selective disease verification based on other observed patient information, we advise the use of the mathematical correction methods described above.[16 17 19]

Table 4 | **Sensitivity, specificity, and predictive values of elbow extension test* depending on whether correction was made for differential verification. All values are percentages (95% confidence intervals)**

Analysis	Sensitivity	Specificity	Negative predictive value	Positive predictive value
No correction	96.8 (95.0 to 98.2)	48.5 (45.6 to 51.4)	97.2 (95.5 to 98.3)	45.8 (42.9 to 48.7)
Correction†	91.8 (88.0 to 95.7)	47.2 (44.2 to 50.2)	92.3 (88.6 to 95.9)	45.8 (42.6 to 49.0)

*Data from Appelboam et al, 2008.[11]

†Corrected for partial verification (accuracy with respect to radiography) by method of Begg and Greenes, 1983.[16]

There is insufficient knowledge to make general statements about what proportion of missing reference standard results might be acceptable and at which point correction methods will become unreliable. Following various statistical guidelines,[18-21 28 29] we recommend the use of correction methods even with small rates of missing verification data. Even small proportions of missing outcomes may yield biased accuracy estimates of the index test(s) or model under study if the non-verified sample is highly selective.

What upper limits of missing reference test data can still be corrected for is even harder to say.[4] Recently Janssen et al showed that, even for large amounts of missing data, imputation leads to less biased results than simply ignoring the (selectively) non-measured subjects.[28] The authors warn that this possibility for imputation depends on how selective or different the observed and non-observed subjects are and how many results remain to build "good enough" imputation models. In any case, authors applying correction or imputation methods for addressing partial verification should provide insight in both issues—how many subjects had missing reference test values and how different were the verified and non-verified patients by comparing both groups on their observed characteristics.[29 30]

If the preferred reference test is not possible and thus missing in complete subgroups, applying a different, usually inferior, reference test will obviously produce different information about the disease status. In such cases, the results should be reported separately for each reference test to provide more clinically informative and unbiased measures of diagnostic accuracy.[3] If in these situations one still wants to quantify the accuracy of the diagnostic index test or model with regard to the same underlying target condition, one should also correct for possible imperfections of the applied reference tests.[26]

Contributors: KGMM is guarantor for the article and heads a research team aimed at improving methods for quantification of the diagnostic and prognostic value of medical tests, biomarkers, and other devices. KJMJ and JAHdG are clinical epidemiologists in his team. PMMB and JBR lead the Biomarker and Test Evaluation (BiTE) programme to develop and appraise methods for evaluating medical tests and biomarkers, and spearheaded the STARD initiative to improve the reporting of diagnostic accuracy studies. AWSR's PhD thesis was on sources of bias and variation in diagnostic accuracy studies and she currently works to update QUADAS, a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. ND is engaged in research in the area of methods for diagnostic studies.

Funding: We acknowledge the support of the Netherlands Organization for Scientific Research (projects 9120.8004 and 918.10.615).

Competing interests: All authors have completed the ICJME unified disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that they have no financial or non-financial interests that may be relevant to the submitted work.

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

1 Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, ed. *The evidence base of clinical diagnosis*. 2nd ed. BMJ Books, 2002:39-60.

2 Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA, ed. *The evidence base of clinical diagnosis*. 2nd ed. BMJ Books, 2002:1-18.

3 Rutjes AW, Reitsma JB, Irwig LM, Bossuyt PM. Sources of bias and variation in diagnostic accuracy studies. In: Rutjes AWS, ed. *Partial and differential verification in diagnostic accuracy studies*. Rutjes, 2005:31-44.

bmj.com Previous articles in this series

- How to obtain the confidence interval from a P value (*BMJ* 2011;343:d2090)
- Verification problems in diagnostic accuracy studies: consequences and solutions (*BMJ* 2011;343:d4770)
- Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials (*BMJ* 2011;343:d4002)
- Interpreting and reporting clinical trials with results of borderline significance (*BMJ* 2011;343:d3340)
- Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting (*BMJ* 2011;342:d1548)

4 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:iii, ix-51.

5 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.

6 Lee J, Aronchick JM, Alavi A. Accuracy of F-18 fluorodeoxyglucose positron emission tomography for the evaluation of malignancy in patients presenting with new lung abnormalities: a retrospective review. *Chest* 2001;120:1791-7.

7 Pode D, Shapiro A, Lebensart P, Meretyk S, Katz G, Barak V. Screening for prostate cancer. *Isr J Med Sci* 1995;31:125-8.

8 Elhendy A, van Domburg RT, Poldermans D, Bax JJ, Nierop PR, Geleijnse ML, et al. Safety and feasibility of dobutamine-atropine stress echocardiography for the diagnosis of coronary artery disease in diabetic patients unable to perform an exercise stress test. *Diabetes Care* 1998;21:1797-802.

9 Drum DE, Christacopoulos JS. Hepatic scintigraphy in clinical decision making. *J Nucl Med* 1972;13:908-15.

10 Kline JA, Israel EG, Michelson EA, O'Neil BJ, Plewa MC, Portelli DC. Diagnostic accuracy of a bedside D-dimer assay and alveolar dead-space measurement for rapid exclusion of pulmonary embolism: a multicenter study. *JAMA* 2001;285:761-8.

11 Appelboam A, Reuben AD, Benger JR, Beech F, Dutson J, Haig S, et al. Elbow extension test to rule out elbow fracture: multicentre, prospective validation and observational study of diagnostic accuracy in adults and children. *BMJ* 2008;337:a2428.

12 Buller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KG, Oudega R, et al. Safely ruling out deep venous thrombosis in primary care. *Ann Intern Med* 2009;150:229-35.

13 Bleeker SE, Moons KG, rksen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. *Acta Paediatr* 2001;90:1226-32.

14 The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA* 1990;263:2753-9.

15 Little RA, Rubin DB. *Statistical analysis with missing data*. Wiley, 1987.

16 Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;39:207-15.

17 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.

18 Harel O, Zhou XH. Multiple imputation for correcting verification bias. *Stat Med* 2006;25:3769-86.

19 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.

20 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087-91.

21 Rubin DB. *Multiple imputation for non response in surveys*. Wiley, 1987.

22 Kosinski AS, Barnhart HX. Accounting for non-ignorable verification bias in assessment of diagnostic tests. *Biometrics* 2003;59:163-71.

23 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.

24 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.

25 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.

26 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.

27 Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. *J Clin Epidemiol* 2003;56:501-6.

28 Janssen KJ, Donders AR, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol* 2010;63:721-7.

29 Mackinnon A. The use and reporting of multiple imputation in medical research—a review. *J Intern Med* 2010;268:586-93.

30 Van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* 2006;59:1102-9.

Learning from failure

Last year I had the humbling privilege of experiencing a year's fellowship at the Institute for Healthcare Improvement (IHI) in Boston in the United States.

Early in my fellowship year I received a visit from a long time medical friend, an internationally renowned rheumatologist from the UK. The same day I had attended an IHI course on executive leadership. I had heard a speaker, by coincidence also a rheumatologist and former hospital chief executive, recount how standardisation of methotrexate initiation has the potential to improve safety and efficiency. Full of enthusiasm, I repeated this to my visitor over dinner anticipating a positive reception.

"Why would I want to deliver cookbook medicine?" came the response. The remainder of the evening was polite, but the fun had drained away. I had failed to convey the concept effectively.

Where had I gone wrong? The next day I reflected on the events and realised that I had omitted some key elements.

The IHI lecturer had actually started his story by recounting how he had been sitting in clinic when a nurse came in and asked, "Why do you all do it differently? It makes it so much harder for us."

"What do you mean?" he asked.

"All six of you initiate methotrexate differently. The receptionist has to have a list because you all have different follow-up intervals. The phlebotomist has lists for frequency and order sets because you all want different blood tests at varying intervals. The pharmacist has lists of all your different starting doses and titration rates. It's so complicated. We have to recheck everything so carefully because we're frightened we'll make a mistake."

"I hadn't realised," he replied. "I'll ask the others about it at our next staff meeting."

On opening the discussion, he found all his colleagues were confident that they were delivering good care but shared his surprise at the variation. One offered to look into the evidence and report back to the group. At the next meeting she reported that all their regimens were acceptable based on current evidence. They discussed the matter and agreed to try a single standard regimen with the understanding that they all retained complete clinical freedom to vary from the standard if they felt it necessary.

A few weeks later, the same nurse entered the clinic room between patients. She simply said, "Thank you."

"No, thank you," he replied. "You've improved my working life. It's so much easier writing 'Methotrexate standard initiation regime' than having to fill out all those forms. Why didn't we do this years ago?"

Clinician engagement with quality improvement is a challenge. Successful improvement is more likely if two conditions are met—firstly, when clinicians want to change rather than being told to change, and, secondly, when clinicians are given the freedom to vary from agreed standards if they feel it would benefit their individual patient. In these circumstances capturing the reason for variation is invaluable information for informing regular update of standards.

My failure to convey the potential benefit of standardisation to my friend was among the most valuable moments of my entire fellowship—failure is a powerful learning opportunity.

Tom Downes consultant physician and geriatrician, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK Tom.Downes@sth.nhs.uk

Cite this as: *BMJ* 2011;343:d5152