CCTC Clinical Chemistry Trainee Council | AACC Better health through laboratory medicine.

PEARLS OF LABORATORY MEDICINE

www.traineecouncil.org

TITLE: Optimal Reporting of Diagnostic Accuracy Studies

PRESENTER: Daniël A. Korevaar

Slide 1:
Hello, my name is Daniël Korevaar. I am a fourth year PhD candidate at the Department of Clinical Epidemiology, Biostatistics, and Bioinformatics at the Academic Medical Center, University of Amsterdam in the Netherlands. Welcome to this Pearl of Laboratory Medicine on "Optimal Reporting of Diagnostic Accuracy Studies".

Slide 2: Diagnostic Accuracy Studies: Design
A large majority of decisions made in clinical practice depend on the results of medical testing. Medical tests can be considered any method for obtaining information about the health status of a patient, and can be used for diagnosis, screening, staging, monitoring, surveillance, prediction, or prognosis. Laboratory tests comprise a fair proportion of the available medical tests.

Diagnostic accuracy studies evaluate the ability of medical tests to correctly classify patients as having or not having a specific target condition. This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future.

Figure A on the left shows the typical design of a diagnostic accuracy study. A series of study participants suspected of having a specific target condition are identified through a single set of eligibility criteria. First, they all undergo an index test; this is the test whose accuracy is under evaluation. Then, they all undergo a reference standard; this is the best available method for establishing the presence or absence of the target condition. Sometimes a reference standard is referred to as the 'gold standard.' The results of the index tests and reference standard are then cross-classified.

As shown in Figure B on the right, some diagnostic accuracy studies are case-control studies, which means that multiple sets of eligibility criteria are used for the identification of study participants: patients in which the presence of the target condition has already been established are included as cases, along with a group of controls that do not have the target condition. These can be healthy controls, or patients with other conditions. Both cases and controls then undergo the index test.

Slide 3: Diagnostic Accuracy Studies: Results

If the results of the index test are categorized as either positive or negative, these can be cross-classified with the results of the reference standard in a 2x2 table, as exemplified in Figure A on the left. Discrepancies are considered to arise from index test misclassifications: false positive index test results in cell B, and false negative index test results in cell C. From this table, diagnostic accuracy measures can be calculated. Sensitivity, for example, is the proportion of patients with the target condition that have a positive index test; specificity is the proportion of patients without the target condition that have a negative index test.

In some cases, the index test can take more than two values; for example, if the results are continuous, as is the case for many laboratory tests. Authors then often report a receiver operating characteristic curve, or ROC curve, which illustrates the combination of sensitivity and specificity for each possible test positivity cut-off. An example of an ROC curve is shown on the right in Figure B.

Slide 4: Diagnostic Accuracy Study: Example

This is an example of a diagnostic accuracy study published in *Clinical Chemistry*. The patient cohort consisted of “patients suspected of having acute hemispheric stroke.” The index test was “plasma glial fibrillary acidic protein.” The reference standard was the “final diagnosis at hospital discharge [..] on the basis of all available clinical data, brain imaging, laboratory testing, and other examinations.”

Because the index test produces a continuous outcome, the authors reported an ROC curve, shown on the right. They also report the clinical sensitivity and specificity at a positivity cut-off of 0.29 µg/L; I have added this point to the figure.

Slide 5: Interpreting Diagnostic Accuracy Studies

This may seem simple, but interpreting results of diagnostic accuracy studies can be challenging. First of all, many types of bias can occur which usually leads to overestimation about the accuracy of the tests. A source of bias can, for example, include study design. Case-control studies have been shown to produce higher accuracy estimates than can be achieved in clinical practice. The reason for this is that the contrast between patients with and without the target condition in these studies is frequently at the extreme ends compared to routine clinical practice. A test will obviously have less difficulty and greater accuracy in differentiating, for example, between patients with severe Alzheimer’s disease and healthy adolescents, compared to detecting Alzheimer’s disease in a group of patients that are all suspected of having this target condition.

Another source of bias can occur when the readers of the reference standard are not blinded to the results of the index test, or vice versa. Particularly in cases where the index test or reference standard contains a level of subjectivity, such as in many imaging tests, knowledge of the results of the other test will usually lead to an increased agreement and, therefore, inflate accuracy estimates. This is called test review bias. There are many other sources of potential bias that cannot all be discussed here.

In addition to these sources of bias, there are sources of variation in diagnostic accuracy. The accuracy of a test is not a fixed property, but tends to vary depending on the clinical context and setting in which the test is applied, on patient characteristics such as disease severity and co-morbidity, on disease prevalence, on previous testing, and on how the index test is performed and interpreted. So, in contrast to what many physicians believe, the same test can have varying sensitivity and specificity if it is used in different clinical settings.

Slide 6: Reporting Diagnostic Accuracy Studies

Individuals who read diagnostic accuracy studies are only able to assess the validity and applicability of study findings if these potential sources of bias and variation are reported. These individuals can include other researchers, clinicians, or patients, but also systematic reviewers or guideline developers that try to summarize the available evidence and make recommendations for specific clinical settings based on this information.

Unfortunately, numerous evaluations have shown that reporting of diagnostic accuracy studies is often insufficiently informative. Suboptimal reporting is now generally considered one of the major sources of waste in biomedical research, because it reduces the value and usefulness of scientific output.

To improve the quality of reporting of diagnostic accuracy studies, the STARD statement was developed. STARD stands for STAndards for Reporting Diagnostic accuracy. STARD was first published in 2003 and an update was recently developed by 85 experts in the field of diagnostic testing and published in 2015 in *Clinical Chemistry*.

Slide 7: STARD 2015 Update

STARD 2015 provides a list of 30 essential items that should be reported to make sure that a study report is sufficiently informative. The list of items is a one-page document that follows the general structure of a research paper; therefore, critical items are provided for the Title, Abstract, Introduction, Methods, Results, and Discussion sections. Although essential, most items on this list are not difficult to incorporate in a study report or an appendix. Some of the items apply to any medical research paper, while others specifically apply to diagnostic accuracy studies.

I will now go through some of the items, explain why they are essential to report, and give examples of complete reporting from previously published reports of diagnostic accuracy studies in *Clinical Chemistry*. An Explanation and Elaboration document that provides a detailed description and rationale for each item will soon be available. All STARD documents can be found at http://www.equator-network.org/.

Slide 8: STARD Item #1: Title or Abstract

Item 1 is “Identification as a study of diagnostic accuracy using at least one measure of accuracy” in the Title or Abstract.

It is often difficult to find diagnostic accuracy studies in research repositories such as PubMed. By referring to measures of diagnostic accuracy in the Title or Abstract, this will make searching for relevant studies much easier because search strategies can then focus on such measures.

In the example, Mak and colleagues correctly report the terms “diagnostic accuracy,” “sensitivity,” and “specificity” in their Title. Their study can now be easily retrieved in PubMed, by searching for one of these terms in combination with the term “ceruloplasmin,” which was the index test.

Slide 9: STARD Item #9: Participants

Item 9 is in regards to the study participants: “Whether participants formed a consecutive, random, or convenience series.”

Included study participants can be either a consecutive series of all patients evaluated for eligibility and satisfying the inclusion criteria, or a sub-selection of eligible individuals. A sub-selection can be a truly random selection of all eligible individuals, or it can be a convenience sample; for example, if participants are enrolled only on specific days, during specific office hours, or when attending a specific physician. In a convenience series, participants may not provide a fair representation of the targeted population. This may jeopardize the applicability of the study results, and has been shown to be an important source of potential bias.

In the example, Devaux and colleagues explicitly report that they included “consecutive patients with a myocardial infarction referred for emergent percutaneous coronary intervention.” This implies that the way patients were selected is unlikely to be a source of bias.

Slide 10: STARD Item #12a: Test Methods

Item 12a: “Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory.”

As indicated, some tests produce continuous outcomes. These outcomes are often reclassified by defining a positivity cut-off, so that a 2x2 table can be produced, and measures of sensitivity and specificity can be calculated. Patients that exceed this positivity cut-off are considered ‘positive’; the other patients are considered ‘negative.’ It is important that authors report this positivity cut-off, so that future studies can aim to reproduce it, and clinicians can apply it in practice. It is also important to report whether this cut-off was pre-specified or exploratory. Pre-specified cut-offs can be based on previous studies, clinical practice, or recommendations by the manufacturer, for example. Exploratory cut-offs should be interpreted with much more caution, as they often lead to biased accuracy estimates and are difficult to reproduce, especially if established in small studies.

In the example, Foerch and colleagues evaluated the diagnostic accuracy of plasma glial fibrillary acidic protein, or GFAP, which produces continuous test results. In the Methods section, they reported, “We used ROC curve analysis to calculate diagnostic accuracy of GFAP.” They also reported, “We predefined a GFAP plasma concentration of 0.29 µg/L [..] as the cut-off.” So a positivity cut-off was pre-defined, at which sensitivity and specificity were

calculated. They also appropriately indicate that this cut-off was selected based on previous explorative studies.

Slide 11: STARD Item #13a: Test Methods

Item 13a: "Whether clinical information and reference standard results were available to the performers/readers of the index test."

Many tests require human interpretation. As indicated earlier, reading the index test may influence the final conclusions if the reader is aware of the results of the reference standard, which could lead to bias.

In the example, Lui and colleagues evaluated the accuracy of a serum microRNA expression profile for diagnosing pancreatic cancer. They state "the investigators performing the molecular analysis on the blood samples were blinded to the patients' clinical diagnosis." This implies that the risk of test review bias is low.

Slide 12: STARD Item #19: Results

Item 19: "Flow of participants, using a diagram."

Estimates of diagnostic accuracy may be seriously biased if not all eligible participants undergo the desired reference standard, or if many participants have missing or inconclusive test results. The extent to which missing and inconclusive results occur in diagnostic accuracy studies varies, and there can be many underlying causes. A test may fail, for example, because of technical reasons or an insufficient sample.

In itself, the frequency of missing or inconclusive test results is an important indicator of the feasibility of the test, and typically limits the overall clinical usefulness. The way researchers deal with such results in their analysis also varies. Some, for example, exclude participants with such results, which could lead to overestimations of the test's accuracy. Others include participants with such results as false positives or false negatives, which could lead to underestimations of the test's accuracy.

The most informative way to give insights in these potential sources of bias is to report a flow diagram which describes the flow of participants throughout the study. Ideally, such a diagram is detailed enough for readers to reproduce the final 2x2 table and recalculate reported estimates of diagnostic accuracy, thereby also improving the reproducibility and trustworthiness of the presented findings.

Slide 13: STARD 2015 Flow Diagram Prototype

This is the STARD 2015 prototype of such a flow diagram. It illustrates the number of individuals that were potentially eligible in a particular study, those that were actually included, which index tests and reference standards the participants underwent, what the results of these tests were, and what was done in case of missing or inconclusive test results.

Slide 14: STARD Item #24: Test Results

Item 24: “Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals).”

Diagnostic accuracy studies provide estimates of the ‘true’ accuracy of medical tests. The smaller the number of included participants, the larger the uncertainty will be that the calculated accuracy estimates actually represent the ‘true’ values. Therefore, it is recommended that authors not only report point estimates of accuracy, but also provide 95% confidence intervals around these estimates. Not doing so invites over-optimism about the performance of the test.

In the example, Debray and colleagues evaluated the diagnostic accuracy of the blood lactate-to-pyruvate molar ratio. In the Abstract, they report on a point estimate of specificity of 71%. They also report a 95% confidence interval around this point estimate, which illustrates that the “true” specificity of the test under investigation has a 95% chance of lying between 20% and 96%. Because the authors provided these confidence intervals, the reader is immediately aware that this estimate of specificity is very imprecise, and should be interpreted with caution.

Slide 15: Utilization of STARD 2015

Who can or should use the STARD 2015 list of essential items? Authors can use it to make certain that the report of their diagnostic accuracy study is thorough and complete. They can use the list as a structured guidance when writing the report or they can use it before submitting the report to a journal to make sure that all the essential items are reported, and incorporate those that are missing. Peer reviewers and journal editors can use it to assess the completeness of submitted reports of diagnostic accuracy studies, and identify essential items that are missing.

What are the advantages of using STARD 2015? If all essential applicable items are reported, the study will be more easily identified, easier to reproduce, and of value to the medical and scientific community. This is likely to improve the visibility and usefulness of study reports, thereby positively influencing its scientific impact. Adherence to reporting guidelines has also been positively associated with citation rates and journal impact factor. More than 200 journals now explicitly endorse STARD, including, for example, Lancet, JAMA, and *Clinical Chemistry*. This means that these journals require or recommend the use of STARD in their instructions to authors. *Clinical Chemistry*, for example, reports: “For studies of diagnostic accuracy of tests, complete the STARD Checklist for Evaluations of Diagnostic Accuracy electronically upon submission.”

Slide 16: Conclusions

In conclusion, suboptimal reporting of clinical diagnostic accuracy studies is considered to be one of the major sources of research waste, but one that is 100% preventable. Authors, peer reviewers, and journal editors should make every effort to ensure published study reports are sufficiently informative. STARD 2015 aims to improve the quality of reporting of diagnostic accuracy studies by providing a detailed list of 30 essential items that, if reported, facilitate the identification and reproducibility of the study, the assessment of risk of bias, and maximizes the clinical utility and applicability of the study.

Slide 17: References

Slide 18: Disclosures

Slide 19: Thank You from www.TraineeCouncil.org
Thank you for joining me on this Pearl of Laboratory Medicine on “Optimal Reporting of Diagnostic Accuracy Studies.”