Method comparison studies: an introduction to acceptability criteria

In this third article in his series on method comparisons, Stephen MacDonald moves on from experimental design and analysis, the sources of samples and the number required, to focus this month on what differences are expected and acceptable, and what other factors need further investigation before implementation of an assay.

When comparing methods our goal is to assure ourselves of the consistency of results between methods so that methodologies can be used interchangeably, or one can replace another without adversely affecting patient outcomes. This is an interesting concept for a number of reasons. It makes sense that we would not want to introduce an assay that is less clinically useful than what we currently have.

So, should our benchmark be to perform as well as the current method? Should we be aiming for better? Ideally, our goal is to achieve performance to specifications derived from clinical outcome studies. What if these studies were themselves performed on assays that are less sensitive or specific than what we have at our disposal now? Does that potentially change patient outcome? What specifications could we use?

Performance specifications

At its simplest, the assay must measure an association between the test and the condition or disease of interest. This is such a fundamental principle that it is decreed in EU regulation to determine the scientific validity of any analyte. Furthermore, the regulation ([EU] 2017/746) defines requirements for assays providing continuous (numerical) and dichotomous (positive/negative, diseased/non-diseased, detected/not-detected) measures.

The choice of what performance specifications we use has been an evolving process over the past 50 years. The latest framework has built upon the Stockholm and Milan conferences of 1999 and 2015, previously referenced in both the measurement uncertainty and statistical quality control series of articles.

Initially, five models of quality specifications were proposed in 1999. Four of them were further subdivided. This has been condensed to a more manageable three models covering a more conceptual approach. Although the framework is not meant to be viewed as a hierarchy of standards, there is an apparent order to the levels of evidence used in each of the three models. The minimum standard is based on the current state of the art. Sources for such information may come from performance in proficiency testing schemes or other schemes where a large assessment of performance data are available. Peer-reviewed publications referencing methodology development and performance characterisation are included as state-of-the-art knowledge.

Biological variation data are common to all historical and current frameworks, and are a very useful measure to guide performance standards, being linked

Some assays are established as screening tests, while others are more specifically used for diagnosis.

Decisions to introduce a new test are often approached by the scientific and clinical staff who have a natural interest and desire to expend the repertoire.

directly to the behaviour of the measurand in the system we are testing. There are a number of sources of such information, ranging from the well-established westgardqc.com website (which itself aggregates much of the data found in other summaries of biological variation) to individual studies published in the literature. Its application comes with a word of caution though. Published studies, particularly for measurands performed less often, may be smaller than we would like. Results from such studies, particularly when they are the only ones available, must be interpreted in the context they were derived.

Also, biological variation data invariably are generally performed in otherwise well, healthy volunteers. These data are essential to understand the underlying physiological variation of potential patients, but the complexity of studying biological variation in the diseased state means that data in that realm are lacking. When such studies are used, it is highly recommended that the source of data used for the performance specifications is clearly stated, why it was chosen and why that was deemed the most appropriate. If an assay is claimed to be at minimum, desirable or even optimal performance, that must be shown to be referencing a robust source of data to make the claim appropriate.

Finally comes clinical outcome data – subdivided as either direct or indirect. Direct clinical outcome data are notoriously difficult to obtain, particularly for the less-common measurands. The data must also be critically appraised as being appropriate. The scarcity of direct clinical outcome data has led to the need for indirect measures of clinical outcome. One such method is the allowable misclassification rate. This has been implemented with cardiac troponin in recent times. This quantifies misclassification error using duplicate measurements of the measurand.

Patient clinical outcome (treatment pathway followed) is classified based on each replicate. If there is a difference, the impact is directly attributable to assay imprecision. In this specific case, the false-positive rate can be limited to 1% if the troponin assay bias and imprecision is kept below 10%. It is this that can be used as a performance specification for assays under test to achieve.

Screening, diagnosis, treatment and monitoring

Assays can be used for any or all of the above purposes. The clinical requirements of the assay may differ depending on the clinical context. Some assays are established as screening tests, others are more specifically used for diagnosis, and yet others are used only for monitoring treatment such as drug therapies. Indirect clinical outcome data incorporate both analytical and clinical criteria. The clinical application may be different between, and even within, assays. So what is the difference?

Analytic performance is the ability of the assay to measure the analyte of interest. Clinical performance is the ability of the assay to use that measurement to make a distinction between different disease states of an individual, or to guide clinicians towards appropriate clinical interventions. We may need to quantify one before we can interpret the other.

The analytical data may be described by imprecision, reproducibility, linearity among others. Terms such as sensitivity and specificity particularly have been at the forefront of our minds, and that of the general public, due to SARS-CoV-2 and COVID-19, and relate to clinical performance. Less well publicised, but just as important, are predictive values (both positive and negative) and ratios (such as likelihood and odds ratios).

So, what should we consider first? Experiments to investigate analytical performance of assays should be performed first. The results of these experiments will inform our interpretation of diagnostic accuracy. Established assays permit interpretation of the diagnostic accuracy from analytical performance.

Ensure consistency with manufacturers' claims

What is expected?

Manufacturers are required to use a continual assessment process to evidence both the analytical and clinical validity of assays. The end goal of that assessment is to provide the performance evaluation to users of the assays as a baseline performance expectation. The *in vitro* diagnostic regulations (IVDR) replace previous EU legislation (97/79/EC) governing *in vitro* diagnostic devices. This came into place in May 2017. Manufacturers must provide more extensive and rigorous clinical evidence for their assays. These will include clinical performance studies which in themselves are extremely useful for the laboratory when finalising performance specifications for method comparisons. The documentation provided must also be more extensive and the details of such studies should be made available to the laboratories.

How do we use manufacturer specifications?

Laboratories should be able to reproduce (or better) the minimum quality requirement provided in the manufacturer documentation. However, we accept this is a starting point and should be viewed as a minimum. Verification studies including imprecision, reproducibility, bias assessment, calibration, traceability, and recovery experiments also provide

Manufacturers are required to use a continual assessment process to evidence both the analytical and clinical validity of assays

indispensable data for the most important part of the analytical phase of method comparability studies – measurement uncertainty estimation.

Measurement uncertainty

Measurement uncertainty (MU) must be used to compare analytical methods. It is impossible to compare methods, and to validate them as being appropriate, if we do not know what the expected measurement uncertainty is for each. Measurement uncertainty bridges analytic performance and clinical interpretation. All relevant uncertainty contributors are considered for inclusion into the preliminary MU budget for the assay under test. Data for the index method should already be on file. The important thing to note is that it does not involve any additional testing or experiments during the study. It is just applying the data in a way that means they can be applied directly to the comparability. Examples are assessment of bias, calibrators and traceability of results.

Bias assessment

Bias assessment comes in two forms. First there is identification of bias between our two methods under test – this assumes the index method as the 'true' value. Second is assessment of bias in either method, individually, from the 'true' value. We already know we cannot ever determine the true value but we can determine whether we are biased against international standards. Our index method will have been assessed regularly for bias using consensus values applied by external quality assessment (EQA) schemes. Bias should be corrected if possible, using an international standard to confirm the extent of the bias, and the uncertainty of that correction incorporated into the MU.

Calibrators

Calibrators, if used, and their associated uncertainty, are intrinsic to the assessment of measurement uncertainty – and the impact may be different depending on the method. It is essential that this is fully documented and, if different calibrators are used for each method, a comparison of the traceability chains, and of their performance for each methodology, should be explicit to ensure that the results are interchangeable.

Traceability of results

In the absence of the gold-standard methodology (which should be traceable through a chain to the SI system of units), traceability of the assay under test is limited to the index method. Reports should explicitly outline at the outset what metrics are used and how they are interpreted with respect to their traceability chain.

Armed with the data from our measurement uncertainty evaluations for both methods, we can be confident in knowing that if we observe a difference above the threshold we have set, we can attribute that to a variance other than as a consequence of MU (ie a genuine analytical difference between methods). That being said, the implication in the clinical setting is yet to be established, and the clinical comparison follows.

Clinical performance

Sensitivity and specificity

Sensitivity and specificity are always quoted. They are a useful starting point for our diagnostic assessment. We will cover their derivation in the worked examples in the coming articles in this series. Importantly, we must accept that we have not verified these data, which is the point of the comparison study. We accept that they are an essential part of the initial assessment, for a proof of concept. Without that, the assay would not be available on the market for us to use. The limitations of these data are important. Often, clinical studies may be quoted, but, even though we consider

them to be a strong source of evidence for performance, such studies are not infallible.

Clinical studies may be subject to selection bias. Often they are not representative of real-world clinical practice. 'Normal' volunteers/donors are often used as a control; confirmed 'diseased' patients as the diseased group. We expect these samples to be from the extremes of the clinical spectrum we are likely to encounter. This focuses the assay range under test to those extremes and neglects the areas between.

Numeric assay results will be expected to show a statistical difference in the means (average) of the two groups. Qualitative assays will show different proportions of positive patients when comparing control to diseased groups. The selection of 'very positive' cases and 'very negative' cases may have an influence on this. In the real world we see patients with a full spectrum of possible results, and often they are not in the extremes.

Patients also present with co-morbidities. How do these co-morbidities interact with the performance of our clinical results? Some of these co-morbidities affect our assay method, affecting sensitivity and specificity. These issues may not have been accounted for during validation so we must be prepared to account for these issues as part of our method comparison study.

Our reference method may or may not be affected and the same can be said for our assay under test. Differences in results may be genuine, and they may even be desirable if unwarranted effects are seen previously to be a problem in our index method. It is for this reason that we really must prepare our definition of what differences are expected, acceptable and what others need further investigation.

Prevalence

We have seen during the current pandemic that assay performance must be considered in the context of external factors, including the disease prevalence, at the time of testing if using diseased samples in comparison studies. Our statistical inference is impacted by the probability of encountering positive samples in the sampling population. We will discuss this in the coming articles, but moving from high to low prevalence significantly changes the probability of a 'genuine' positive result being able to be interpreted as such. The reduction of prevalence may lead us to interpret a positive result as being more likely to be a false positive (reduced specificity) rather than a true positive (high sensitivity). The interpretation of results may be a balancing act, and it is important to have an awareness of what the impact of changing patient prevalence has on the availability of verification samples and interpretation of the results.

> If considering different methodologies as a back up to a primary assay, method comparability is essential

What is the impact of disagreement?

It is clear why we need to set limits for our expected agreement. These may be set by evidence we have or can reference. As long as we can justify the criteria as being appropriate we can assess our results. This leads to another question: What happens if the results do not agree by these criteria? Is it the end of the road? Well, that depends.

If there is a difference, method-specific reference ranges may be required. This becomes even more essential, not to mention problematic, if there is a direct clinical cut-off. It is undesirable that we would deviate from established (in some cases international) clinical decision limits as a consequence of poor method comparison study.

Service resilience and continuity of care is also important. If considering different methodologies as a back up to a primary assay, method comparability is essential. The worst-case scenario would be for reference ranges and/or clinical action points to be changing back and forth as a consequence of one assay being unavailable and an alternative being used.

Conclusions

Often, decisions in the laboratory to introduce a new test are first approached by the scientific and clinical staff who have a natural interest and desire to expand the repertoire. The procurement process is agnostic to this and takes the scientific validity of the implementation to be a given. The other aspects of implementation, including but not limited to procurement, supply chain, costing and billing will determine whether the assay is successful in being implemented. It is worth, from the outset, having this at the forefront of your thinking so that when you have the scientific evidence from the method comparison study you are ready for the rest of the process that will inevitably follow. All we can do is provide the scientific data to establish the incoming method as appropriate.

The steps here are simple. Find a source for performance specifications that you can test the assay against. This may be from manufacturers, peer-reviewed publications, guidelines, biological variation databases or clinical outcome studies. Whichever you choose, you must be prepared to defend. Once you have them, we go back to our experimental data and see how we did. Clearly, the choice of specification needs to be in place prior to any experiments being run, as it will inform what types of samples are required. Along with our experimental design previously discussed, we can simply test the samples and assess.

The methods used to quantify the agreement are next. In the following few articles we will use a case study approach to incorporate our previous articles and start to implement the analysis and interpretation methodology we have at our disposal.

Further reading

- Fraser CG. The 1999 Stockholm Consensus Conference on quality specifications in laboratory medicine. *Clin Chem Lab Med* 2015; **53** (6): 837–40. doi: 10.1515/cclm-2014-0914.
- Horvath AR, Bossuyt PM, Sandberg S *et al.*; Test Evaluation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine. Setting analytical performance specifications based on outcome studies – is it possible? *Clin Chem Lab Med* 2015; **53** (6): 841–8. doi: 10.1515/cclm-2015-0214.
- Panteghini M, Ceriotti F, Jones G, Oosterhuis W, Plebani M, Sandberg S; Task Force on Performance Specifications in Laboratory Medicine. Strategies to define performance specifications in laboratory medicine: 3 years on from the Milan Strategic Conference. *Clin Chem Lab Med* 2017; **55** (12): 1849–56. doi: 10.1515/cclm-2017-0772.
- Pathology uncertainty (https://pathologyuncertainty.com).
- Sandberg S, Fraser CG, Horvath AR *et al.* Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015; **53** (6): 833–5. doi: 10.1515/cclm-2015-0067.

Dr Stephen MacDonald is Principal Clinical Scientist, The Specialist Haemostasis Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0QQ (tel 01223 216746).