Technical Guidance Series (TGS)
for WHO Prequalification – Diagnostic Assessment

Principles of performance studies

TGS–3
Preface

WHO Prequalification – Diagnostic Assessment: Technical Guidance Series

WHO Prequalification is coordinated through the Department of Essential Medicines and Health Products. WHO prequalification of in vitro diagnostic medical devices (IVDs) is intended to promote and facilitate access to safe, appropriate and affordable IVDs of good quality in an equitable manner. The focus is on IVDs for priority diseases and their suitability for use in resource-limited settings. WHO Prequalification undertakes a comprehensive assessment of individual IVDs through a standardized procedure that is aligned with international best regulatory practice. It also undertakes post-qualification activities for IVDs to ensure their ongoing compliance with prequalification requirements.

Products that are prequalified by WHO are eligible for procurement by United Nations agencies. The products are then commonly purchased for use in low- and middle-income countries.

IVDs prequalified by WHO are expected to be accurate, reliable and able to perform as intended for the lifetime of the IVD under conditions likely to be experienced by a typical user in resource-limited settings. The countries where WHO-prequalified IVDs are procured often have minimal regulatory requirements, and the use of IVDs in these countries presents specific challenges. For instance, IVDs are often used by healthcare workers who do not have extensive training in laboratory techniques, in harsh environmental conditions, in the absence of extensive pre- and post-test quality assurance capacity, and for patients with a disease profile that differs from the profiles encountered in high-income countries. Therefore, the requirements of WHO Prequalification may differ from the requirements of high-income countries, or those of the regulatory authority in the country of manufacture.

The Technical Guidance Series (TGS) was developed following a consultation held on 10–13 March 2015 in Geneva, Switzerland. The consultation was attended by experts from national regulatory authorities, national reference laboratories and WHO prequalification dossier reviewers and inspectors. The guidance series is a result of the efforts of this and other international working groups.

This guidance is intended for manufacturers interested in WHO prequalification of their IVD. It applies in principle to all IVDs that are eligible for WHO prequalification for use in WHO Member States. This guidance should be read in conjunction with relevant international and national standards and guidance.

The TGS guidance documents are freely available on the WHO website.
WHO Prequalification of IVDs

Contents

Acknowledgements .................................................................................................................. 2
List of contributors.................................................................................................................. 2

1 Abbreviations ..................................................................................................................... 3

2 Definitions .......................................................................................................................... 3

3 Purpose of this document ................................................................................................. 6

4 Introduction ......................................................................................................................... 7

5 Types of studies .................................................................................................................. 7

6 General considerations for analytical and clinical performance studies ................. 10
   6.1 Intended use ................................................................................................................. 10
   6.2 Reference standard .................................................................................................... 10
   6.3 Study validity (internal and external) ........................................................................ 10
   6.4 Bias .............................................................................................................................. 12
   6.5 Specimen type, collection and handling .................................................................... 15

7 Performance study protocol ............................................................................................. 16
   7.1 Study rationale ............................................................................................................ 16
   7.2 Ethical considerations ............................................................................................... 16
   7.3 Study objectives ......................................................................................................... 17
   7.4 Study method .............................................................................................................. 17

8 Performance study outputs ............................................................................................. 20
   8.1 Example table, clinical performance study: diagnostic sensitivity ..................... 21
   8.2 Example table, clinical performance study: diagnostic specificity .................... 21
   8.3 Example table, analytical performance study: precision ..................................... 22
   8.4 Example table, analytical performance study: interfering substances ............ 22
   8.5 Example table, analytical performance study: cross-reacting infections, diseases or medical conditions ................................................................. 23

9 Conclusion ......................................................................................................................... 23

10 References ......................................................................................................................... 24
Acknowledgements

The draft document *Technical Guidance Series for WHO Prequalification – Diagnostic assessment: Principles for performance studies* was developed with support from the Bill & Melinda Gates Foundation and UNITAID. This draft was prepared in collaboration with Dr E Cowan, Washington DC, United States of America (USA); Dr M Lanigan, Geneva, Switzerland; and D Healy and R Meurant, WHO; and with input and expertise from Dr S Hojvat, Maryland, USA and Dr L Kestens, Institute of Tropical Medicine Antwerp, Antwerp, Belgium. This document was produced under the coordination and supervision of Ms R Meurant and Ms I Prat, Prequalification Team – Diagnostic Assessment, WHO, Geneva, Switzerland.

List of contributors

The draft technical specifications document was posted on the WHO website for public consultation on 19 May 2016. Various stakeholders – including manufacturers submitting to WHO prequalification of IVDs, IVD manufacturing industry associations, various national and international regulatory bodies, and IVD standards organizations – were informed of the consultation in order to solicit feedback. A 2-month response period was provided.

Comments were received from the following:

Dr J Duncan, London, United Kingdom; European Federation of Laboratory Medicine Working Group for Test Evaluation, Working Group Chair: Professor C Cobbaert, Leiden University Medical Center, Department of Clinical Chemistry and Laboratory Medicine, Netherlands; Japanese Committee for Clinical Laboratory Standards Vice Chair & WG2 Delegate: Mr H Koide, Tokyo, Japan; Dr M Lanigan, Geneva, Switzerland; Médecins Sans Frontières International Office, (Dr C Kosack, Laboratory Working Group Leader, Dr AL Page, Epidemiologist at Epicentre and Ms E Tran, International Coordinator for MD/IVD quality), Geneva, Switzerland; and Dr S G. Schumacher, FIND Headquarters, Geneva, Switzerland.
1 Abbreviations

CI  confidence interval
CLSI  Clinical And Laboratory Standards Institute
CV  coefficient of variation
FDA  US Food and Drug Administration
GHTF  Global Harmonization Task Force
IMDRF  International Medical Device Regulators Forum
IFU  instructions for use
ISO  International Organization for Standardization
IVD  in vitro diagnostic medical device
mL  millilitre
mmol  millimole
NPV  negative predictive value
PPV  positive predictive value
QC  quality control
S/Co  signal to cut-off ratio
SD  standard deviation
SOP  standard operating procedure
STARD  Standards For Reporting Diagnostic Accuracy Studies Initiative
WHO  World Health Organization

2 Definitions

The definitions given below apply to the terms used in this document. They may have different meanings in other contexts.

**Analytical performance study:** A study undertaken to assess the ability of the in vitro diagnostic medical device (IVD) to measure a particular analyte. *Adapted from (1)*

**Clinical performance study:** A study undertaken to establish or confirm the clinical performance of an IVD. *(1)*

**(Clinical) reference standard:** The best currently available criteria for establishing the presence or absence of the condition, event or characteristic of interest using a single
method or combination of methods, including laboratory tests, imaging tests, pathology and clinical information (including follow-up).

Adapted from Clinical And Laboratory Standards Institute (CLSI) definition for “diagnostic accuracy criteria” (2, 3)

Note 1. A reference standard divides the intended-use population into only two groups (condition present or absent); it does not consider the outcome of the new test under evaluation. The reference standard can be a single test or method, or a combination of methods and techniques, including clinical follow-up. If a reference standard is a combination of methods, the algorithm specifying how the different results are combined to make a final positive or negative classification (which may include the choice and ordering of these methods) is part of the standard. (4)

Note 2. The determination of what constitutes the “best available method” and whether that method should be considered a “reference standard” is established by opinion and practice within the medical, laboratory and regulatory community.

Intended use: The objective intent of the manufacturer regarding the use of a product, process or service as reflected in the specifications, instructions and information provided by the manufacturer. (1)

Matrix: All components of a material system, except the clinically relevant analyte forms. Adapted from (5)

Result and interpretation: This document distinguishes between the terms “result” and “interpretation” as outputs of an IVD. A “result” is taken to mean the signal or output of a test, which is generally the first reading from an IVD. For example, a test may give a signal of 0.5; hence, with a cut-off value of 0.25, the signal cut-off ratio (S/Co) would be calculated to be 2. Each of these values would be considered to be results. For a rapid diagnostic test that shows the presence or absence of one or more of each of the test and control bands, the presence of those bands would be considered to be results.

“Interpretation” is taken to mean the final output from using an IVD and is derived from interpreting test results according to the instructions for use (IFU). In both of the examples above, the interpretation might be “positive”, “reactive” or similar.

Specimen and sample: Here, “specimen” refers to material collected directly from an individual. Where a study material is generated by diluting, pooling, spiking and so on of one or more specimens, the term “sample” is replaced by a term such as “contrived material”, “quality control (QC) panel member” or “dilution”, as appropriate.

Where possible, the use of the term “sample” on its own is confined to its statistical meaning.
WHO Prequalification of IVDs

**Specimen type:** Refers to the different types of specimen collected for use with an IVD. In this context, examples of specimen types include urine, oral fluid, whole blood, serum or plasma. (6)

**Study validity:** The degree to which the inference drawn from a study is warranted when taking into account the study methods, the representativeness of the study sample and the nature of the population from which it is drawn.

Note: Two varieties of study validity are distinguished, internal validity and external validity:

**Internal validity** – Bias can arise from multiple sources, in particular the choice and handling of patient selection, index test, reference standard, and flow and timing, as outlined in QUADAS-2 and related documents. The index and comparison groups are selected in such a manner that the observed differences between them on the dependent variables under study may be attributed only to the hypothesized effect under investigation. (7-9)

**External validity (generalizability) –** A study is externally valid, or generalizable, if it can produce unbiased inferences regarding a target population (i.e. beyond the subjects in the study). This aspect of validity is only meaningful with regard to a specified external target population. For example, the results of a study conducted using only white male subjects might or might not be generalizable to all human males, or to human females. Compared with internal validity, the evaluation of generalizability usually involves much more subject-matter judgement. (10)
3 Purpose of this document

The purpose of this technical guidance document is to identify the key principles that apply when conducting and reporting the study design, results and conclusion of analytical and clinical performance studies that support performance claims for in vitro diagnostic medical devices (IVDs) undergoing assessment for WHO prequalification. The document then summarizes the principles that generally apply to most performance studies.

The elements discussed here support recommendations from international regulatory harmonization groups including the International Medical Device Regulators Forum (IMDRF; previously the Global Harmonization Task Force, GHTF), and the Asian Harmonization Working Party. They are consistent with the requirements of mature regulatory authorities such as the US Food and Drug Administration (FDA), the European Union and Health Canada; they are also consistent with the requirements of standards setting bodies such as the International Organization for Standardization (ISO) and the Clinical and Laboratory Standards Institute (CLSI). In addition, this document mirrors approaches taken by the Standards for Reporting Diagnostic Accuracy Studies initiative (STARD), which is a checklist of essential elements for the design and reporting of diagnostic evaluations (2). STARD was created because of concerns about the poor quality of design and reporting of clinical performance studies in the literature, and it has been adopted by journals as a reporting standard. This WHO document also draws on the FDA document Statistical guidance on reporting results from studies evaluating diagnostic tests (11).

Guidance on the execution of best practice in specific performance studies can be found in more detailed guidance and standards. The WHO guidance document TGS-1 Standards applicable to the WHO prequalification of in vitro diagnostics (12) provides a comprehensive checklist of standards and guidance documents applicable to the IVDs assessed for WHO prequalification, including those relevant to analytical and clinical performance studies.

The scope of this document does not extend to the demonstration of clinical utility; that is, the effectiveness or benefits of an IVD, relative to or in combination with other measures, as a tool to inform clinical intervention in a given population or health-care setting. To demonstrate clinical utility, a separate set of studies is required. Clinical utility studies usually inform programmatic strategy and are thus the responsibility of programme managers, ministries of health and other related bodies in individual WHO Member States.
4 Introduction

The WHO prequalification assessment process for IVDs requires submission and assessment of a product dossier. A central part of such a dossier is a description and documentation of the studies conducted to establish the performance of the IVD. This information should be included in Section 7 (“Product performance specification, and associated validation and verification studies”) of the product dossier submitted for WHO prequalification assessment.

The manufacturer must carry out relevant investigations to support the intended use. These investigations should cover analytical and clinical sensitivity and specificity, accuracy (bias and precision, repeatability and reproducibility), linearity, detection limits and traceability, as appropriate. In addition, WHO requires investigations to assess the potential effects of interfering factors and the provision of evidence in support of claims regarding specimen, reagent and product stability, and robustness. Studies in support of the intended use should take into account the intended specimens, the intended user and the intended setting of use. Communicating the design, results and conclusions from these studies in a clear and organized fashion is essential to an efficient WHO prequalification assessment.

5 Types of studies

Two broad types of studies describe the performance of IVDs and are required as part of the dossier submission for WHO prequalification of IVDs: analytical performance studies and clinical performance studies. As defined by GHTF (1):

- **analytical performance studies** evaluate “the ability of an IVD medical device to detect or measure a particular analyte”; and
- **clinical performance studies** demonstrate “the ability of the IVD medical device to yield results that are correlated with a particular condition/physiological state in accordance with (the) target population and intended user”.

Thus, the analytical performance studies estimate the intrinsic properties of the IVD, whereas the clinical performance studies determine the expected performance of the IVD in an intended use setting by intended users, using the intended specimen type. Table 4-1 summarizes and highlights the characteristics of each type of study.

The number and type of performance studies that a manufacturer should perform to validate an IVD will depend on the design and nature of the IVD itself. As part of its quality management system, a manufacturer should conduct a full risk analysis of the IVD being developed. Through this process, the risk associated with the manufacture, intended use and so on of the IVD can be assessed. Based on this assessment, relevant standards, references and regulatory requirements can be identified and validation studies designed as a means to mitigate those risks (13). Nevertheless, an IVD product dossier submitted for prequalification would be
expected to include at least those analytical and clinical performance studies summarized in Table 5-1. Further information on each of these studies can be found in the WHO prequalification document *Instructions for compilation of a product dossier* (14).

Analytical and clinical performance studies differ markedly in the type of evidence they generate and the performance claims that such evidence is intended to support. Nevertheless, both categories of performance studies are underpinned by several common key principles that are discussed here.

The evaluation of the analytical and clinical performance characteristics should be conducted according to international standards or recommendations. For example, precision and bias can be verified according to CLSI EP05-A3 (15) and CLSI EP09-A3 (16).

In the absence of appropriate justification and mitigating circumstances, WHO prequalification expects performance evidence to be generated using the IVD that is intended for prequalification, and not surrogates or closely-related products.
**WHO Prequalification of IVDs**

Table 5-1: Analytical and clinical performance studies to support IVD performance claims

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Analytical performance studies</th>
<th>Clinical performance studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td>• Establish intrinsic performance capabilities</td>
<td>• Establish expected performance for intended use with intended users</td>
</tr>
<tr>
<td><strong>What is evaluated</strong></td>
<td>• Specimen type (suitability, collection, storage and transport stability)</td>
<td>• Clinical sensitivity</td>
</tr>
<tr>
<td></td>
<td>• Equivalence between specimen types</td>
<td>• Clinical specificity</td>
</tr>
<tr>
<td></td>
<td>• Analytical performance characteristics:</td>
<td>• PPV</td>
</tr>
<tr>
<td></td>
<td>• accuracy</td>
<td>• NPV</td>
</tr>
<tr>
<td></td>
<td>• trueness and bias</td>
<td>• End-user verification of labelling and IFU (self-testing)</td>
</tr>
<tr>
<td></td>
<td>• precision (repeatability and reproducibility)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Analytical sensitivity (limit of detection, detection of variants)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Analytical specificity (interference and cross-reactivity)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Measuring range of the assay</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validation of assay cut-off</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validation of assay reading time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Traceability of calibrators and control materials</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Validation of assay procedure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• IVD and reagent stability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Shelf-life</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• In-use stability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Shipping stability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Robustness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Human factors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Verification of labelling and IFU (non-self-testing)</td>
<td></td>
</tr>
<tr>
<td><strong>Types of specimens for studies</strong></td>
<td>• Pedigreed repository specimens</td>
<td>• Intended-use specimens collected from intended-use population</td>
</tr>
<tr>
<td></td>
<td>• Commercial reference panels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Contrived specimens to produce a target concentration or level of reactivity to challenge product</td>
<td></td>
</tr>
<tr>
<td><strong>Who performs testing</strong></td>
<td>• Trained user (in some cases intended user)</td>
<td>• Intended user</td>
</tr>
</tbody>
</table>

IFU, instructions for use; IVD, in vitro diagnostic medical device; NPV, negative predictive value; PPV, positive predictive value
6  General considerations for analytical and clinical performance studies

6.1  Intended use

The intended use describes how the test is to be used and by whom, for what condition, with what specimen type, for what patient or individual (e.g. age, race, gender, geography or clinical condition), and what is to be detected. The intended use also includes the function of the test that describes the circumstances under which an individual or patient would be tested. Functions may include:

- screening (e.g. for surveillance or safety of blood supply);
- aiding the diagnosis and determination of a patient’s disease course and prognosis;
- monitoring patient therapy or following their progress after treatment;
- staging or aid to staging of disease; and
- disease differentiation or prediction.

The intended use of an IVD will largely dictate the type and scope of both analytical and clinical performance studies.

6.2  Reference standard

Meaningful conclusions about the IVD performance can only be derived from performance studies that make use of specimens in which the true analyte or clinical status has been determined to a high level of confidence. The true status of a specimen or subject should be determined using an appropriate reference standard (the best available method for establishing the presence or absence of the target condition). In some cases, the reference standard may be a single state-of-the-art test; in others, a validated testing algorithm may be required. In situations where this is not feasible, clinical presentation and surrogate markers may provide indirect evidence in support of the presence or absence of a target condition, but the limitations of this approach should be carefully considered (17). Neither a testing algorithm nor a single reference test should in any way include the test under evaluation or a variation of it.

6.3  Study validity (internal and external)

The validity of a study is defined as the “degree to which the inference drawn from a study is warranted when account is taken of the study methods, the representativeness of the study sample, and the nature of the population from which it is drawn” (7). Study validity can be further defined as comprising both internal and external components.
Internal validity

Internal validity is the absence of systematic error; it describes how well an observed phenomenon can be attributed to a hypothesized effect. A study is internally valid or unbiased if it can produce unbiased inferences regarding a target parameter (e.g. sensitivity and specificity).

Example of internal validity:

**Stability** – it is appropriate to investigate the effect of the intended storage conditions on the retained lots of an IVD stored for increasing lengths of time. The internal validity of such a study is based on:

- the extent to which storage conditions and durations are defined; whether results are reliable and the reliability of those results can be understood (e.g. replicate testing at each time point and temperature condition are reported); and
- how results are analysed towards a final conclusion (e.g. pass and fail criteria are specified that address the study aim).

External validity

A performance study can be said to have external validity if its results can be generalized to a specified and relevant target population beyond the sample tested. With clinical performance studies, external validity can also be thought of as the extent to which results from a study reflect the real-world performance of the test in the intended-use population (7).

Examples of external validity:

- **Stability** – Determination of the stability of an IVD is externally valid if testing is conducted using multiple lots or batches of the IVD at a validated batch size drawn from (or representative of) routine manufacture. The study must also demonstrate that the storage conditions were relevant to the intended-use setting in order to address the intended operating environment.

- **Claimed specimen types** – external validity is demonstrated by understanding the relationship between specimen matrices for all claimed specimen types. Thus, for an IVD that is claimed for use with whole blood, serum and plasma, if testing is only conducted using serum, it is necessary to understand whether or not results can be generalized to both plasma and whole blood, and what evidence there is to support this.

With clinical performance studies, external validity can also be ensured by using the final version of the IVD according to the final IFU. If changes must be made, they should be validated to demonstrate that they do not have an adverse impact on test safety or performance by ensuring that: testing is conducted by multiple test
operators at multiple geographically distinct test sites (typically a minimum of three sites and testing three independent product lotsa)

- with the appropriate level of training for the particular study; and
- appropriate studies are conducted in the intended-use setting to address the intended operating environment.

A well-designed study, whether it examines analytical or clinical performance, must be able to demonstrate both high internal and external validity. Designing a study so as to minimize experimental bias is fundamental to ensuring study validity.

6.4 Bias

Bias refers to the extent to which a result deviates from a correct value. This may manifest, for example, as an assay bias, where a test value from an IVD differs from that expected by similar testing using a reference method or material. It is important to strive for traceability of test results to internationally recognized standards and reference materials of higher order, and the results should be shown to be commutable.

Bias may also arise from one or more flaws in the design of a study that lead to conclusions that do not adequately reflect the truth about the performance of an IVD (7). As noted above, bias can arise from multiple sources, in particular the choice and handling of patient selection, index test, reference standard, and flow and timing (7-9). Several important sources of experimental bias are summarized in Table 6-1.

\[a\] Lots with different critical reagents (e.g. biological reagents prepared in different syntheses, growths or purifications; other risk-defined critical reagents from different manufactured lots, or different suppliers if applicable).
Table 6-1. Types of bias

<table>
<thead>
<tr>
<th>Type of bias</th>
<th>Explanation</th>
<th>Methods for minimizing bias</th>
</tr>
</thead>
</table>
| Spectrum composition bias | When an IVD is tested using specimens that do not represent either the intended use or intended user population. | This type of bias can be minimized by:  
• ensuring that specimens come from individuals at all stages of the condition that the test is designed to detect;  
• testing specimens that contain potentially interfering substances as well as those from individuals with potentially interfering medical conditions representative of the testing population; and  
• ensuring that specimens are tested from individuals representing an appropriate demographic diversity (e.g. age, sex and ethnicity) within the context of the intended use and setting. |
<p>| Selection bias | When any specimens that are unlikely to perform well in a validation study are removed from a testing panel that otherwise represents the intended use and testing population. Example: A manufacturer may have developed an enhanced version of an existing IVD, with the variation consisting of reagent formulations that increase analytical sensitivity. However, the study design consists of testing only specimens that were positive in the less sensitive original IVD. Selecting specimens in this way does not provide a true challenge of the claim that the enhanced IVD has greater analytical sensitivity; it only demonstrates that the sensitivity is as good as the original IVD. The same bias could be introduced in any IVD where specimens are chosen so that they are significantly above the S/Co. | This type of bias can be minimized by applying transparent, objective and scientifically rigorous inclusion criteria that do not unduly favour the IVD being evaluated. Ideally, patients for clinical performance studies would be prospectively enrolled consecutive patients meeting pre-specified selection criteria. Selection bias can also be addressed by ensuring that the testing panel contains specimens that would challenge the IVD (e.g. low-titre specimens, and those from unselected blood donors, including first time donors, to be used for determination of specificity) (18). |
| Sampling bias | When too few specimens are tested to be representative of the target population. Example: A bias of this type can overestimate IVD performance by suggesting a low risk of false negativity where insufficient numbers of specimens have been tested to detect this possibility. | Sampling bias can be minimized by ensuring that statistically meaningful numbers of specimens, sufficient to detect a phenomenon should it exist, are tested in a performance study. |</p>
<table>
<thead>
<tr>
<th>Type of bias</th>
<th>Explanation</th>
<th>Methods to minimize bias</th>
</tr>
</thead>
</table>
| Workup bias                         | When there is a poor or insufficient method of characterization that does not allow determination of the true analyte status of a specimen recruited to a testing panel.  
Example: A study may attempt to determine the clinical sensitivity of an IVD by comparing test results to those from a single comparator assay. In this case, further testing to determine true analyte status will only occur for those specimens where results are discordant between the IVD and comparator assay. Moreover, such an experimental design does not take into account and cannot detect that test results may be both concordant and incorrect. | Workup bias can be minimized by using an appropriate reference standard or a validated testing algorithm on all specimens used in a study. The manufacturer has an obligation to ensure that the reference standard chosen is validated and that the validation is undertaken in a competent laboratory; for example, one functioning under ISO 15189 (19) or equivalent.                                                                                   |
| Review, observer, or information bias | When test operators who are aware of a prior test result or the clinical status of the individual from whom the test specimen was obtained introduce bias.  
Example: For a visually read IVD, if a tester is aware that a specimen is from an infected individual, that test operator may be expecting a positive test result. Likewise, if the test operator conducts both the reference test and the test under evaluation, then the operator may not be able to give an objective test interpretation, especially in the case of a weak positive specimen. | This type of bias can be minimized by ensuring that test operators are blinded to any prior test results or to the clinical status of the individual from whom the specimen was derived. This can be done by assigning a specimen code to replace a patient name or patient number and randomising testing order of the specimens. The testing technician or a clinical investigator would not know the key linking the specimen code to an individual patient history until the study was completed.  
It is important to ensure that test operators record results to the highest detail practicable. For example, for a qualitative rapid test (providing results of: “positive”, “negative” or “indeterminate”), at least semi-qualitative test results should be recorded (e.g. band intensities scored as –, +, ++, ++++,+++++). Results recorded in this way are important for study validity in that they allow changes in IVD performance to be better understood (e.g. signal degradation over time) than is the case for qualitative statements such as “positive”, “negative” or “all specimens passed”. |
6.5 Specimen type, collection and handling

The numbers and types of specimens used in performance studies will depend largely on the studies themselves, particularly on whether it is analytical or clinical performance that is being investigated. International guidance also provides recommendations (13). Consideration should be given to the ability of an IVD to detect all claimed analytes; for example, for an IVD intended to detect HIV-1 including Group O and HIV-2, performance should not consist solely of testing using HIV-1 antibody-positive specimens.

Specimens used in analytical studies will vary depending on the study objectives, but each study should make use of specimens that provide a level of reactivity that demonstrates how well the test performs at its limits. Ideally, specimens should be of the same matrix intended for use with the test (e.g. serum, plasma, finger-prick whole blood or oral fluid). However, low-reactive specimens close to the cut-off value, which can be of the greatest value in testing the limits of performance, may be difficult to obtain or be in short supply. If this is the case, contrived specimens (e.g. negative specimens in the corresponding matrix spiked to a low level of reactivity with the test analyte, or dilutions of a high-concentration specimen) may be used in a study, provided that the approach has a comprehensive scientific justification.

The choice of sample specimen type will be dictated by the intended use of the IVD (and the intended specimen for use with the IVD). Clinical evidence must be presented for all claimed specimen types. If a full clinical study is performed on only one of several claimed specimen types, this approach must be justified. Specimens for clinical performance studies typically come from three possible sources:

- Specimens taken prospectively from patients with appropriate disease signs and symptoms, with the intention that the specimens be used in a particular clinical performance study. These specimens may be tested immediately (fresh) or may be aliquoted and stored refrigerated or frozen for testing at a later time. If tested at a later time, specimen storage conditions (e.g. temperature, duration and the effect of specimen freeze-thaw cycles on the specific test analyte) must be consistent with those determined as part of analytical studies conducted during earlier stages of product development.
- Leftover specimens collected for routine diagnostic testing that would otherwise be discarded, or specimens collected for research purposes. Knowledge of specimen storage and handling before use of leftover or research-use specimens is important, as are any ethical considerations related to the patient source.
- Archived specimens that were collected in the past and were stored for extended periods of time in repositories. These specimens would be made available for use by those conducting analytical and clinical performance studies, or for use in product research and development. As above, specimens should only be used if their storage has been consistent with storage requirements (e.g.
duration, temperature and freeze-thaw cycles) determined for specimens during analytical testing of the IVD.

Regardless of the route of acquisition, particular care must be taken to ensure both that specimen integrity is maintained during the course of a study, and that the acquisition of specimens does not introduce one or more types of bias such as selection bias (see Section 6.4).

7 Performance study protocol

All studies, analytical or clinical, must be based on a detailed and comprehensive study protocol. Information on such protocols can be found in a number of guidance documents and standards (e.g. 20).

The specific content of a study protocol for different investigations will depend on the characteristic being validated, which in turn will depend on the risk management and planning that has been undertaken. However, most analytical and clinical performance studies share several common features. In general, protocols for studies investigating either analytical or clinical performance characteristics should include the following features, discussed below: study rationale, ethical considerations, study objectives and study method.

7.1 Study rationale

The study protocol should include an explanation of why the study is being conducted. For example, a study may have been instigated in order to address a risk-mitigation or regulatory requirement identified as part of the manufacturer’s risk analysis of product development. The study rationale will have a bearing on issues such as whether the study receives approval by an ethics committee.

7.2 Ethical considerations

Clinical performance studies that utilize prospectively collected specimens should ensure that the rights, safety and well-being of subjects participating in a clinical performance study are protected in accordance with the Declaration of Helsinki (21). That is, each clinical performance study should generate new data, the benefits to health must outweigh risks to study participants and any risks must be minimized; also, confidentiality must be respected. An ethics committee should review, approve and monitor studies to ensure that human rights and welfare are protected.

Clinical performance studies should also make users of the IVD or subjects recruited to the study aware of the limitations of the IVD, including warnings and precautions relevant to its use. Procedures should also be in place that allow the reporting of adverse events. Where appropriate, clinical end-points or requirements for patient or subject follow-up should be
identified. It should be noted that informed consent and ethics approval may be required in certain jurisdictions for clinical performance studies that use leftover or stored specimens.

7.3 Study objectives

The study protocol should include a description of the specific experimental objectives or aims the study is intended to address. Clearly defined study objectives form a basis from which an experimental method can be devised and the validity of the subsequent experimental findings can be judged. In general, any performance study (whether analytical or clinical) will have the broad objective of attempting to determine the reliability of one or more analytes in one or more specimen types. Performance studies submitted to WHO for prequalification assessment (whether analytical or clinical) must address all analytes an IVD is intended to detect and in all specimen types they are intended to be detected. Where this is not possible, justification for not performing certain studies or not including certain specimens should be provided.

7.4 Study method

A detailed description should be provided of the experimental method that will be used to address the stated aims and objectives. Highly summarized descriptions (e.g. “sensitivity will be determined”) should be avoided. Also, sufficient information should be provided to allow an independent assessor to understand and examine the scientific validity of the experiment, including how experimental biases have been minimized (e.g. by specimen or tester blinding). Protocols should be in place to ensure that sufficient numbers of all reagents, QC specimens, ancillary components and other supplies are collected, labelled and stored appropriately for the whole of the study. The method should also include the following, discussed below:

- descriptions of test methods, test kits or other reagents and required materials, and how they will be used in the performance study;
- a description of how results from tests will be recorded and interpreted; and
- the numbers and types of specimens or samples that will be used and how these were or will be acquired.

7.4.1 Descriptions of test methods, test kits or other reagents and required materials, and how they will be used in the performance study

The lot numbers of kits and reagents should also be recorded. It is important to ensure that standard operating procedures (SOPs) are in place for each of the testing methods to be used, and that all instruments have been validated. Moreover, technical staff should be made aware of any hazards involved in carrying out the study.

Note 1. Results of (clinical or analytical) performance studies submitted to WHO prequalification should be generated from the same version of the product (including assay and IFU) as that intended to be prequalified (sometimes known as the “lock-down” or
“frozen” design). Where this does not happen, full justification should be provided.

Note 2. For clinical performance studies, all sites should use the same version of the product (assay and IFU). Any deviations should be recorded, reported and explained. If clinical performance studies submitted to prequalification have been generated by a version of the product that is changed in any manner from that submitted to WHO, a detailed justification and validation that the modification has no impact on the performance results must be supported by validation studies of the modification. Notwithstanding this consideration, certain changes (e.g. substantial modification of a critical reagent) constitute a new product.

7.4.2 A description of how results from tests will be recorded and interpreted

Enough details should be provided for verification of the performance and for the independent reviewer to repeat the statistics. Raw data should be available on request. The acceptance criteria and their justification, including references shall be recorded. All results (including where a test is invalid) should be recorded to the highest level of detail practicable, regardless of the final output of a test. For example, if a rapid diagnostic test gives reactive results, every effort should be made during experimentation to record band intensities in at least a semiquantitative fashion. All results from each individual test specimen should be recorded; for example, it is not sufficient to record “all specimens were reactive”.

7.4.3 The numbers and types of specimens or samples that will be used and how these were or will be acquired

Analytical and clinical performance studies may use different types of materials or specimens that may be acquired in different ways. This section discusses the different types of specimens required for analytical and clinical studies, and the need for study oversight and monitoring, what is required for testing protocols, data collection and management, and data analysis.

7.4.3.1 Specimens for analytical studies

Specimens should be chosen so that their reactivity in an IVD is able to demonstrate the limits of IVD performance. Ideally, specimens should be of the same matrix intended for use with the test (e.g. serum, plasma, finger-prick whole blood or oral fluid). However, low-reactive specimens close to the cut-off value, which can be of the greatest value in testing limits of performance, are often difficult to obtain or may be in short supply. If this is the case, it is possible to evaluate more limited specimen matrices (as long as all specimen matrices effects are known), contrived specimens (e.g. negative specimens spiked to a low level of reactivity with the test analyte) or dilutions of a high-concentration specimen, provided that they are in an appropriate specimen matrix. The use of contrived specimens
or of an alternative specimen matrix must be justified and accompanied by supporting evidence.

7.4.3.2 Specimens for clinical studies

Clinical specimens should be acquired from a population likely to represent the intended use, user and setting, and in numbers that are sufficient to represent IVD performance in that population. Specimens should be chosen that are consistent with the performance characteristics of an IVD (e.g. are presumed to have analyte concentration within the measuring range of the assay, do not contain known interfering substances and so on, as appropriate). However, when choosing specimens, care must be taken to avoid introducing a selection bias.

Where specimens or personal data are collected specifically for the clinical performance study, or where specimens or data can be traced back to an individual, documented informed consent must be obtained for the data to be used in support of the WHO prequalification dossier submission.

The risk assessment conducted as part of product development should also include a component that that accounts for any potential risks posed (to user or patient, or both) by the product during the course of the clinical study.

7.4.4 Testing protocol

The testing protocol should include:

- inclusion and exclusion criteria for both the specimens used in a study and the results and analyses that arise from those specimens;
- a mechanism for both reporting and justifying exclusion of particular specimens and results, should exclusions occur; and
- a description of how the “true” analyte status of a specimen will be determined (e.g. using a reference standard) and how testing discrepancies will be resolved.

7.4.5 Data collection and management

Worksheets or some equivalent method of recording the information about the study, should be used during the course of the experiment. There should also be a description of how the collected information will be reviewed and approved by appropriately qualified and trained staff. Data collection methodology must be defined for each study and should include:

- recording the lot and item numbers of each component and IVD tested;
- recording any unexpected events noticed while the test is being performed; and
- collecting and storing raw data in a secure manner.
7.4.6 Data analysis

The study protocol should provide a detailed description of how study data will be analysed. The description should include the statistical methods that will be used to analyse study results and how these are appropriate for addressing the study objectives.

7.4.7 Study oversight and monitoring

It is important to establish managerial control and review of the progression of a planned study. Clinical performance studies will typically be conducted by a testing laboratory external to, and independent of, the manufacturer of an IVD. It is expected that the external testing site will appoint a clinical investigator who will be responsible for the overall conduct of a clinical investigation ensuring the well-being of subjects and adherence to the study protocol (including proper reporting of results and maintenance of records).

Regardless of where a performance study (clinical or otherwise) is conducted (whether in-house or by an external body), it is good practice to conduct study monitoring. This is a process whereby one or more appropriately qualified and experienced individuals ensure adherence to the study protocol and ethics, the generation of accurate and complete study data, and the documentation and scrutinizing of any protocol deviations (should they occur). The manufacturer should instigate a monitoring process for each study, whether analytical or clinical, performed by a third party on the manufacturer’s behalf.

8 Performance study outputs

The output of a performance study, whether analytical or clinical, will be one or more study reports (e.g. one or more interim progress reports, culminating in a final report at the completion of testing). The study report (whether interim or final) should provide at least the following:

- an executive summary – this should include a summary of the experimental protocol (as described in detail above) as it was intended and as it was actually performed, to ensure that it aligns with the study validity principles outlined in the study protocol;
- for clinical performance study reports, a discussion on the study population demographics, to allow a clear understanding of limitations of the studies – this will be part of the limitations on the use of the test (e.g. age limits) and will also address study bias;
- lot numbers involved and the location of the manufacturing documentation;
- criteria for all the testing (including physical, chemical and QC panels at the start and end of the assigned life of the components), and location of the records of all original testing data and records of storage conditions;
• any deviations from, or additions to, the study protocol, and justifications for these, including specimen exclusions, collection procedure as it was actually performed, and so on;
• tabulated or graphical summaries of the evidence in support of the performance claim being validated – any such table or graph should be accompanied by an explanation of how the experimental evidence supports the performance claim, as well as any inherent limitations to conclusions that can be drawn from the study;
• full study data (as an addendum) to support any summary evidence – annexes should be included, giving raw or intermediate results that allow verification of the summary (statistical) results; and
• a final conclusion stating whether or not the study’s stated objectives had been satisfactorily addressed and the consequences this has for product development and validation.

WHO encourages retention of photographic records, machine printouts and electronic data, or physical retention of membranes from opened cassettes, as appropriate. Records should be retained for the period of time equivalent to the commercial lifetime of the IVD, but not less than 2 years, as explained in paragraph 4.2.4. of (8).

The way that study results are summarized will depend on the study design and the results. Some examples of typical performance parameters for both qualitative and quantitative IVDs, and appropriate ways to present study findings, are presented below.

The tables below are for illustrative purposes and although each example table below purports to present summary data in an easy to read format, WHO expects any summarized results to be accompanied by appendices of full study data in a prequalification dossier.

### 8.1 Example table, clinical performance study: diagnostic sensitivity

Table 8-1. Summary of results for clinical trial Xxxx, for determination of diagnostic sensitivity

<table>
<thead>
<tr>
<th>Study site</th>
<th>Number of specimens tested</th>
<th>Number of specimens reactive by reference method</th>
<th>Number of valid tests</th>
<th>Number of specimens reactive in the IVD</th>
<th>Number of specimens falsely nonreactive</th>
<th>% sensitivity</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
</tbody>
</table>

### 8.2 Example table, clinical performance study: diagnostic specificity

Table 8-2. Summary of results for clinical trial Xxxx, for determination of diagnostic specificity

<table>
<thead>
<tr>
<th>Study site</th>
<th>Number of</th>
<th>Number of</th>
<th>Number of valid</th>
<th>Number of specimens</th>
<th>Number of</th>
<th>% specificity</th>
<th>95% confidence</th>
</tr>
</thead>
</table>
8.3 **Example table, analytical performance study: precision**

Table 8-3. Summary of assay precision (repeatability)

<table>
<thead>
<tr>
<th>QC panel member</th>
<th>Number of replicate tests</th>
<th>S/Co</th>
<th>Within-condition % CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative control</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>QC-1 (low-titre positive)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>QC-2 (mid-range positive)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>QC-3 (high-titre positive)</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
</tbody>
</table>

QC quality control, S/Co signal to cut-off ratio, SD standard deviation, CV coefficient of variation

Table 8-4. Summary of assay precision (reproducibility) for QC panel member QC-1 (low-titre positive)

<table>
<thead>
<tr>
<th>Results for QC panel member QC-1 (low-titre positive)</th>
<th>Number of replicate tests</th>
<th>S/Co</th>
<th>Between-condition % CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall reproducibility</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>Between-day</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>Between-operator</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>Between-lot</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>Between-instrument</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
</tbody>
</table>

QC quality control, S/Co signal to cut-off ratio, SD standard deviation, CV coefficient of variation

8.4 **Example table, analytical performance study: interfering substances**

Table 8-5. Summary of test results for determination of analytical specificity: endogenous substances

<table>
<thead>
<tr>
<th>Specimen ID</th>
<th>Test results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unspiked specimen</td>
</tr>
<tr>
<td>ID-1</td>
<td>(Value)</td>
</tr>
</tbody>
</table>
ID identification

Notes to Table 8-5:

- Spiking should be done to a low level of reactivity, to detect subtle effects that may be clinically significant (i.e. that change a result from reactive to nonreactive or vice versa).
- It is sometimes difficult to obtain specimens with elevated levels of a potentially interfering substance; in such cases, exogenous spiking with the potential interferent may be acceptable.
- If there are any conditions for which an effect is observed, these should be noted in the IFU.

8.5 Example table, analytical performance study: cross-reacting infections, diseases or medical conditions

Table 8-6. Summary of test results for determination of analytical specificity: unrelated infections, diseases or medical conditions

<table>
<thead>
<tr>
<th>Infection/disease/medical condition</th>
<th>Number of specimens tested</th>
<th>Test results</th>
<th>Reference test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism 1</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
<tr>
<td>Medical condition 1, etc.</td>
<td>(Value)</td>
<td>(Value)</td>
<td>(Value)</td>
</tr>
</tbody>
</table>

9 Conclusion

The ultimate goal of product analytical and clinical performance studies is to provide scientifically sound data that allow a determination that, in the hands of the end-user, the benefits of using a new IVD outweigh the risks. This goal is achieved by generating analytical and clinical data to support a determination of safety and performance that:

- establishes and supports claims in the IFU (for intended use, by an intended user in an intended setting of use);
- provides instructions about reagents, instruments and specimens;
- details the test method and limitations and warnings of the procedure;
- sets out expected values; and
- supports claims of specific performance characteristics.

Providing the information to WHO for the prequalification of IVDs assessment in a way that is as complete, organized and systematic as possible allows the assessment process to be completed in the shortest amount of time.
10 References

1  GHTF. Clinical evidence for IVD medical devices – key definitions and concepts. Global Harmonization Task Force (GHTF); 2012


4  CLSI. Verification and validation of multiplex nucleic acid assays; approved guideline MM17-A. Wayne, PA: Clinical and Laboratory Standards Institute (CLSI); 2008


10 CLSI. Studies to evaluate patient outcomes; approved guideline GP45-A. Wayne, PA: Clinical and Laboratory Standards Institute (CLSI); 2004


The Technical Guidance Series for submission to WHO Prequalification – Diagnostic Assessment is developed to assist manufacturers in meeting prequalification requirements for their IVD. Further information on this guidance and other Technical Guidance series documents email diagnostics@who.int