UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

Defining and Using Reference Standards for New Diagnostic Tests for Neglected Tropical Diseases

Report of 2023–2024 project

Matthew Denwood, Abbey Olsen
University of Copenhagen

2024-03-21

Content

1 Overview

Control and elimination programs for neglected tropical diseases (NTDs) are currently facing a number of challenges that require urgent solutions, including dependence on imperfect diagnostic tests for routine diagnosis and program monitoring. Several new diagnostic tests are under active development, but there is currently no clear guidance on how the performance of these tests should be evaluated. This project, "Defining Reference Standards for New NTD Diagnostic Tests", was carried out at the University of Copenhagen, with funding from the Task Force for Global Health, to address these challenges. This report provides an overview of concepts and recommended methods for estimating the performance of new diagnostic methods, as well as guidance for how to approach the evaluation of diagnostic tests under the wide range of applications relevant to NTDs. The approaches outlined are influenced by requirements identified by the target product profiles (TPP[1]) produced by the diagnostic and technical advisory group (DTAG) of the World Health Organization (WHO), as well as previous work identified from the wider scientific literature. However, the scope of this report is limited to evaluation of the diagnostic performance of tests; issues relating to improving the performance of tests, logistical challenges of testing in the field, economic/ethical challenges with diagnostic testing, and technical laboratory issues relating to the development of new diagnostic tests are therefore not considered.

We begin with a brief overview of the literature relevant to diagnostic test evaluation from the perspectives of statistical methodology and applications to NTDs. Section 3 provides definitions of key terms that underlie the concepts detailed in section 4. A complete guide to the relevant approaches and statistical methods is given in section 5, and a brief review of the epidemiology, diagnostic test availability, and current targets for control for each NTD is given in section 6. The overall recommendations are summarised in section 7 along with a flow chart in the Appendix.

[1] https://www.who.int/observatories/global-observatory-on-health-research-and-development/analyses-and-syntheses/target-product-profile/links-to-who-tpps-and-ppcs

2 Background

The issue of diagnostic test evaluation arises in practically every field of medicine (Bolboacă, 2019; Chikere et al., 2019; Singh, 2014), veterinary medicine (Branscum et al., 2005; Greiner and Gardner, 2000; Toft et al., 2005), and wildlife disease ecology (Helman et al., 2020). The typical aim is to estimate the test sensitivity (probability of a positive test conditional on a positive disease status) and specificity (probability of a negative test conditional on a negative disease status), as these reflect the information provided by the test, and can be combined with the expected prevalence to generate positive and negative predictive values representing the probability that the individual has a positive status conditional on the test result (Monaghan et al., 2021). Additional metrics have also been suggested to measure the performance of diagnostic tests, including Youden's J (Youden, 1950), which is simply calculated as the sum of sensitivity and specificity minus one, and is a useful metric that can be used to maximise overall test accuracy (Smits, 2010). Sensitivity and specificity can also be expressed as positive and negative likelihood ratios, as well as being combined with the prevalence of disease to provide positive and negative predictive values, which provide more relevant information in a clinical context.

Traditional methods involved the use of a "gold standard" test (or combination of tests) that can be used to generate the presumed truth from which the sensitivity and specificity of new diagnostic tests can be determined (McKenna and Dohoo, 2006). The term "gold standard" originates in the monetary practice of linking paper money directly to gold reserves, which was instrumental in the economic policy of several countries in the late 19th and early 20th centuries (Cooper et al., 1982). The original etymology of a "gold standard test" is therefore a "benchmark" against which other tests can be compared consistently (Cardoso et al., 2014; Claassen, 2005; Versi, 1992), but the term has become somewhat synonymous with the concept of a "perfect test" in some scientific fields (Walter, 1988). This may be because estimates obtained from comparison to imperfect gold standards are known to be biased (Hawkins et al., 2001; Thibodeau, 1981) and are therefore of limited practical utility. The term has also been used to describe a test that is "good enough for a specific purpose" (McKenna and Dohoo, 2006), which implies a degree of subjectivity in when a test can and cannot be considered a "gold standard". However, a similar lack of consensus is also encountered when defining reference tests (Cook, 2012), which suggests that the challenge with this approach is more fundamental than a simple disagreement in terminology. We also note that current

guidelines suggest to use the term ”reference standard” rather than “gold standard” when reporting diagnostic accuracy studies (Kostoulas et al., 2017).

The work of Hui and Walter (1980) demonstrated a method of obtaining unbiased estimates of sensitivity and specificity for two imperfect diagnostic tests simultaneously, which represented a major breakthrough in diagnostic test evaluation (Cheung et al., 2021; Johnson et al., 2019). This approach became known as latent class modelling (Walter, 1988), but is also referred to as evaluation of diagnostic tests “in the absence of a gold standard” or “when there is no gold standard” (Aws et al., 2007; Black and Craig, 2002; Chikere et al., 2019; Jones et al., 2010; Rutjes et al., 2007; Toft et al., 2007, 2005). The original formulation was for two conditionally independent tests in two populations, but the framework has also been extended to multiple populations and multiple tests (in one or more populations), and to relax the assumption that the tests are independent conditional on the latent class (Adel and Berkvens, 2002; Branscum et al., 2005; Georgiadis et al., 2003; Liu et al., 2022; Menten et al., 2008a; Qu et al., 1996; Wang et al., 2017; Yang and Becker, 1997). Methods have also been developed to fit latent class models and latent class regression models in a more general sense (e.g. Linzer and Lewis, 2011), and zero-inflated models can be used to separate the zero observations frequently encountered in over-dispersed count data “true zero” from “false zero” observations (Martin et al., 2005). These models are available in general-purpose statistical modelling packages (Brooks et al., 2017) and have also been applied to NTDs (Bakuza et al., 2017).

An important limitation of almost all latent class models is the requirement for data from multiple populations with varying true prevalence. Consistent diagnostic test performance across these populations is an important assumption for the standard models (Toft et al., 2005), although it is possible to modify the model to relax some of these (Stærk-Østergaard et al., 2022). However, mis-specified latent class models can also result in biased estimates and/or a lack of identifiability when complex dependence structures are fit to datasets with a limited number of observations (Albert and Dodd, 2004), and it is strongly recommended to follow standardised guidelines for reporting of results from studies implementing such models in order to avoid the possibility of erroneous interpretation (Kostoulas et al., 2017). Methodological development in the field of latent class models is currently a very active area of research, in particular different formulations of conditional dependence models and the use of traditional Hui-Walter model formulations compared to

formulations using random effects to constrain the number of parameters (Keddie et al., 2023; Wang et al., 2017).

The evaluation of diagnostic tests is also highly relevant to NTDs, although the typical use of diagnostic tests varies widely depending on the disease. For some NTDs, mass drug administration (MDA) is implemented within communities without individual diagnoses, whereas other diseases require intensive disease management due to factors like the need for individual diagnosis, the complexity of treatment, or the lack of a suitable drug for mass administration (Gass, 2020; Macpherson et al., 2015; Peeling and Mabey, 2014; Taylor, 2020). Accordingly, there are various diagnostic tools available for the detection of NTDs, ranging from microscopy, immunoassays and molecular methods to specific laboratory techniques for identifying pathogens/analytes in samples (Peeling and Mabey, 2014; Taylor, 2020; Bharadwaj et al., 2021; Choi et al., 2022). New diagnostic tools are also being developed to improve accuracy, speed, cost-effectiveness, and field usability, and there is a pressing need to evaluate the performance of these tests so that they can be evaluated for use as part of eradication and control programs and/or in clinical settings. This process faces many NTD-specific challenges, for example sensitivity of diagnostic tests for microparasitic infections has been shown to vary across individuals and settings as a function of infection intensity (Coffeng et al., 2023; Kazienga et al., 2022).

The motivation for developing new diagnostic tests also varies among NTDs; in some cases, practical considerations such as a lack of availability of tests or the desire for patient-side tests is the predominant driving force, while in others, the low sensitivity of existing reference tests has motivated the development of more sensitive tests. For example, current tests for Buruli ulcer are largely based on the identification of *Mycobacterium ulcerans* in smear microscopy, culture, histopathology, and PCR, which are not readily available or implementable in endemic regions (van der Werf, 2018; Röltgen et al., 2019). For Schistosomiasis, a stool smear test is the standard method applied in epidemiological surveys, but a urine-based rapid assay to detect circulating antigen has also been developed. Although this test has shown lower specificity and discordant results compared to the standard stool smear test (Coulibaly et al., 2013; Straily et al., 2022), the operational utility of a rapid assay could be highly advantageous even despite a reduction in diagnostic performance. Ethical and cultural challenges with existing tests can also be a motivating factor for developing better-tolerated alternatives (Orish et al., 2022). Several NTDs also rely on

diagnostic tests based on clinical observations, with variable performance dependent on the experience of the clinician (Walker et al., 2020). A further potential motivation for the development of new tests is therefore to focus on more easily standardisable approaches, although the performance of clinical tests can also be improved by standardisation exercises (Engelman et al., 2020, 2018). For dracunculiasis, traditional testing approaches using poorly specific PCR techniques required species confirmation by DNA sequencing, which can take several weeks to process. Although substantial reductions in prevalence have been achieved in endemic countries through surveillance and control programs (Lemma et al., 2020), delays in obtaining test results can present a challenge to disease eradication. In contrast, a new assay has removed the need for confirmatory DNA sequencing, thereby generating results within 24 hr (Coker et al., 2022). However, the absence of standardised methods for evaluating these new tests has contributed to a relatively slow adoption of new diagnostic approaches. For example, the World Health Organization recommendations for the control of schistosomiasis rely on identification of schistosome eggs in stool or urine samples despite the development and recent availability of additional diagnostic approaches (Chala, 2023; Utzinger et al., 2015).

There are a large number of studies that have attempted to evaluate the performance of new tests for NTDs. For example, the sensitivity and specificity of two rapid diagnostic tests for leishmaniasis were estimated for clinically suspected cases using microscopic examination as an assumed perfect reference standard (Eyayu et al., 2022); this study also used the kappa statistic (Cohen, 1960; McHugh, 2012) to investigate agreement between the test results. The approach of evaluating against an assumed perfect reference standard test such as microscopy or PCR has also been applied in other studies for leishmaniasis (Mbui et al., 2013; Pedrosa et al., 2013; Ayelign et al., 2020; Salam et al., 2021; Rezaei et al., 2022). Diagnosis of trachoma is typically made using clinical diagnostic indicators based on the WHO grading system, but other methods such as microscopy, immunoassays, culture, and molecular methods are also used (Solomon et al., 2004; Meyer, 2016), and some studies have evaluated diagnostic tools such as nucleic acid amplification by using PCR as an assumed gold standard (Yang et al., 2009; Harding-Esch et al., 2011). Newly developed diagnostic tools for Buruli ulcer have also been compared with assumed gold standards such as PCR (Herbinger et al., 2009), although the estimated sensitivity and specificity vary substantially depending on the chosen reference standards, and a high probability of misclassification has been demonstrated (Amewu et al., 2022). Meta-analysis has also been used to summarise estimated test

performance based on published studies (Gurung et al., 2019), but these methods cannot correct for bias underlying the original published estimates. Composite reference tests have been used for many NTDs in situations where a combination of results from individual tests was assumed to represent a perfect reference standard, including for leishmaniasis (Humbert et al., 2019) and schistosomiasis (Hoekstra et al., 2022), as well as Buruli ulcer (Amewu et al., 2022), where expert panels have also been used as a form of composite reference test (Eddyani et al., 2018). However, work evaluating a composite reference standard for trachoma concluded that combining serial interpretation of individual tests into a single composite reference test result unnecessarily discards information and is not guaranteed to lead to improved test performance unless all tests have perfect specificity. The study therefore recommended the use of alternative statistical methods rather than relying on composite reference standards (Schiller et al., 2016). We also note that composite reference standards often include several related tests with strongly correlated results, which results in a combined diagnostic ability that is substantially worse than would be expected from fully independent tests.

Latent class models have also been utilised in the context of NTDs. Boeleart et al. (1999) used a latent class model to estimate the performance of a direct agglutination test for leishmaniasis compared with clinical data and lymph node aspirates, and further work has also been done using latent class models for this disease (Menten et al., 2008b; Machado de Assis et al., 2012). Boeleart et al. (2004) compared a number of tests for leishmaniasis to direct-observation tests using a latent class model, and showed that these estimated specificities were substantially higher than those derived using a direct-observation test as an assumed-perfect reference standard. This study concluded that the two immunological tests could potentially replace the standard parasitological test; this conclusion would likely not have been made based on the biased estimates of specificity obtained from a simple comparison to a microbiological test used as a reference test. Latent class models have also been used to overcome challenges in implementing laboratory-based methods or diagnostic tools for Buruli ulcer in the field. Mueller et al. (2016) used a latent class model to classify patients into groups based on laboratory test results, which provided results that were then used to generate a clinical prediction score to guide treatment decisions. Studies using latent class analysis to estimate the sensitivity and specificity of PCR and two clinical diagnostic indicators for trachoma showed that PCR is more accurate than clinical examination in identifying trachoma, and that the clinical test based on WHO guidelines lacks specificity (See et al., 2011; Koukounari et al.,

2013). Latent class modelling has also been applied to compare the performance of newly developed immunological assays to existing diagnostic methods (Wiegand et al., 2018), and to schistosomiasis, where the diagnostic performance of urine-based rapid assays was estimated against direct-observation and PCR results (Fuss et al., 2018). A more recent study evaluated the performance of 11 different diagnostic tests for schistosomiasis, including parasitological tests, nucleic acid amplification tests, and immunological methods, on different sample matrices (faeces, blood, and urine) originating from low endemicity areas (Mesquita et al., 2022). Latent class models have also been used to determine the performance of different diagnostic tests for *Strongyloides stercoralis,* as well as to determine an optimal cut-off value for one of the diagnostic tests being evaluated (Tamarozzi et al., 2023).

3 Definitions and Key Concepts

LATENT CLASS MODEL (LCM): a general class of statistical methods that, in the context of this report, refers to models designed to estimate the performance of multiple diagnostic tests concurrently without an assumption that any of the tests are perfect. Latent class models can be formulated and implemented in different ways, and current best statistical practice should be expected to change over time. We take the approach of describing these in the general sense so that the underlying concepts can be translated to different implementations of the general class of models.

TARGET CONDITION: a binary representation of the condition of interest, which reflects the desired practical use of the test and forms the basis for interpretation of sensitivity and specificity. Depending on the context, this may reflect clinical disease, infection status, or previous exposure to a pathogen – multiple target conditions may even be used for the same NTD but test performance should be expected to vary between these. Target conditions based on continuous measures such as infection intensity for microparasitic infections are also valid but must be converted to a binary classification in order to be used in the context considered here. A clear statement of the target condition is an essential requirement when reporting results of diagnostic test evaluation.

TARGET POPULATION: the conceptual population representing the context in which the diagnostic test will be used in practice, which (along with the target condition) provides a basis for interpretation of sensitivity and specificity. In the context of NTDs this will almost always reflect populations with endemic disease - alternative target populations may also be used in some situations but estimates of test performance should be expected to vary between target populations. A clear statement of the target population is an essential requirement when reporting results of diagnostic test evaluation.

CASE DEFINITION (OR LATENT CLASS): in the context of LCMs, the case definition is the implicitly defined "true positive status" that is defined by the model based on the combination of diagnostic tests (and samples) included in the study. As this is effectively unknown to the user, it is frequently referred to as the "latent class", but we primarily use the terminology "case definition" to reduce the use of statistical jargon. Understanding the case definition is a critical part of interpreting results of LCM; mis-matches between the stated case definition and target condition, as well as mis-matches

between the samples included in the study and the stated target population, result in biased interpretation of the sensitivity and specificity estimates produced. When estimating test performance by comparison to an assumed-perfect reference test, then the case definition is implicitly defined as the result of the reference test. A thorough analysis of the case definition is an essential requirement when reporting results of diagnostic test evaluation – we provide a framework for doing this in section 4.

CASE AND NON-CASE: an individual that is defined as positive (case) and non-positive (non-case) according to the case definition. In the context of LCM, case vs non-case is also a latent quantity i.e. cannot be observed directly.

SENSITIVITY: the probability of a positive test result given a case.

SPECIFICITY: the probability of a negative test result given a non-case.

YOUDEN'S J: an index calculated as the sum of sensitivity and specificity minus one, representing one method of summarising the diagnostic accuracy of the test to a single number. Youden's J ranges on a scale from 0 (test has no diagnostic information) to 1 (a perfect test).

POSITIVE/NEGATIVE LIKELIHOOD RATIO: an alternative presentation of sensitivity and specificity, given by $LR_+ = sensitivity/1 - specificity$ and $LR\_ = 1 - sensitivity/specificity$

TEST ACCURACY: we use this generic term as shorthand for "sensitivity and specificity" (and therefore also Youden's J and positive/negative likelihood ratios) to indicate the overall diagnostic capability of a test.

CONDITIONAL INDEPENDENCE: where the probability of observing a positive test result from one test is unaffected by the observed result of another test, conditional on the case status. This is typically assumed with standard latent class models.

CONDITIONAL DEPENDENCE: where two tests are correlated due to a commonality, so that the probability of observing a positive test result from one test is affected by the observed result of

another test, even after accounting for the case status. The commonality may either be analytical target (e.g. detection of eggs or IgG antibodies), or due to sampling structure (e.g. re-use of the same sample rather than taking independent samples for each test). It is more technically correct to use the term "lack of conditional independence", but we also use the shorter term "conditional dependence" for convenience. It is important to note that a lack of conditional independence reduces the combined test accuracy of two tests applied to the same individual. A biologically motivated analysis of potential sources of conditional dependence is an essential step of using LCM and reporting results of diagnostic test evaluation.

PREVALENCE (OR TRUE PREVALENCE): the unknown proportion of individuals/samples in the group/population sampled that are a case. The prevalence is constant for all diagnostic tests applied to the same individuals.

OBSERVED PREVALENCE: the observed proportion of individuals/samples in the group/population sampled that have tested positive using a specified test. The observed prevalence may vary between diagnostic tests applied to the same individuals due to differences in performance of the diagnostic tests.

POSITIVE/NEGATIVE PREDICTIVE VALUE (PPV AND NPV): the probability that a positive/negative test result reflects a individual with case definition status positive/negative, reflecting both the performance of the test and prevalence in the population. PPV and NPV are more useful in a clinical context than sensitivity and specificity because they provide a direct interpretation for the probability that an individual is truly positive. PPV and NPV are calculated using Bayes' theorem:

$$PPV = \frac{sensitivity\,.\,prevalence}{sensitivity\,.\,prevalence + (1 - specificity)\,.\,(1 - prevalence)}$$

$$NPV = \frac{specificity\,.\,(1 - prevalence)}{specificity\,.\,(1 - prevalence) + (1 - sensitivity)\,.\,prevalence}$$

REFERENCE STANDARD: the best-performing diagnostic test available for a specific case definition, with known (but likely imperfect) sensitivity and specificity relative to the target condition.

PERFECT TEST: a test with perfect performance, i.e. sensitivity of 100% and specificity of 100%.

GOLD STANDARD: an arguably ambiguous term that may mean "reference standard" or "perfect test", depending on context. We follow other authors in recommending the use of these more clearly defined terms instead (Kostoulas et al., 2017).

STARD-BLCM: this stands for "Standards for Reporting of Diagnostic Accuracy Studies that use Bayesian Latent Class Models", which refers to guidelines designed to enhance the quality and transparency of reporting in studies that use latent class models to estimate diagnostic test performance (Kostoulas et al., 2017) .

95% CI: this may refer to either 95% confidence intervals (for frequentist methods) or 95% credible intervals (for Bayesian methods), which represents uncertainty in the corresponding estimate of a given parameter (such as sensitivity, specificity, or Youden's J). Appropriate calculation and communication of statistical uncertainty (using 95% CI and other mechanisms) is essential for all statistical analyses, including methods for diagnostic test evaluation.

4 General Considerations

This section introduces generic issues that are common to the majority of use cases for NTDs as a basis for consistent definition and interpretation of diagnostic test performance.

4.1 Illustrative tests and scenarios

We begin by defining the five theoretical test types and their performance characteristics (Table 4.1) and six scenarios that we will use for illustrative purposes (Table 4.2).

Table 4.1: Test characteristics used for illustrative purposes.

TEST G1	A hypothetical perfect test with 100% sensitivity and 100% specificity. This is included for illustrative purposes only; we note that a perfect test is not expected to exist in practice.
TEST A	A reference test with high specificity but potentially poor sensitivity, such as a direct-detection test for an adult parasite. Imperfect sensitivity is primarily due to the absence of eggs in the faecal sample taken from an infected individual. Specificity may be less than 100% due to mislabelling of samples, misidentification of the pathogen, or both.
A1:	*A relatively well-performing test with 80% sensitivity and 99% specificity*
A2:	*A poorer test with 50% sensitivity and 99% specificity*
TEST B	A reference test with relatively high sensitivity but lower specificity than TEST A, such as an antibody test. Imperfect sensitivity may be due to the absence of an immune response in some infected individuals, and imperfect specificity may be due to the presence of antibodies in uninfected individuals (either vaccinated, cross-reacting, or previously infected). TEST B is conditionally independent of TEST A because the targets of the two tests are completely distinct.
B1:	*A relatively well-performing test with 90% sensitivity and 95% specificity*
B2:	*A poorer test with 80% sensitivity and 90% specificity*
TEST C	A new test with characteristics to be estimated. TEST C is conditionally independent of both TEST A and TEST B, meaning that TEST C is designed to detect a target that is unrelated to the presence of either the pathogen or the antibody target used by TEST B. An example might be a test to detect juvenile parasites, where this would not also be expected to detect adult parasites.
C1:	*A relatively well-performing test with 90% sensitivity and 95% specificity*
C2:	*A poorer test with 80% sensitivity and 90% specificity*
TEST D	A new test with characteristics to be estimated. TEST D is conditionally independent of TEST A but is correlated with TEST B because it is designed to detect another aspect related to the immune response of the individual. An example might be a second antibody test. We do not consider conditional independence of TEST D and of TEST C because our theoretical scenarios do not use these tests concurrently.
D1:	*A test with 80% sensitivity and 90% specificity and a strong correlation with TEST A*
D2:	*The same test (80% sensitivity and 90% specificity) but where the correlation with TEST A is weaker (but non-zero)*

The theoretical possibilities for evaluating diagnostic tests across all NTDs are virtually limitless, but we chose the test performance profiles and scenarios described in Tables 4.1 and 4.2 to define the particular examples that we will refer to throughout this report. Although biological interpretations are given for the sake of clarity, the reader is reminded that these are for illustration only, and other biological scenarios would also fit into these test types. However, the note regarding conditional independence of tests A and B is particularly important: these two tests must be regarded as reflecting completely different aspects of a "case", so that their only commonality is the status of the individual according to this case definition (i.e. they are independent conditional on the latent state implicitly used by the model).

SCENARIO 1	Comparison of test C to test G
SCENARIO 2	Comparison of test C to test A or B
SCENARIO 3	Comparison of test C to tests A and B
SCENARIO 4	Comparison of test D to test A
SCENARIO 5	Comparison of test D to test B
SCENARIO 6	Comparison of test D to tests A and B

Table 4.2: Test scenarios used for illustrative purposes.

Each of the scenarios can be further subdivided according to the number of distinct populations available (either one, two or ≥three, where the performance of each test is consistent across populations). Additional complexity may also be encountered in practice due to additional dependencies (for example if test D may be correlated with test A and/or test B), the availability of more than three test results per individual, and the use of different tests in different populations. We explore these issues together with the practical implications of varying test performance among populations later in this section.

4.2 Conditional dependence and the case definition

Although estimates obtained from latent class models are in general unbiased with respect to diagnostic test performance, they are highly susceptible to bias resulting from a failure to properly consider conditional dependence of the available tests and the correspondence between the desired target condition and case definition implicitly defined by the model. Unfortunately, this important source of bias is frequently overlooked when interpreting results obtained from these models,

particularly when models are implemented by relatively inexperienced statistical practitioners - this issue was a strong motivating factor for developing the STARD-BLCM guidelines (Kostoulas et al., 2017). We recommend using directed acyclic graphs (DAGs) as a way of exploring the issue in biological terms, as outlined below.

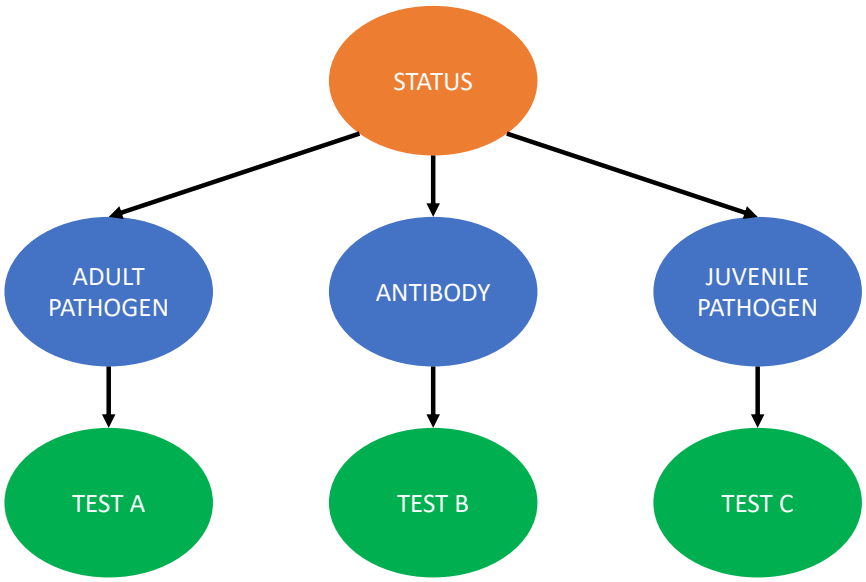

Figure 4.1: Directed acyclic graph illustrating the connection between three conditionally independent tests (data obtained under Scenario 3). Observed test results are shown in green, unobserved intermediate states are in blue, and the implicit latent state is in orange.

The DAG displayed in Figure 4.1 illustrates the relationship between Tests A, B, and C, which are conditionally independent because each target is a distinct aspect of the target condition. A latent class model applied to data obtained from Scenario 3 would therefore be expected to return unbiased estimates for the performance of these tests relative to the latent state broadly intersecting the presence of adult parasite, juvenile parasite, and antibody production. In contrast, the DAG displayed in Figure 4.2 illustrates the relationship between Tests A, B, C, and D, where there is a conditional dependence between Tests B and D arising from their common target related to the immune response. A latent class model applied to data obtained from Scenario 5 would therefore produce biased estimates of the performance of all tests relative to the latent state identified above unless this correlation is accounted for. The bias arises due to the increased correlation between tests B and D: the stronger agreement arising due to a common detection of antibody status is

incorrectly inferred as increased performance of both tests. In turn, the performance of Tests A and C are underestimated due to the observed lower agreement with tests B and D.

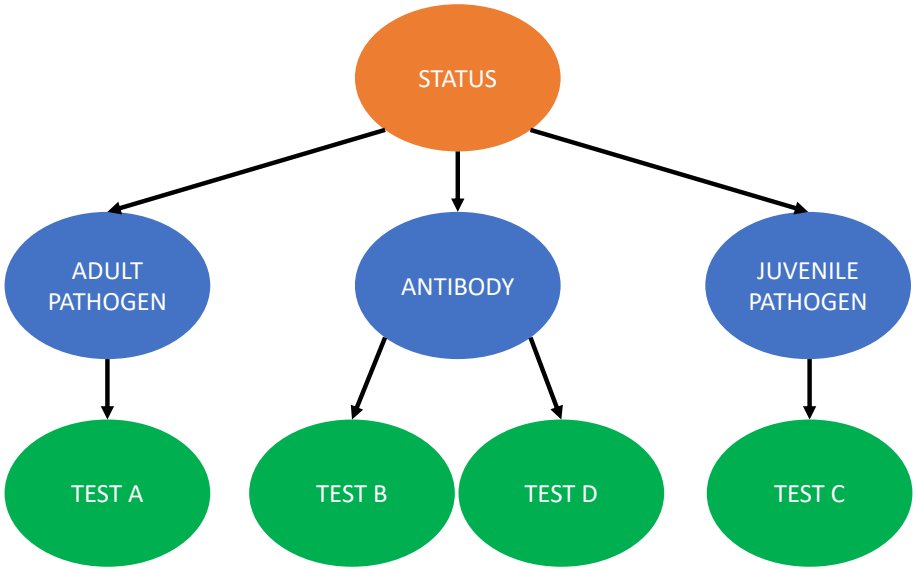

Figure 4.2: Directed acyclic graph illustrating the connection between four tests, where Tests B and D exhibit a correlation, i.e. they are conditionally dependent (data obtained under Scenario 6). Observed test results are shown in green, unobserved intermediate states are in blue, and the implicit latent state is shown in orange.

The unobserved intermediate states in these DAGs suggest that the interpretation of the latent class is the intersection between adult parasite, juvenile parasite, and antibody production. We note that this is a very broad case definition that may not suit all applications, but it is at least consistent between these two DAGs, and latent class models fit to both datasets would be expected to produce equivalent results as long as the conditional dependence is accounted for. The DAG shown in Figure 4.3 illustrates the change in the latent state caused by restricting the range of conditionally independent tests used by the model to a subset of those used in Figure 4.2. Fitting a latent class model to data from the two antibody tests B and D results in a change to the latent state implicitly used by the model away from the broadly defined case definition and towards a more specific case definition based on seropositivity. An important consequence of this change to the latent state is that the interpretation of the sensitivity and specificity changes; in both cases the apparent performance of Tests B and D would be expected to increase relative to estimates obtained from Scenarios 3 or 6. This is due to the difference between the probability of detecting antibody conditional on (1) a

relatively broad case definition and (2) a case definition based on immune response. Where the two hypothetical tests are run based on the same physical blood sample, the change to the latent class (and resulting increase in estimated performance of the tests) is further amplified.

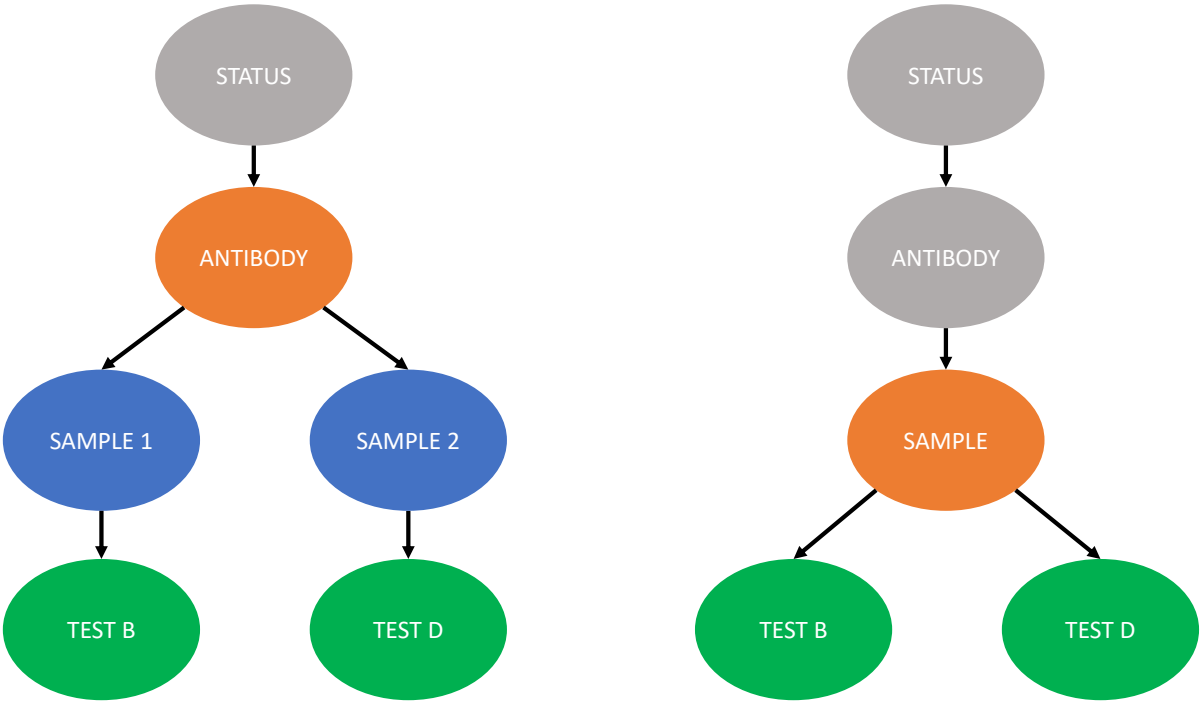

Figure 4.3: Directed acyclic graphs illustrating the implicit case definition arising from the use of Tests B and D (data obtained under Scenario 5), where the two tests are either based on independent samples (left) or the same sample (right). Observed test results are shown in green, unobserved intermediate states are in blue, the implicit latent state is shown in orange, and inestimable states are shown in grey.

Although Tests A and C are defined to be conditionally independent for the purposes of this report, it could be argued that in reality the two tests exhibit correlation due to the temporality of the infection process. In this case, exclusion of the test for juvenile pathogens modifies the latent state towards more long-term or historical infection, whereas exclusion of the antibody test modifies the latent state towards more recent infection (Figure 4.4). As for the DAG shown in Figure 4.3, this change to the implicit latent state will also affect the estimated performance of the tests.

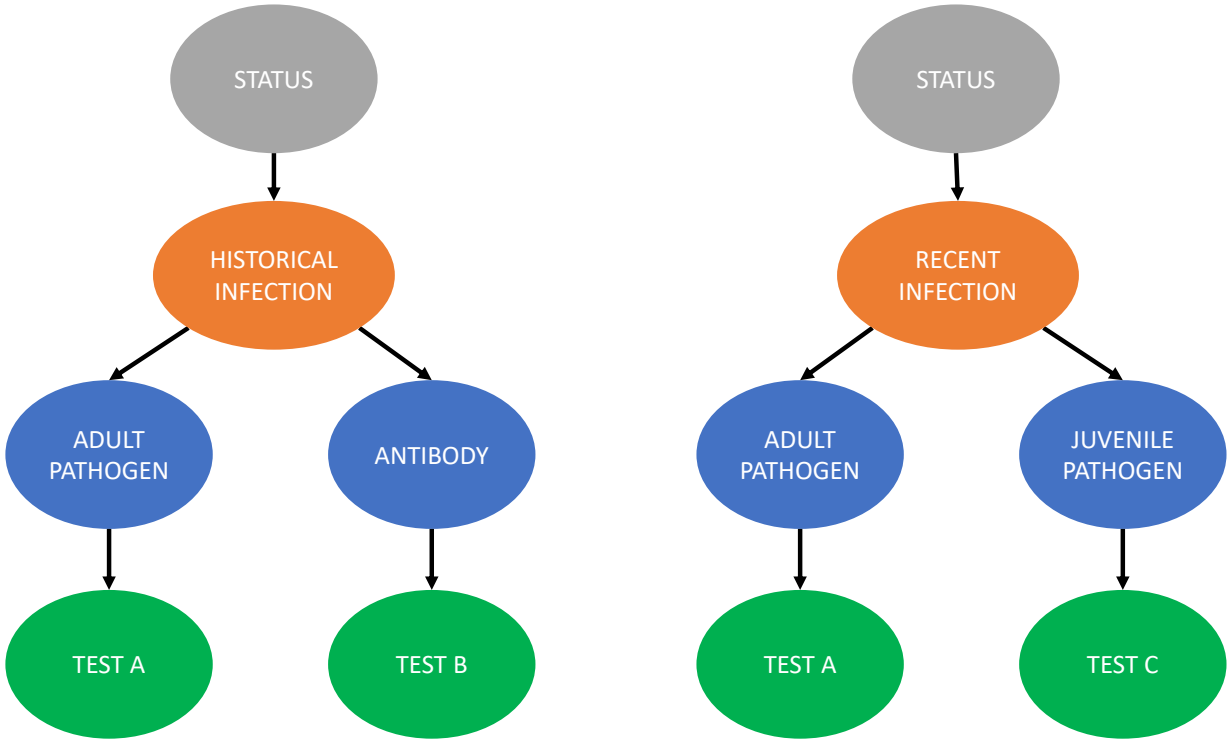

Figure 4.4: Directed acyclic graphs illustrating the change in implicit latent state due to the exclusion of results from Test C (left) and Test B (right) from the latent class model (data obtained under Scenario 3). Observed test results are shown in green, unobserved intermediate states are in blue, the implicit latent state is shown in orange, and inestimable states are shown in grey.

As shown in the illustrative examples above, careful consideration of the latent class implicitly defined by the model is essential in order to be able to interpret the results of latent class models. Although modelling software will typically report the estimates as “sensitivity” and “specificity”, the onus is on the user to remember that the estimates actually reflect the probability of observed test outcomes conditional on the latent state used by the model, so if this does not match the desired case definition then the estimates will be biased. This bias can only be corrected by modifying the profile of diagnostic tests included in the study and must be considered before collecting data. A further source of bias is failure to account for conditional dependencies between tests, although in most circumstances this can be corrected by adjusting the model specification to match the biology of the system. Wherever possible, we recommend the specification and use of a standardised set of tests in order to ensure that case definitions are consistent across studies. Where this is not currently practicable, we strongly recommend the use of DAGs as a way to help visualise the system being studied.

4.3 Varying case definition

Section 4.2 illustrates that the choice of diagnostic test combination determines the latent class estimated by the model, and therefore the estimated prevalence, sensitivity, and specificity. However, diagnostic tests should be selected based on the case definition (within the confines of practical considerations such as the availability and cost of tests) in order to ensure that this matches the implicit latent class as closely as possible. Different case definitions will naturally arise from different applications of diagnostic tests, whether this is individual-level testing to determine treatment decisions, community-level surveillance to determine the prevalence at different stages of pathogen eradication, or ultimately establishing freedom from disease. The case definition is also highly dependent on the type of population under consideration (e.g. artificial samples *vs.* field samples, and hospitalised patients *vs*. the wider community in an endemic setting). It is therefore extremely important that the case definition be included when reporting diagnostic test performance, because the sensitivity/specificity should always be assumed to change when the case definition changes.

Testing for soil transmitted helminths (STH) is one example where the case definition will change according to the phase of control/eradication. During Phase 1, the infection intensity can be expected to be high, which will result in a higher sensitivity of egg detection methods and an associated case definition of "high-intensity infection". As the parasite burdens reduce in response to mass drug administration during Phases 2 and 3, sensitivity of the same egg detection methods will be reduced because fewer eggs are expected in the stool of infected individuals, corresponding to a case definition of "low-intensity infection". During Phase 4, surveillance for the parasites may rely more intensively on antibody detection methods (perhaps using children as a sentinel population), so the case definition will shift towards "seroprevalence". Each of these case definitions vary in terms of the desired profile of diagnostic tests and expected performance of these tests.

We have so far only discussed case definitions that are relevant to diagnostic tests applied in the field, but case definitions based on biobank and/or artificial samples are also potentially valid choices, depending on the application. One example would be to aid in the initial development and preliminary evaluation of diagnostic tests based on reference samples, although the performance of the diagnostic tests will typically be much better for these reference samples than would be

expected for a case definition based on samples taken in the field. Another example would be to estimate the performance of tests for diseases that are nearing elimination thresholds, where it may be almost impossible to find a sufficient number of cases in the field. In these situations, it may be necessary to rely on a case definition based on biobank samples, with the same caveat that sensitivity may be higher than for a case definition based on field data due to the fact that biobank samples are typically selected for high positivity based on existing tests.

Due to the importance of the factors discussed above, the objective of a test evaluation study should also be considered as being intrinsically linked to the case definition. Consequently, this is a key aspect of the STARD-BLCM guidelines. We define two general objectives that we believe cover the majority of use cases for NTDs in Table 4.3; Objective 1 will be relevant whenever the performance of the reference tests themselves must be estimated, whereas Objective 2 allows for easier comparison of sensitivity and specificity estimates between studies where either a standardised reference test or a standardised panel of reference tests is used consistently.

OBJECTIVE 1	Concurrent estimation of the performance of two or more tests where neither test has previously been evaluated for this target condition, or where available estimates are subject to substantial uncertainty either due to small sample sizes or the use of sub-optimal statistical methods.
OBJECTIVE 2	Estimation of new tests relative to one or more well-characterised reference tests, where reliable estimates of the performance of the reference test(s) (obtained using appropriate statistical methods) are available for this target condition.

Table 4.3: Possible study objectives for use with NTDs.

4.4 Types of diagnostic test

Diagnostic tests for NTDs can be broadly subdivided along three axes:

1. The sample matrix required, i.e. blood, skin, faeces, etc.
2. Processing requirements, i.e. tests requiring highly specialised laboratory equipment, tests requiring basic laboratory equipment (e.g. microscopes), patient-side tests requiring multiple visits (e.g. diethylcarbamazine patch test for onchocerciasis), and patient-side rapid tests that provide an immediate result.

3. The degree of standardisation of the test, i.e. reliance on skills/experience of operator
4. The nature of the result provided by the test, i.e. qualitative, semi-quantitative, or count-based.

Both sample matrix and processing requirements strongly impact the practical utility of the test, but these considerations are outside the scope of this report. Test standardisation is extremely important and can be expected to vary considerably depending on the nature of the test (e.g. clinical scoring *vs.* patient-side rapid test kit), expected degradation of the sample matrix during transport/processing, as well as the degree of technical variation between equipment and reagents etc. used as part of the test. These issues are discussed further in Section 4.4. In the following, we discuss how the fundamental nature of the test result can affect interpretation, and the relevance of this for defining reference standards.

Qualitative tests

Qualitative test results can be binary, or on an ordinal scale with categories ranging from e.g. "strongly negative" to "strongly positive", which may be based on a scoring system leading to an overall grading scale. Examples of qualitative tests include most patient-side rapid test kits, microbiological presence/absence tests, and most tests based on clinical scoring systems. For test results on ordinal scales and using scoring systems, clear rules must be given for how to map the underlying scale or scoring system onto a dichotomous (negative/positive) interpretation that should be used to make clinical/operational decisions. This includes the special case of negative/inconclusive/positive, where we recommend that inconclusive results should typically be classified as positive. This is consistent with the idea that a diagnostic test is an aid to make a decision, so inconclusive results are useless because they do not allow a decision to be made. By re-classifying inconclusive results as positive (or negative) this allows a decision to be made, albeit at the cost of reduced specificity (or sensitivity). Alternatively, the test procedure can be defined so that inconclusive results are re-tested, if the end result of the procedure is always a valid binary classification. It is important to note that there is an implicit trade-off between sensitivity and specificity in the choice of mapping between ordinal scale/scoring system and binary outcome, and the trade-off may be chosen to prioritise sensitivity over specificity or vice versa, depending on the situation. It is therefore advisable to use different mapping systems for different situations, but the

overall test should be named differently to correspond to the precise mapping used in order to avoid ambiguity.

Semi-quantitative tests

Results from semi-quantitative tests have some kind of numeric connection to the intensity of infection within the individual being tested, so that values at the extreme ranges of the test results can be interpreted as representing stronger evidence of a case or non-case. However, this relationship is typically non-linear and highly complex, and methodological research into how best to analyse the continuous outcome of these tests is ongoing. Examples of semi-quantitative tests include ELISA optical densities, PCR cycle thresholds, and gene copy estimates produced by qPCR. In general, there are two sources of variation: technical sources reflecting slight variations in laboratory conditions between test runs; biological sources, including concentration of the analyte in the sample matrix. Some test types come with additional considerations, such as the imperfect correspondence between the number of gene copies and number of pathogens (or pathogen eggs) for qPCR. The sensitivity of tests measuring antibody status are also affected by the dynamic nature of circulating antibody concentrations during active infection and following elimination of the pathogen. Results from semi-quantitative tests are interpreted by dichotomising the outcome based on a pre-specified cut-off value. This cut-off value is typically chosen to maximise the overall performance of the test, but an explicit trade-off is made between sensitivity and specificity that may not be suitable as a "one size fits all" approach. It is therefore advisable to provide different cut-off values for the same procedure so that e.g. sensitivity can be prioritised over specificity (or vice versa) to suit the purpose for testing, although changing the cut-off value should be considered a change to the overall test procedure for operational purposes. Clear agreement should also be reached on how to implement the cut-off in terms of the procedure for rounding of the observed value and the expected interpretation when the (rounded) value is exactly equal to the cut-off value specified.

Count-based tests

Several tests for parasitic infections (including many NTDs) rely on the microscopic enumeration of discrete units of either parasites (adults or larvae) or their eggs. The result may be reported as the exact count observed (or a derived quantity, such as the number of eggs per gram of faeces for STH) or simply a binary indication of the absence/presence of one or more parasites/eggs, in which

case the outcome can be considered as qualitative. Tests involving direct examination of tissue samples for microbiological organisms also typically fall under this category. Where counts are recorded, a binary interpretation of negative/positive can be derived by applying a cut-off, such as zero *vs.* non-zero counts (although higher cut-off values can also be employed to distinguish low-intensity from high-intensity infection, for example). These tests can be considered as a special case of (semi-)quantitative tests but have a number of specific features that merit separate discussion. It is important to note that the expected range of counts depends on a number of factors, including:

(1) the presence/absence of the pathogen
(2) the distribution of parasites/eggs within and between hosts (both mean and variation)
(3) the volume of sample matrix examined
(4) additional biological/laboratory factors such as faecal consistency, preservation, and correct identification of parasites/eggs in the sample, as well as technical variation in the laboratory process.

All of these factors should be considered to form part of the overall diagnostic performance of the test, so that the probability of detecting one or more parasites/eggs from an infected individual depends on the intensity of infection within the community. As a result, the sensitivity will vary across different clinical settings and endemicity levels, regardless of the cut-off value chosen. The relationship between the overall sensitivity (for a given cut-off value) and the distribution of counts in infected individuals can be defined using a parametric distribution such as the negative binomial, but this requires a strong (and typically unverifiable) assumption that the true distribution of counts matches the parametric distribution and parameterisation chosen. More sophisticated methods of analysing the raw count data, for example the use of zero-inflated models, may offer solutions to some of these problems. However, these models also make strong assumptions regarding the true distribution of counts from infected individuals and represent an ongoing area of methodological research.

Composite tests

Composite tests represent a combination of multiple individual tests, which is expected to increase the overall performance of the diagnostic procedure relative to any of the tests individually. A major challenge is in the choice to interpret the test results in serial (in which case specificity is increased

at the cost of sensitivity), parallel (in which case sensitivity is increased at the cost of specificity), or a more complex decision tree based on different test result combinations (that may weight sensitivity and specificity more evenly). In addition, collapsing the observed combination of individual test results to a single value loses valuable information that could be leveraged by an appropriate statistical model (Schiller et al., 2016). Although composite tests are relatively easy to interpret and can have improved performance compared to any single test, we therefore recommend that the full individual-level test results be preserved as the primary end point of the testing process and starting point of analysis.

4.5 Consistency of test performance

One of the key assumptions underlying the use of latent class models for diagnostic test evaluation is that the performance of diagnostic tests is consistent across the different populations used for the analysis. By extension, when comparing a new test to one or more well-characterised reference test (Objective 2) it is also important that the performance of the reference test(s) is consistent between the study population and the population in which they have previously been characterised. The degree of standardisation can be expected to vary considerably depending on the nature of the test (e.g. clinical scoring *vs* patient-side rapid test kit), the combination of varying transport times and expected degradation of the sample matrix during transport/processing, as well as degree of technical variation between equipment and reagents etc used as part of the test. It is therefore essential to ensure that laboratory methods are standardised, rapid patient-side test kits are from the same batches, and subjectivity in qualitative clinical tests is minimised by standardising criteria for clinical scoring systems. Where possible, balancing the assignment of laboratories/kits/clinicians between populations is also recommended.

It is equally important to consider the potential impact of the populations themselves on diagnostic test performance. Populations are commonly defined based on discrete geographical areas, although in theory these can be defined on any basis that does not involve the test results themselves, such as individuals from households with a known infectious contact *vs* randomly selected from the wider community, or by using known risk factors as proxy indicators for elevated *vs* reduced probability of infection. However, care should be taken to avoid the situation where the selection criteria for the population will also lead to increased sensitivity and/or specificity of one or more test in one population relative to the others, which may occur where either the severity of infection intensity

(spectrum bias) and/or cross-reactions with different disease processes (difference in analytical specificity) may vary systematically between populations. One example where this would be expected to occur is defining a population based on presence of obvious lesions, or separating a community based on demographic characteristics that are known to influence performance of diagnostic tests (such as children *vs* adults for several NTDs including STH). Of particular importance to several NTDs is the potential for inconsistent specificity due to cross-reactions with pathogens that may be present at one site but absent at another. Another relevant example is the potential for sensitivity of egg detection methods to vary based on consistency of stool, which could be a problem if populations have a different expected stool consistency.

In the next section, we present methods of validating the assumption of constant sensitivity and specificity across populations. However, these methods are more useful in situations where multiple populations are available, as it is possible to systematically exclude each population in turn and still have sufficient information to run the model. For this reason, we recommend that prospective study designs include a minimum of three populations wherever possible, and that the populations be chosen based on any criteria that are expected to result in varying prevalence with consistent diagnostic test performance. Where this is not possible, inclusion of two populations gives substantially more flexibility in analysis method than a single population, and a minimum of three diagnostic tests will be required in situations where only a single population is available.

5 Recommended Statistical Methods

This section outlines three statistical approaches that can be used to analyse data according to the scenarios and objectives presented in Section 4.1. Due to the wide range of potential applications, it is not possible to recommend a single "one-size-fits-all" approach, but efforts have been made to streamline the options as far as possible. Worked examples of analyses using simulated data are provided via our website (https://www.costmodds.org/testeval/ntd). This page will be periodically updated to account for future advances in methodological approaches, as well as with expanded worked examples and links to relevant case studies as these become available. It is possible to implement the methods discussed in a variety of different software packages, but we recommend the use of the statistical programming language R (R Core Team, 2023), and references to software packages below assume that R is being used for the analysis.

5.1 Method A: comparison to a perfect reference test

Method

The traditional method of calculating sensitivity and specificity of a comparator test is by simple comparison to a reference test. As it is implicitly assumed to be perfect, only a single reference test is required (Scenarios 1, 2, 4, & 5), although a composite reference test could be used (for Scenarios 3 & 6 we assume parallel interpretation, i.e. that a positive result from either Test A or Test B infers a positive result overall). This method is theoretically applicable to Objective 2 and multiple comparator tests can be evaluated using independent analyses. Any number of populations can be used with this method.

Implementation

Calculation uses the following well-known formulae:

$$Se = \frac{TP}{TP + FN} \qquad Sp = \frac{TN}{TN + FP}$$

Where TP, FP, FN, and TN are calculated from a 2x2 table generated from the observed combinations of test results. Estimation of 95% CI is important to account for the difference in strength of evidence between small and large studies; we recommend a Bayesian approach using a Beta distribution as a conjugate prior to define the posterior:

$$Posterior_{Se} \sim Beta(TP + 1, FN + 1)$$
$$Posterior_{Sp} \sim Beta(TN + 1, FP + 1)$$

Highest posterior density (HPD) intervals can then be calculated from these posterior distributions using standard methods (e.g. the hpd function of the TeachingDemos package (Snow, 2020)). Alternatively, where a minimum of five positive and five negative samples are observed then frequentist methods can be used (e.g. the prop.test function).

Recommendation

We note that the overall performance of the reference test implicitly includes all steps in the diagnostic process, including:

i. Establishing an objective, target condition, and diagnostic test to be used as a reference.
ii. Identifying a relevant population (e.g. random sample vs. clinical suspects).
iii. Obtaining appropriate samples.
iv. Potential labelling issues and degradation of the samples during transport.
v. Technical/laboratory sensitivity/specificity of the test.

Therefore, while it may be feasible to assume that a reference test is perfect for case definitions based on artificial samples (where (i) is straightforward and (ii), (iii) & (iv) are not relevant), this will almost never be true for case definitions based on performance of the tests in the field. It is also not sufficient for such a test to have perfect specificity alone: both sensitivity and specificity must be perfect to qualify. In practice, we suggest that this situation is likely to be extremely rare in the context of NTDs, so do not recommend the use of Method A in this context.

5.2 Method B: comparison to well characterised reference tests

Method

Comparison of a new test with unknown sensitivity and specificity to one or more reference tests with known (but imperfect) performance (Scenarios 2-6) is likely to be a common application, where it is desirable to use the simplest possible statistical method that can be expected to reliably produce unbiased estimates of sensitivity and specificity of the comparator test. We propose using a method that combines a simplified latent class model with the use of posterior positive probability (PPP) values as proposed by Olsen et al (2022). These PPP values represent the posterior

probability of the latent class conditional on the combination of observed reference test results and true prevalence estimates in the relevant populations, which is an extension of the usual concepts of positive predictive value (PPV) and negative predictive value (NPV). Method B allows for multiple reference tests to be included, and imperfect test performance is accounted for, although conditional dependences terms between the comparator test and any reference test are not accounted for. This method is theoretically applicable to all six scenarios (with any number of populations) for Objective 2, and multiple comparator tests can be evaluated using the same PPP value estimates. A minimum of two populations is required.

The conceptual steps to Method B are as follows:

1. The performance of the reference tests is taken from standard recommended values. The case definition used to generate these values must match that to be used for the comparator test, otherwise this method is not applicable. The estimates should include 95% CI in order to propagate statistical uncertainty.

2. A relatively simple statistical model is fit to the observed reference test results in order to estimate the (unknown) true prevalence in the populations, accounting for the known performance of the reference test(s). This model is based on the following equations:

$$\mathrm{P}(+_t)_i = prev_i \,.\, se_t \, + (1 - prev_i) \,.\, (1 - sp_t)$$

   Where $\mathrm{P}(+_t)_i$ refers to the probability of a positive result for reference test *t* (i.e. either test A or test B), se_t and sp_t refer to the (known) performance of reference test *t*, and $prev_i$ refers to the true prevalence in population *i* (which is the parameter vector to be estimated). In cases where conditional dependence terms between reference tests needs to be accounted for, the equations can be modified using a similar approach as described in Section 5.3.

3. PPP value estimates are obtained from the known reference test performance and true prevalence estimates using extensions of formulae for PPV and NPV. These are shown below assuming two conditionally independent reference Tests A and B (Scenarios 3 & 6):

$$PPP_{R,i} = \begin{cases} \dfrac{(1-se_A).(1-se_B).prev_i}{(1-se_A).(1-se_B).prev_i + sp_A.sp_B.(1-prev_i)}, & A^-, B^- \\ \dfrac{se_A.(1-se_B).prev_i}{se_A.(1-se_B).prev_i + (1-sp_A).sp_B.(1-prev_i)}, & A^+, B^- \\ \dfrac{(1-se_A).se_B.prev_i}{(1-se_A).se_B.prev_i + sp_A.(1-sp_B).(1-prev_i)}, & A^-, B^+ \\ \dfrac{se_A.se_B.prev_i}{se_A.se_B.prev_i + (1-sp_A).(1-sp_B).(1-prev_i)}, & A^+, B^+ \end{cases}$$

Where $PPP_{R,i}$ refers to the PPP for reference test result combination R in population i, $prev_i$ refers to the true prevalence in population i, se_A and sp_A refer to sensitivity and specificity of test A, respectively, and A^- and A^+ refer to negative and positive results from test A, respectively (and equivalent subscripts are used for test B). These equations can be extended to any number of reference tests (including a single test), and can be modified to account for conditional dependence using a similar approach as described in Section 5.3.

4. The performance of the comparator test is then estimated relative to these PPP values using a second model based on the following equation:

$$\mathrm{P}(+ \mid R, i) = PPP_{R,i}.\ (1-se)\ + \left(1 - PPP_{R,i}\right).\ sp$$

Where *se* and *sp* refer to the performance of the test to be estimated. This model can be run for each comparator test independently, based on a single set of $PPP_{R,i}$ estimates.

5. Statistical uncertainty in the estimate (95% CI) can be obtained by numerical integration, using a combination of the Beta conjugate prior method shown for Method A with bootstrap sampling of $PPP_{R,i}$ over a distribution of reference test performance values. Alternatively, the more complex statistical inference method described in section 5.3 can be used.

<u>Recommendation</u>

Method B is relatively simple, works with two or more populations, and preserves the information contained within the precise combination of observed test results from multiple reference tests, which is beneficial compared to collapsing the results into a single composite reference test result. Interpretation of the underlying latent class is fixed to that used to define the performance of the reference test, which simplifies comparison of estimates produced from different studies where the

same reference test was used provided that the case definition is consistent between studies. This is a desirable property for a standardised method but does increase the susceptibility of this method to bias arising from mis-matched case definitions between the study used to evaluate the new tests and that reported for the reference tests. Method B also assumes that the comparator test is not strongly correlated with the overall combination of reference tests used; this must either be biologically justified or reduced in impact by including multiple reference tests. Method C should be used as a fall-back option in cases where conditional dependence on reference tests is suspected and/or in situations where there is doubt regarding the transferability of reference test performance estimates between studies.

5.3 Method C: concurrent estimation of performance for all tests

Method

Relaxing the assumptions made by Method B requires the use of more complex latent class models. There are different approaches to fitting models, although each works on the basis of simultaneously estimating the prevalence of the "latent state" along with the sensitivity and specificity of the two tests relative to this latent state. We begin by describing the original two-sample, two-population model of Hui and Walter (1980), which is the simplest to implement and understand.

Paired test data are obtained from two diagnostic tests (i.e. Scenarios 1, 2, 4 & 5) applied to individuals in two populations, and the number of paired test result combinations (-/-, +/-, -/+, +/+) in each population is tabulated. As the total number of individuals in each population is fixed, there are 6 degrees of freedom in the data (three each for the 2x2 tabulations in the two populations). A model is then defined using the following equations:

$$\mathrm{P}(-/-)_i = prev_i \,.\, (1 - se_1) \,.(1 - se_2) + (1 - prev_i) \,.\, sp_1 \,.\, sp_2$$
$$\mathrm{P}(+/-)_i = prev_i \,.\, se_1 \,.(1 - se_2) + (1 - prev_i) \,.\, (1 - sp_1) \,.\, sp_2$$
$$\mathrm{P}(-/+)_i = prev_i \,.\, (1 - se_1) \,.\, se_2 + (1 - prev_i) \,.\, sp_1 \,.(1 - sp_2)$$
$$\mathrm{P}(+/+)_i = prev_i \,.\, se_1 \,.\, se_2 + (1 - prev_i) \,.(1 - sp_1) \,.(1 - sp_2)$$

Where $prev_i$ is the (unknown) prevalence in population i, and *se* & *sp* denote the (unknown) sensitivity and specificity of the two tests. There are six parameters to be estimated (se_1, sp_1, se_2,

sp_2, $prev_1$, $prev_2$), however an additional constraint is imposed to ensure that each $se+sp \geq 1$ (i.e. Youden's J is ≥ 0). These equations also apply to the use of three or more populations, where each additional population increases the number of degrees of freedom by 3 and number of parameters by 1, so that more precise estimates for diagnostic test performance are obtained. For use with Objective 2, the availability of existing estimates for the performance of the reference test can be accounted for by specifying appropriate prior distributions for the relevant parameters. Accordingly, we recommend fitting these models within a Bayesian framework using Markov chain Monte Carlo – this can be achieved from within R using several different software packages, including Just Another Gibbs Sampler (JAGS; Plummer, 2003) or Stan (Stan Development Team, 2023).

Where sufficiently strong prior information is used, the model can be modified to estimate the conditional dependence between the reference test and comparator test using the following equations:

$$\mathrm{P}(-/-)_i = prev_i \,.\, ((1-se_1)\,.\,(1-se_2)+cov_{se}) + (1-prev_i)\,.\,(sp_1\,.\,sp_2+cov_{sp})$$
$$\mathrm{P}(+/-)_i = prev_i \,.\, (se_1\,.\,(1-se_2)-cov_{se}) + (1-prev_i)\,.\,((1-sp_1)\,.\,sp_2-cov_{sp})$$
$$\mathrm{P}(-/+)_i = prev_i \,.\, \left((1-se_1)\,.\,se_2-cov_{se}\right) + (1-prev_i)\,.\,(sp_1\,.\,(1-sp_2)-cov_{sp})$$
$$\mathrm{P}(+/+)_i = prev_i \,.\, (se_1\,.\,se_2+cov_{se}) + (1-prev_i)\,.\,((1-sp_1)\,.\,(1-sp_2)\,+cov_{sp})$$

Where the parameters se_1 and sp_1 (corresponding to the reference test) are fixed, and cov_{se} and cov_{sp} refer to the conditional dependence terms to be estimated. Care must be taken to ensure that these new terms are constrained so that the resulting probability terms remain valid. The use of prior information is particularly relevant when either sensitivity or specificity is strongly believed to be high based on biological reasoning, as might be the case for specificity of tests based on direct observation of pathogens (provided that the selected case definition involves presence of the pathogen). However, incorrect inference for the performance of all tests included in the model can be expected if inappropriate priors are used for any of the tests. A review of methods for elicitation of priors based on expert knowledge is outside the scope of this report, but several different software solutions are available to assist with the process including the PriorGen package (Pateras and Kostoulas, 2023).

The equations can also be extended to the use of three conditionally independent tests for 1 or more population, or to the use of three tests with pairwise conditional dependencies for 2 or more

populations (Scenarios 3 and 6). However, care must be taken to ensure that the number of parameters required to be estimated is feasible given the data and prior distributions available. With 3 tests and 2 populations, it is theoretically possible to estimate all sensitivity, specificity and pairwise conditional dependence terms (the model has 14 parameters and 14 degrees of freedom), but in practice it is better to justify inclusion of at most two (and preferably one) of these pairwise conditional dependencies on biological grounds. More tests can be added but this further complicates a model with all possible pairwise conditional dependence terms (which for four tests requires a minimum of four populations to estimate 24 parameters on 28 degrees of freedom), and the requirement for increasing numbers of populations increases the chances of violating the crucial assumption regarding constant test performance across populations. In practice, we strongly recommend pruning the relevant conditional dependence terms based on biological reasoning. Furthermore, the benefit in terms of information introduced by adding new tests will be partially (and sometimes even entirely) offset by the addition of conditional dependence terms. The simplest approach is therefore to purposefully select two or three tests that can be considered to reflect different aspects of the desired case definition. In practice, where the number of different tests theoretically available is limited, the case definition specified will be a compromise between the theoretically most desirable case definition and the assumed intersection of the most independent of the available tests. This interpretation of the implicit latent class in relation to the desired case definition is extremely challenging and requires familiarity with the technical limitations of the modelling approach combined with domain-specific experience of the disease system and mechanisms of action of the relevant diagnostic tests in relation to the available data. Where only two tests are chosen (or available), the case definition will be the intersection between these two tests, and regardless of the number of populations we do not recommend attempting to use conditional dependence terms because they are effectively indistinguishable from changes to the case definition. Other potential modifications include the use of different test profiles in different populations and allowing test sensitivity and/or specificity to vary for some tests in some populations. Further extensions of the same principle are also possible to multiple latent states, representing (for example) negative status, active infection status, and post-infection (antibody positive) status. These models may also be useful for applications involving potential co-infections that are only partially distinguishable using available tests, for example *S. mansoni* and *S. haematobium.* However, this is an area of ongoing methodological research, and should not be attempted unless collaborating with an experienced statistical practitioner.

Although the challenges involved in determining the appropriate model structure are substantial, obtaining a fit of the chosen model to data is relatively straightforward. The easiest method is to use the `template_huiwalter` function of the `runjags` package (Denwood, 2016) to generate a model for an arbitrary number of tests to be run using JAGS from within R. The conditional dependence terms to include are specified as an input to the function (for `runjags` version ≥2.2.3), the resulting model also includes the necessary parameter value checks, and the model can be run to obtain parameter estimates directly. Alternatively, the same model can be specified manually using a more modern estimation framework such as Stan or Template Model Builder, which can also be used from within R (Kristensen et al., 2016; Stan Development Team, 2023) and should give qualitatively identical results for these relatively simple models, provided that appropriate procedures are undertaken to assess convergence and validity of the Monte Carlo approximation. Regardless of the estimation framework, we recommend this model formulation for a small number of conditional dependence terms (a maximum of 1-3, depending on the number of tests), but if additional conditional dependence terms are required then alternative formulations of conditional dependence models should be explored to ensure that overall results are robust to the model specification. A valid alternative to these more complex models is to simplify the problem into two stages: (1) estimate the performance of two or three well-performing tests representing different aspects of the disease system (and where the latent class is therefore easiest to interpret); (2) run separate models (or use Method B) to estimate the performance of the remaining tests. This reduces the complexity of the problem, including interpretation of the results, at the potential cost of wider 95% CI than might be obtained from a combined model where the additional assumptions can be assumed to be valid.

Regardless of the method used to implement the model, the empirical fit to data should be assessed (and reported) by comparing the predicted probabilities to observed tallies in order to ensure that the observed data is plausible under the model assumptions. We also recommend calculating PPP values within the model (as described for Method B, but including all test results) so that the performance of the tests can be re-estimated for each individual population in order to ensure that the key assumption of constant test performance is met. Further details on implementing these model fit assessment methods are given on the accompanying website. A complete sensitivity analysis should also be undertaken, including re-running the model with different subsets of the

data (e.g. dropping one population at a time) to ensure that inference is robust. If moderate or strong priors have been used for one or more parameters, then a comparison must be made of these prior distributions to the corresponding posterior distributions to ensure that there is no prior-data conflict. It may also be advisable to re-run the model with weaker priors where relevant and practical to do so.

Recommendations

There are two major challenges with Method C. The first challenge is the difficulty in providing universal sample size requirements: these are highly dependent on the number of populations, variation in true prevalence, number of tests, degree of conditional dependence between tests, and the performance of the tests. We therefore strongly recommend undertaking a bespoke sample size evaluation procedure as part of the process of planning prospective studies. The second and arguably more important challenge lies in the complexity of understanding the relationship between the model assumptions and underlying biological processes. This includes selection of appropriate tests to match the desired case definition, specification of an appropriate model to match the assumed conditional dependences between tests, and interpretation of the results. We note that the most appropriate approach will always depend on the precise application, number and type of tests available, and number of populations in the data.

The most common use case for this approach is likely to be the initial evaluation of reference tests to generate unbiased estimates of sensitivity and specificity that can then be used in future analyses leveraging simpler methods. Along with estimates (and 95% CIs) for the performance of the tests and the strength of any conditional dependence terms, we also strongly recommend that the end result of this process should include a clear and unambiguous description of the case definition, populations and applications to which this case definition is relevant, and recommendations for sample sizes required for future applications of Method B (Objective 2). Method C will also be useful in situations where Method B is used as the primary analysis, but the assumption regarding consistent performance of the reference test(s) between studies should be verified. However, this requires that a sufficient number of populations is available to allow the LCM to be run on different subsets of the data.

Overall, although these models are not technically difficult to implement, this complexity in interpretation represents a potential danger when used by practitioners without experience in implementing and interpreting their output. However, their use is necessary in situations where robust estimates for the performance of reference tests is not available (objective 1). In these cases, the STARD-BLCM guidelines should be followed carefully (Kostoulas et al., 2017), including giving careful consideration to the implicit latent class, and when in doubt we strongly encourage collaboration with statistical practitioners that are experienced in the use of these methods. Examples of how to perform some of these model fit evaluation methods are given on the accompanying website, and the HARMONY COST action "Novel tools for test evaluation and disease prevalence estimation" and associated Society for Advanced Methods in Epidemiology and Diagnostics can be used both a source of additional resources and a network of experts who may be able to assist with analysis of more complex datasets.

5.4 Alternative objectives

Here we include a brief overview of methods that may be used to address additional objectives relating to estimation of diagnostic test performance. Worked examples of the methods outlined are given on the accompanying website.

Determining non-inferiority

Rather than simply estimating the performance of the new test, the primary objective of some studies may be to establish that a new diagnostic test performs equivalently to (or better than) an existing diagnostic test. This is analogous to a non-inferiority trial for demonstrating that the efficacy of a new clinical intervention is no worse than the efficacy of an existing intervention minus a pre-specified non-inferiority margin (Mulla et al., 2012; Walker and Nowacki, 2011). Pre-requisites of this objective are (1) that a single objective be defined in terms of sensitivity and/or specificity (for example should both sensitivity and specificity be non-inferior or is it sufficient for the new test to more simply demonstrate a higher Youden's J), and (2) that reference tests with known performance already exist (although estimation of the performance of these reference tests may also be undertaken as part of the same field study, we recommend treating the non-inferiority test as a subsequent analysis). This implies a variation of objective 2, and the performance of the new test can therefore be estimated using either Method B or Method C depending on the combined diagnostic performance of the available reference test(s). An estimate and 95% CI for the difference

in performance (for example Youden's J) between the two tests can then be calculated by subtracting the Monte Carlo samples for the new test from Monte Carlo samples taken from the distribution of uncertainty for the existing test, to arrive at a distribution representing uncertainty in the difference in performance. The probability that the new test is non-inferior to the existing test is then calculated by examining the empirical cumulative distribution function (CDF) of this distribution to the critical value of zero minus the pre-specified non-inferiority margin. When conducting a non-inferiority trial for a new test B against a test A that was itself validated using a non-inferiority trial against an original test, it is important that the thresholds used to evaluate test A are held constant when evaluating test B (and any future tests) in order to avoid 'bio-creep' (Everson-Stewart and Emerson, 2010).

Optimisation of test cut-off values

Although outside the scope of this project, we note that the statistical methods presented can also be used to optimise cut-off values for tests that produce a (semi-) quantitative output such as (semi-) continuous values (optical density or cycle threshold), count data (faecal egg counts or parasite counts), or gene copies (qPCR). Method B can be used to obtain PPP values based on the reference test(s) and then estimate the sensitivity and specificity that would be expected when dichotomising the comparator test results using a range of different cut-off values. All possible valid cut-off values can be obtained by extracting the set of unique observations in the data, sorting these into ascending order, and then defining the cut-off values as the mid-point between the 1st and 2nd value, 2nd and 3rd value and so on. Results can be presented either as a receiver operating characteristic (ROC) curve, and/or as simpler plots of the estimated sensitivity, specificity and Youden's J depending on the cut-off value chosen. One caveat of this procedure is that the resulting estimates of sensitivity and specificity are biased by the re-use of the same data to estimate the optimal cut-off and the performance conditional on this optimised cut-off, although this limitation can be avoided by splitting the dataset into calibration and evaluation subsets.

6 NTD-Specific Considerations

This section provides a basic summary of the epidemiology and diagnostic test situation for all listed NTDs[2] along with their status under the global targets for 2030[3], and is intended to serve as a reference point for application of the methods described in the main body of the report. The information provided in the summaries was obtained from a combination of the WHO target product profiles (TPP[4]), the WHO road map for NTDs 2021-2030, as well as several interview/discussions with representatives of the DTAG subgroups and other relevant experts. We are extremely grateful to these individuals for their insights.

Buruli ulcer

Buruli ulcer is a skin disease caused by an environmental bacterium (*Mycobacterium ulcerans*), which produces a toxin that causes (initially painless) localised or diffuse swellings, plaques or oedema. Without treatment, the lesion progresses to ulceration within four weeks, resulting in long-term disability in approximately 25% of affected patients. The mode of transmission to humans is unknown but is likely possible from a wide range of sources, so the main relevance of diagnostic testing is in early identification of the disease to facilitate rapid treatment, consisting of antimicrobial treatment combined with a prolonged course of wound care. The WHO road map for NTDs 2021-2030 specifies a target of *control.*

Diagnosis in endemic settings is typically made on clinical grounds, including the use of the WHO Buruli ulcer scoring scheme, which is based on clinical signs and demographic characteristics of the patient. Diagnostic tests are available, including culture, direct smear examination, histopathology, and PCR, where PCR is considered the reference test and can be expected to have high sensitivity and close to 100% specificity, provided that the entire lesion was swabbed. Challenges include the lack of availability of PCR in endemic areas, where lower-sensitivity microscopy is more commonly available. The DTAG target for Buruli ulcer is therefore focussed on development of

---

[2] https://www.who.int/health-topics/neglected-tropical-diseases; accessed February 2024

[3] https://www.who.int/teams/control-of-neglected-tropical-diseases/ending-ntds-together-towards-2030/targets

[4] https://www.who.int/observatories/global-observatory-on-health-research-and-development/analyses-and-syntheses/target-product-profile/links-to-who-tpps-and-ppcs

rapid point-of-care tests and standardisation of PCR testing. Particular challenges for defining reference standards for Buruli ulcer include the potentially variable sensitivity and specificity of diagnostic tests depending on the experience of the operator. This applies in particular to clinical diagnosis, but also to PCR and microscopy where improper swabbing of a lesion may reduce sensitivity.

Chagas Disease

Chagas disease is caused by infection with a protozoan parasite (*Trypanosoma cruzi*) via a wide range of transmission routes including vector-borne, food-borne, congenital and via contaminated organs and blood products. The disease is typically chronic and asymptomatic, although muscle and nerve damage can lead to heart failure and sudden death. The WHO road map for NTDs 2021-2030 specifies a target of *elimination as a public health problem.*

Diagnostic tests for Chagas include antibody tests, detection of *Trypanosoma* spp. by PCR, and serological rapid diagnostic tests. Challenges include lack of field validation of rapid diagnostic tests and the lack of direct parasitological diagnostics. Co-infections and co-morbidities are common. Particular challenges for defining reference standards for Chagas include the chronic nature of the disease, which may result in e.g. variable sensitivity antibody testing depending on the stage of infection. The presence of co-infections could potentially also impact specificity.

Dengue and Chikungunya

Dengue and chikungunya are vector-borne viral diseases that both cause fever, joint and muscle pain, headache, rash, and lymphopenia. Differentiation of the two is extremely difficult clinically, although infection with Dengue virus is frequently asymptomatic, whereas chikungunya infections are typically symptomatic. Both can cause large outbreaks in situations favouring the mosquito vector. The WHO road map for NTDs 2021-2030 specifies a target of *control.*

Available diagnostic tests include antibody ELISA tests and RT-PCR methods to detect the viral agents, although these are of varying sensitivity. Development of new diagnostic tests with improved sensitivity and specificity has been identified as a key objective. Challenges for defining reference standards include differentiation of dengue *vs* chikungunya *vs* other causes of fever etc, which will impact specificity of diagnosis based on clinical signs.

Dracunculiasis

Dracunculiasis is caused by infection with the guinea worm parasite (*Dracunculus medinensis*), which causes painful blisters in the feet of infected individuals. Transmission occurs via ingestion of parasitised water fleas in stagnant drinking water, and contamination of water occurs as a result of affected patients immersing affected limbs in water to soothe the painful lesions, into which larvae are released by the mature worm within the lesion. The parasite also infects animals, and separate TPP are under development for detection of pre-patent infection in animals and detection of the parasite in the environment. Dracunculiasis is one of the two diseases identified by the WHO road map for NTDs 2021-2030 for *eradication.*

Diagnostics for dracunculiasis are currently based on clinical macroscopic findings and PCR testing. Identified aspects for development include serological tests and pond-side tests. The major challenge for defining reference standards relates to varying sensitivity of clinical findings based on stage of development of the parasite/lesion.

Echinococcosis

Cystic and alveolar echinococcosis are tapeworm infections characterised by development of cysts/vesicles in various areas of the body. The disease may be asymptomatic for several years, before causing a variety of symptoms depending on the affected organ system. Humans are accidental intermediate hosts of the usual canid-livestock/rodent life cycle. The WHO road map for NTDs 2021-2030 specifies a target of *control.*

There is currently a lack of standardised diagnostic tests for echinococcosis in humans, dogs and livestock. Diagnostic imaging is the most commonly used method in humans, although serological tests are available. Coproantigen tests are available for use in dogs, but these lack validation. A particular challenge for defining reference standards is the lack of standardisation of the available tests.

Foodborne Trematodiases

Trematode (fluke) infection is caused by a number of parasite species, causing clinical symptoms resulting from pathology in the liver and bile ducts. Infection is via raw or undercooked food contaminated with parasite larvae, and the parasites have a broad host range including wildlife, livestock and domestic animals. The WHO road map for NTDs 2021-2030 specifies a target of *control.*

The available diagnostic methods include parasitological methods (e.g. detection of parasite eggs in stool) and diagnostic imaging. Serological methods and PCR tests are also under development. The relatively long development period of the parasite combined with varying sensitivity of tests depending on the stage of development are the major challenges to defining reference standards.

Human African Trypanosomiasis

Human African Trypanosomiasis (HAT) is caused by infection with the parasites *Trypanosoma brucei* subspecies *gambiense* (West African sleeping sickness; ~95% of cases) and *rhodesiense* (East African sleeping sickness). Their respective geographic ranges are mostly separate, and both are zoonotic, although *gambiense* primarily affects humans whereas *rhodesiense* primarily affects animals (wild and domestic, including cattle). Transmission is primarily vector-borne (by tsetse fly bite) although vertical transmission is also possible. Clinical symptoms include fever, headache, joint pain, and behavioural/motor disturbances once the parasite crosses the blood-brain barrier. The WHO road map for NTDs 2021-2030 specifies a target of *elimination of transmission* for Gambiense HAT and *elimination as a public health problem* for Rhodesiense HAT.

Clinical signs of HAT are non-specific, so diagnosis relies on laboratory methods. Serological tests for antibodies exist for Gambiense HAT, but do not necessarily reflect infection status, and are not available for Rhodesiense HAT. Detection of active infection can be done with point-of-care rapid diagnostic tests, and the card agglutination test for trypanosomiasis (CATT), which is mostly used by specialised teams undertaking active screening. Detection of the parasite by microscopic examination of body fluids is also possible but requires specialist training. Identified future needs include parallel screening methods that can be carried out by non-specialised personnel, which would ideally be applicable to both humans and animals. Four separate TPP have been written for use in different circumstances, reflecting the different challenges involved with individual-level diagnosis of Gambiense *vs* Rhodesiense HAT for the purposes of treatment *vs* use as part of community-level

control programmes. A particular challenge for defining reference standards is distinguishing between active infection (presence of pathogen) and serological positivity (presence of antibodies).

Leishmaniasis

Cutaneous and visceral leishmaniasis is caused by vector-borne (sandfly) protozoan parasites of the genus *Leishmania*. Infections are often asymptomatic but can cause ulcerative skin lesions (cutaneous form) and fever/anaemia (visceral). Both forms are associated with immunosuppression, and coinfection with HIV is a particular concern. The WHO road map for NTDs 2021-2030 specifies a target of *control* for cutaneous leishmaniasis and *elimination as a public health problem* for visceral leishmaniasis.

Diagnosis of cutaneous leishmaniasis is currently based on parasitological tests and/or clinical features, but these lack sensitivity. Rapid tests are available for visceral leishmaniasis but similarly lack sensitivity. PCR tests are available in reference laboratories, and antibody tests are potentially useful for screening of dermal leishmaniasis and mucosal leishmaniasis, but the long-term persistence of antibodies affects specificity relative to active infection. The development of more sensitive and affordable rapid diagnostic tests has been identified as a priority. Issues of particular relevance to defining reference standards include the potentially varying sensitivity of the available tests for cutaneous leishmaniasis and increased possibility of co-infections (and therefore potentially reduced specificity of tests) due to immunosuppressive effects of the parasite.

Leprosy

Leprosy is a bacterial disease (caused by *Mycobacterium leprae and Mycobacterium lepromatosis*) characterised by an extremely long incubation period averaging 5+ years but sometimes as long as 20 years. Direct transmission occurs via contact with respiratory droplet nuclei from infected (and possibly asymptomatic) individuals, and lesions typically develop on cooler areas of the body, in particular the elbow and ear lobes (although this is strain dependent). Clinical symptoms are classified into a number of different forms, but all are a direct result of damage to skin and peripheral nerves. Asymptomatic carriage of up to 20% is common in endemic areas. Leprosy is identified by the WHO road map for NTDs 2021-2030 for *elimination of transmission.*

Historically, diagnoses of leprosy have been made entirely on clinical grounds, based on the presence of one or more of the three cardinal symptoms (localised loss of skin sensation or muscle weakness, and presence of acid-fast bacilli). A number of studies have examined the accuracy of these historical clinical diagnoses, but a major concern is that performance of these clinical tests may be declining due to dwindling clinical expertise. Histological methods are also available in some settings but have low sensitivity. Effective control of leprosy relies on post-exposure prophylaxis, which requires identification of asymptomatic carriers of the infection - this necessitates diagnostic tests that can detect carriage of the pathogen rather than clinical disease. Serological methods allow detection of carriage but are non-specific both due to cross reactions with antibodies to related pathogens and in terms of predicting progression to disease. Although some of the immunodiagnostics are highly specific, their sensitivity is very low for the paucibacillary form the disease. Additional tests include slit-skin smears and PCR tests. New tests (including molecular tests and qPCR tests) are under active development and can be broadly divided into (1) point-of-care tests to detect pathogen-specific analytes, (2) laboratory and point-of-care tests to confirm elimination of the pathogen. TPP for leprosy are similarly divided into tests to be used for clinical disease *vs* asymptomatic individuals. Leprosy presents several challenges to defining reference standards, including the distinction between detection of asymptomatic carriage *vs* clinical disease, and the potentially variable sensitivity and specificity of the available tests. Performance of current tests has typically been measured against clinical diagnosis with varying criteria, so the existing estimates for sensitivity and specificity of these tests may be unreliable.

Lymphatic Filariasis

Lymphatic filariasis is a vector-borne (mosquito) disease caused by infection with three filarial parasites (*Wuchereria bancrofti*, *Brugia malayi*, *B. timori*). A variety of different mosquito species are competent, but human infection is required to complete the life cycle. Generation of microfilariae in a host requires a mating pair of parasites and is a slow process taking up to a year, so many infections are asymptomatic. Clinical symptoms of chronic lymphoedema and hydrocele arise as a result of impaired lymphatic function following damage to lymphatic vessels caused by the pathogen, but often develop some time after the pathogen itself has been eliminated. Lymphatic filariasis is identified by the WHO road map for NTDs 2021-2030 for *elimination as a public health problem.*

Testing for lymphatic filariasis is currently focussed mostly on population-level testing in conjunction with preventive chemotherapy, rather than individual-level tests. The currently most used test is a lateral flow assay to detect adult worm antigen in *W. bancrofti* endemic areas; this has largely replaced the older method of direct detection of microfilariae in blood samples, which is less sensitive and has practical challenges due to the necessity to sample during the night. There are no *Brugia* spp. antigen detection tests, so antibody tests are used in these settings to make program decisions. In addition, antibody tests have been used in research settings and are anticipated to be used for serological surveillance in the future in both *W. bancrofti* and *Brugia* spp. endemic areas. Currently available tests have not been recommended by the WHO for use in diagnostic settings, and typically either detect long-lived antibodies (perhaps years after infection) or are poorly specific due to cross-reactions. Surveillance via testing of the mosquito vector is also possible, and development of new tests is ongoing, including refinement of the antigens used for antibody detection. Major challenges include feasibility of testing in the field using available tests, and the challenges of accurately measuring low prevalence given the imperfect specificity of the antigen test (cross-reactions with *Loa loa* infection are frequently observed) and long-lived nature of antibodies. For this reason, children are often used as sentinel populations in treatment areas.

There are a number of aspects to lymphatic filariasis that present challenges to defining a reference standard. Performance of the diagnostic tests must be considered separately for each stage of the disease process, as presence of the parasite, presence of antibodies and presence of clinical disease cannot be expected to occur contemporaneously. A distinction must also be made between individual-level testing and testing as part of surveillance and control programmes with unknown background prevalence, although the latter is of most practical relevance. A further challenge is the long-lived nature of antibodies, which would be expected to cause a temporal lag between decreasing prevalence of the parasite and decreasing seroprevalence in the population.

Mycetoma, Chromoblastomycosis and Other Deep Mycoses

This group of diseases are caused by chronic infection of the skin and subcutaneous tissue following inoculation of one of a group of fungal pathogens through areas of broken skin. Clinical symptoms vary widely depending on the pathogen involved, although skin lesions are the most common presentation. The WHO road map for NTDs 2021-2030 specifies a target of *control.*

Diagnostics are based on clinical presentation and detection of relevant fungal pathogens from skin scrapings and biopsies. There are currently no rapid diagnostic tests or serological tests available. Due to the importance of early treatment, development of tests for early detection at primary health care level has been identified as a priority. The principle challenge for defining a reference standard is the potentially variable sensitivity of diagnostic tests based on skin scrapings/biopsies depending on the clinical experience of the operator.

Noma

Noma is a gangrenous disease of the mouth that is mostly found in children aged 2-6 in poorer communities. It is caused by a range of different opportunistic pathogens and is not considered to be contagious. Treatment is typically with antimicrobials. Noma was added to the list of NTDs in December 2023, so it not included in the WHO road map for NTDs 2021-2030.

Diagnosis of noma is based on clinical criteria scoring infection on a 0-5 point scale. Other tests are not currently available. There are no particular challenges in defining reference standards for noma as a clinical presentation, although the diverse range of causative pathogens does present a challenge for assessment of new diagnostic tests that may only detect a subset of these.

Onchocerciasis

Onchocerciasis (aka river blindness) is a vector-borne (blackfly) disease caused by filarial parasites *(Onchocerca volvulus*), which is associated with skin irritation, lesions, blindness, and possibly neurological disorders. Mass drug administration with ivermectin kills microfilariae, preventing transmission of infection, but does not kill the long-lived adult worms. Repeated treatment of the entire community once or twice per year over a period of 15-20 years is therefore required to control the parasite. The WHO road map for NTDs 2021-2030 specifies a target of *elimination of transmission.*

There are a number of diagnostic tests available for onchocerciasis. Skin biopsies have high specificity but low sensitivity and can be painful for the patient. Clinical palpation of nodules is a simple and non-invasive test that suffers from low specificity due to the presence of unrelated nodules. The transdermal “DEC patch” works by killing microfilariae in the skin, inducing a reaction that can be visualised two days later. This test has practical limitations and concerns with

specificity due to cross reactions with *Loa loa*, and has not yet been fully evaluated. Serological tests are used as the basis for stopping mass drug administration, but lack of standardisation is an issue. Detection of infected blackflies is also possible. The main issue of concern for defining reference standards is the potential for varying sensitivity and specificity of the available tests, due to inter-operator variation, lack of standardisation, and the possibility of cross-reactions with other endemic diseases.

Rabies

Rabies is a viral infection (caused by lyssaviruses including the rabies virus) with a unique epidemiology characterised by a long incubation period and extremely wide host range. The vast majority of cases in humans are due to bites from infected domestic dogs, and vaccination of domestic dog populations combined with post-exposure prophylaxis is effective at preventing cases in humans. Rabies is targeted by the WHO road map for NTDs 2021-2030 for *elimination as a public health problem.*

Well-established diagnostic protocols and vaccines are available for rabies, although a field-deployable ante-mortem test has been identified as a priority for development. Other than the long incubation period before development of clinical symptoms, there are no particular concerns for rabies in terms of defining reference standards.

Scabies and Other Ectoparasitoses

Mite infections are a common problem (particularly in children) characterised by skin rashes and itching caused by immune responses to parasite eggs that are laid in the skin following close contact with an infected individual. Secondary bacterial infections can occur, which complicates disease progression. Treatment options include topical scabicides and oval ivermectin, although prophylactic treatment of household contacts is also necessary to control the mite, and mass drug administration is typically recommended at a prevalence of 10% or more. Untreated infections may take months to resolve. The WHO road map for NTDs 2021-2030 specifies a target of *control.*

Diagnosis of mite infections is predominantly done on clinical grounds, which has high sensitivity but low specificity, or via direct microscopic visualisation of the mite or eggs, which has low sensitivity but high specificity. Clinical diagnostic criteria vary, although efforts have been made to

find a consensus using Delphi studies. Counts of egg/mites are sometimes reported from direct visualisation methods, although typically results are given qualitatively. Challenges for defining reference standards are that sensitivity of direct visualisation is dependent on the severity of infestation (and is lowest during the asymptomatic phase before pruritis is reported in specific regions of the skin), and is also lower in darker coloured skin compared to lighter coloured skin. Some lesions occur as an inflammatory reaction rather than due to presence of mites/eggs on that area of the skin, so sensitivity will vary depending on if it is defined at lesion level, patient level, or community level (which best matches the focus of interventions).

Schistosomiasis

Schistosomiasis is caused by trematode parasites in the genus *Schistosoma* and are transmitted to humans through contact with larvae-infested water. There are two main forms of the disease, resulting from infection with different species of parasite: urogenital disease caused by *Schistosoma haematobium* that lives in blood vessels surrounding the bladder, and intestinal disease caused mainly by *Schistosoma mansoni* (as well as other species including *S. japonicum* and *S. mekongi*) that live in mesenteric veins of the gastrointestinal tract. Pathology in urogenital schistosomiasis results from egg deposition in the bladder wall, which can cause hematuria, bladder fibrosis and kidney damage. In intestinal schistosomiasis, eggs become lodged in the liver and intestinal wall and can cause (in advanced cases) damage to the liver and spleen. Untreated infection with *S. haematobium* can also lead to female genital schistosomiasis (FGS); a gynaecological condition resulting from deposition of eggs in the female genital tract. Children aged 5-15 are at highest risk of infection, so control focusses on mass drug administration (MDA) of praziquantel to school-aged children, which is effective against adult parasites (although not juveniles). MDA alone is not sufficient to interrupt transmission of the parasite. Other measures such as improvements to sanitation, snail control, and behaviour modification are likely necessary to interrupt transmission. Schistosomiasis is targeted by the WHO road map for NTDs 2021-2030 for *elimination as a public health problem.*

The desired diagnostic target for schistosomiasis is presence of adult worms, as these are responsible for both pathology and onward transmission via the production of eggs. Urine filtration (for *S. haematobium*) and Kato-Katz faecal smear (for *S. mansoni* and others) are widely available methods of enumerating the parasite egg output of infected individuals and are frequently used to

estimate the prevalence and intensity of infection. However, although these methods are highly specific, they may lack sensitivity, especially as prevalence and intensity of infection decrease following MDA (although this limitation can be partially overcome by using appropriate statistical methods to analyse the data). Eggs are highly aggregated in faeces and their production varies on a daily basis. In addition, the consistency of stool will also affect the egg count, but this is typically not considered. Alternative diagnostics include a point-of-care test to detect circulating cathodic antigen (CCA) in urine that has higher sensitivity than microscopy (particularly for *S. mansoni*) but lower specificity, which may partly reflect detection of juvenile worms. Circulating anodic antigen (CAA) tests measure released parasite antigen and detect both species of parasite, although with potentially varying sensitivity. One of these is an Up-Converting Phosphor Lateral Flow (UCP-LF) laboratory-based test that provides a quantitative indication of the level of antigen present and works with blood or urine. A rapid diagnostic test for CAA using fingerstick blood is being developed for field use. PCR tests to detect DNA from parasite eggs in stool are expected to have higher sensitivity, although the tests currently in use are not standardised. The use of a composite test based on an amalgamation of these tests (along with microscopy) has also been proposed. Antibody tests have a relatively high sensitivity but are unable to differentiate between actively infected and cured individuals due to long-term persistence of antibodies. However, these tests can be useful for surveillance in younger age groups and to support stopping MDA. Many existing antibody tests use native antigens prepared from parasite life cycle stages, which is not sustainable. However, tests using recombinant antigens for specific antibody isotypes are under development. The presence of haematuria is a simple and relatively good proxy for *S. haematobium* infection but is not diagnostic for other species of the parasite. For the specific condition of FGS, the primary diagnostic method is visualisation of egg-induced lesions on pelvic examination, but histological biopsy and PCR testing of samples taken from the cervix or vagina have also been used.

Schistosomiasis presents a number of challenges in terms of defining a reference standard. A particular issue is the differential test sensitivity depending on parasite species: most obviously Kato Katz and urine filtration, but also potentially varying sensitivity of antigen-detection tests. Ideally, performance of the tests would be considered separately by species, although this is complicated by the partially overlapping geographical distributions: *S. haematobium* is endemic primarily in Africa, whereas *S. mansoni* is endemic in both sub-Saharan Africa and South America; *S. japonicum* occurs in China, Indonesia, and the Philippines. Community co-infections with *S.*

haematobium and *S. mansoni* are therefore possible in some areas. Schistosomiasis is also a good example where consideration must be given to the testing use case; for example, individual-level treatment decisions for FGS have different requirements compared to population-level prevalence estimation; use of antibody tests in children may be useful to determine interruption of transmission within a community. There is an urgent need for standardised guidance on planning studies to evaluate tests for schistosomiasis (including sample size recommendations) due to the number of new tests under development. Nevertheless, the wide-spread availability of reference tests with high specificity is an advantage.

Soil-Transmitted Helminthiasis

Soil-transmitted helminthiasis is caused by intestinal infection with soil-transmitted helminth (STH) parasites, including *Ascaris lumbricoides, Trichuris trichiura, Strongyloides stercoralis,* and the hookworms *Necator americanus* and *Ancylostoma duodenale*. Transmission of parasite eggs/larva occurs via the faecal-oral route (*A. lumbricoides, T. trichiura*, *A. duodenale*) or skin penetration (*S. stercoralis*, *N. americanus, A. duodenale*), and clinical symptoms include abdominal pain, diarrhoea, malnutrition and anaemia, particularly in children and reproductive-age women with moderate to heavy burdens. Treatment is with anthelmintics, although efficacy varies across compounds and parasites, and anthelmintic resistance is of increasing concern. Control measures include standard hygiene practices and mass drug administration of anthelmintics to school-aged children. STH is targeted by the WHO road map for NTDs 2021-2030 for *elimination as a public health problem.*

The standard mechanism for diagnosis of STH (other than *Strongyloides stercoralis*) is by demonstrating the presence of eggs in stool samples using Kato-Katz. Eggs of the different (major) pathogens are distinguishable, so typically an enumerated count of eggs for each species (in eggs per gram of faeces) is provided. The egg count threshold used to determine prevalence can also vary depending on the use case; mass drug administration is typically justified based on the prevalence of high-intensity infections, but so-called “cure rates” are based on a comparison of pre- *vs* post-treatment egg counts above a threshold of zero. Other egg detection methods are also available, including newer microscopy-based methods including FECPAK and FLOTAC, and qPCR methods to detect DNA of parasite eggs. These methods have both advantages and disadvantages relative to Kato-Katz in terms of logistics and performance, but for some species of parasite embryonation

impacts the relationship between egg detection methods and qPCR results depending on the time interval between sampling and laboratory analysis. Alternative tests mechanisms targeting antigens, antibodies and metabolites are under development. The challenges in terms of defining reference standards for STH share some similarities with those for schistosomiasis; count-based tests involve the use of a threshold for dichotomisation, and diagnostic test performance is complicated by the potential for co-infection with different parasite species with varying fecundity (and therefore effective sensitivity). In addition, the common target of egg counting methods should be considered, particularly when comparing to tests with different mechanisms (including DNA detection methods).

Snakebite Envenoming

Snakebite envenoming occurs as a result of exposure to snake venom, either following a bite or from venom sprayed into the eyes. Clinical symptoms include shock, paralysis, and bleeding disorders that can lead to a number of potentially series sequalae. Snakebite envenoming is targeted by the WHO road map for NTDs 2021-2030 for *control.*

Diagnosis and treatment of snakebite envenoming is mostly syndromic. Diagnostic tests can facilitate differentiation of the species of snake involved, although their use cases are mostly limited to research. There are no issues relevant to defining reference standards for testing.

Taeniasis/Cysticercosis

Taeniasis and cysticercosis are tapeworm (*Taenia solium*) infections caused by ingestion of tapeworm cysts in poorly cooked pork (taeniasis) or ingestion of food contaminated with tapeworm eggs (cysticercosis). Taeniasis is usually asymptomatic (but may cause gastrointestinal disease), whereas cysticercosis is characterised by the development of cysts in muscle tissue, eyes, skin or the nervous system (neurocysticercosis), with clinical signs developing according to the location of the cyst. Taeniasis/cysticercosis is targeted by the WHO road map for NTDs 2021-2030 for *control.*

Diagnosis of taeniasis can be done via demonstration of tapeworm eggs in stool, although this method lacks sensitivity for taeniasis and cannot be used to diagnose cysticercosis. Diagnosis of cysticercosis requires imaging and/or biopsy of the infected tissue. Antibody tests for cysticercosis are available for research use, but do not always detect intracranial cysts. Other than the lack of

availability of diagnostic tests, the only issue relevant to defining reference standards for diagnostic tests is that the two conditions (taeniasis *vs* cysticercosis) must be considered as entirely separate target conditions.

Trachoma

Trachoma is caused by infection of the eye with the bacterium *Chlamydia trachomatis*, causing irritation of the eye (trichiasis) and potentially leading to blindness. Infection can be both by mechanical vectors (files) and by direct transmission following contact with ocular or nasal discharge. Trachoma is targeted by the WHO road map for NTDs 2021-2030 for *elimination as a public health problem.*

Diagnosis of trachoma is typically based on clinical examination, although microbiological culture as well as both DNA-based and RNA-based PCR tests for *Chlamydia trachomatis* are available. Serological tests to support elimination and surveillance have also been identified as a potential area for development. The two potential issues relating to defining reference standards are the persistence of clinical signs following clearance of the causative pathogen, and the potential future need to distinguish between recent infection and longer-term serological status.

Yaws

Yaws is a chronic disease caused by spiral bacteria (*Treponema pallidum* subspecies *pertenue*), which is thought to be transmitted via skin-skin contact. Yaws has a relatively long clinical course, starting with a single infectious lesion (primary yaws) that spontaneously heals leaving the individual infected but asymptomatic (latent yaws). Following this, new infectious lesions may occur and then resolve; several episodes of alternation between this secondary yaws and latent yaws may occur over a period of several years. Finally, without intervention by antimicrobial treatment, the pathogen may cause damage to skin, bones, and cartilage, leading to physical disfigurement (tertiary yaws). Yaws is one of the two NTDs that is targeted by the WHO road map for NTDs 2021-2030 for *eradication.*

Several diagnostic tests are available for yaws, including a PCR test that can be used to demonstrate the presence of the pathogen in active lesions with high sensitivity and specificity, although sensitivity may be lower for less exudative non-symptomatic lesions. Rapid point-of-care tests are

also available, and lesion swabs can additionally be used to test for azithromycin resistance. The major challenge is in detection of individuals with latent yaws: this relies on the use of serological tests that have low specificity due to the persistence of antibodies following bacterial cure, as well cross-reactions with the closely related causative organism for syphilis. The latter is a particular challenge as syphilis antibodies persist at a high level for life following infection, and antibodies can also be transmitted from mother to child. These challenges can be partially overcome by examining paired serum titres pre- and post-treatment in the same individual, with the assumption that a falling titre represents successful treatment. A further challenge is the possibility of latent yaws concurrently with an unrelated dermatological condition such as scabies, which can be mistaken for a yaws lesion. As such, a point-of-care molecular test to differentiate yaws from unrelated dermatological lesions is an identified area for active development.

The principle challenge for defining reference standards lies in the varying test sensitivity between primary/secondary yaws and latent yaws. Assessment of test performance must therefore account for the implicit target condition; this is of particular relevance to the use of the tests for individual-level diagnosis *vs* eradication of yaws at community level. A further challenge is the low specificity of the antibody test, in particular cross-reactions with syphilis, which can be expected to occur with high frequency in some populations.

7 Overall Recommendations

This report includes a comprehensive guide to the key challenges, concepts, and methodological approaches to defining and using reference standards in the context of diagnostic test evaluation for NTDs in practice. A more concise flow chart is also provided (Appendix); this is intended to be used as a decision tree to help determine the most relevant approach to a specific problem.

In summary, we emphasise the following key points:

- It is crucial to define a clear target condition, evaluate the case definition implicit to the choice of diagnostic tests, and ensure that this case definition is consistent across studies where the performance of tests is assumed to be constant. Performance of a diagnostic test must be estimated in the same population(s) in which it is intended to be used.
- Careful planning of a prospective study should always be undertaken, including the selection of tests to be applied and the potential conditional dependence between these. We strongly recommend the use of a DAG as part of this process, in order to help visualise the relevant biological pathways.
- The selection of an appropriate statistical method should be matched to the objectives of the study and practical limitations in terms of availability of diagnostic tests and the likely performance of these tests. Where multiple different approaches to statistical analysis exist, it is good practice to analyse the data using each potentially applicable method to ensure that sensitivity and specificity estimates are consistent. We do not recommend the use of simple comparisons to reference standards (including composite reference standards).

We also note that there are continual methodological developments in the field of latent class modelling, so the optimal approach to implementing the methods described may change over time. For more complex scenarios we strongly recommend collaboration with researchers and statistical practitioners with experience in the development and application of these models, including the network of experts within the Society for Advanced Methods in Epidemiology and Diagnostics arising from the HARMONY COST action (http://harmony-net.eu).

8 Bibliography

Adel, A., Berkvens, D., 2002. Modelling Conditional Dependence Between Multiple Diagnostic Tests , Using Co-Variances Between Test Results 1, 1–6.

Albert, P.S., Dodd, L.E., 2004. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 60, 427–435. https://doi.org/10.1111/j.0006-341X.2004.00187.x

Amewu, R.K., Akolgo, G.A., Asare, M.E., Abdulai, Z., Ablordey, A.S., Asiedu, K., 2022. Evaluation of the fluorescent-thin layer chromatography (f-TLC) for the diagnosis of Buruli ulcer disease in Ghana. PLoS ONE 17, e0270235. https://doi.org/10.1371/journal.pone.0270235

Aws, R., Jb, R., Coomarasamy, A., Ks, K., Pmm, B., 2007. Evaluation of diagnostic tests when there is no gold standard. Health Technol. Assess. 11, 1–86.

Ayelign, B., Jemal, M., Negash, M., Genetu, M., Wondmagegn, T., Zeleke, A.J., Worku, L., Bayih, A.G., Shumie, G., Behaksra, S.W., Fenta, T., Damte, D., Yeshanew, A., Gadisa, E., 2020. Validation of in-house liquid direct agglutination test antigen: the potential diagnostic test in visceral Leishimaniasis endemic areas of Northwest Ethiopia. BMC Microbiol. 20, 90. https://doi.org/10.1186/s12866-020-01780-0

Bakuza, J.S., Denwood, M.J., Nkwengulila, G., Mable, B.K., 2017. Estimating the prevalence and intensity of Schistosoma mansoni infection among rural communities in Western Tanzania: The influence of sampling strategy and statistical approach. PLoS Negl. Trop. Dis. 11, e0005937. https://doi.org/10.1371/journal.pntd.0005937

Bharadwaj, M., Bengtson, M., Golverdingen, M., Waling, L., Dekker, C., 2021. Diagnosing point-of-care diagnostics for neglected tropical diseases. PLoS Negl. Trop. Dis. 15, e0009405. https://doi.org/10.1371/journal.pntd.0009405

Black, M.A., Craig, B.A., 2002. Estimating disease prevalence in the absence of a gold standard. Stat Med 21, 2653–2669. https://doi.org/10.1002/sim.1178

Boelaert, M., El Safi, S., Goetghebeur, E., Gomes-Pereira, S., Le Ray, D., Van der Stuyft, P., 1999. Latent class analysis permits unbiased estimates of the validity of DAT for the diagnosis of visceral leishmaniasis. Trop. Med. Int. Health 4, 395–401. https://doi.org/10.1046/j.1365-3156.1999.00421.x

Boelaert, M., Rijal, S., Regmi, S., Singh, R., Karki, B., Jacquet, D., Chappuis, F., Campino, L., Desjeux, P., Le Ray, D., Koirala, S., Van der Stuyft, P., 2004. A comparative study of the effectiveness of diagnostic tests for visceral leishmaniasis. Am. J. Trop. Med. Hyg. 70, 72–77.

Bolboacă, S.D., 2019. Medical Diagnostic Tests: A Review of Test Anatomy, Phases, and Statistical Treatment of Data. Comput. Math. Methods Med. 2019, 1–22. https://doi.org/10.1155/2019/1891569

Branscum, A.J., Gardner, I.A., Johnson, W.O., 2005. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev. Vet. Med. 68, 145–163. https://doi.org/10.1016/j.prevetmed.2004.12.005

Brooks, M.E., Kristensen, K., Benthem, K.J. van, Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M., 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J. 9, 378–400. https://doi.org/10.32614/RJ-2017-066

Cardoso, J.R., Pereira, L.M., Iversen, M.D., Ramos, A.L., 2014. What is gold standard and what is ground truth? Dent. Press J. Orthod. 19, 27–30. https://doi.org/10.1590/2176-9451.19.5.027-030.ebo

Chala, B., 2023. Advances in diagnosis of Schistosomiasis: Focus on challenges and future approaches. Int. J. Gen. Med. 16, 983–995. https://doi.org/10.2147/IJGM.S391017
Cheung, A., Dufour, S., Jones, G., Kostoulas, P., Stevenson, M.A., Singanallur, N.B., Firestone, S.M., 2021. Bayesian latent class analysis when the reference test is imperfect: -EN- -FR- Analyse bayésienne à classes latentes dans les situations où le test de référence est imparfait -ES- Análisis bayesiano de clases latentes cuando la prueba de referencia es imperfecta. Rev. Sci. Tech. OIE 40. https://doi.org/10.20506/rst.40.1.3224
Chikere, C.M.U., Wilson, K., Graziadio, S., Vale, L., Allen, A.J., 2019. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. PLoS ONE 14, 1–25. https://doi.org/10.1371/journal.pone.0223832
Choi, H.L., Ducker, C., Braniff, S., Argaw, D., Solomon, A.W., Borisch, B., Mubangizi, D., 2022. Landscape analysis of NTD diagnostics and considerations on the development of a strategy for regulatory pathways. PLoS Negl. Trop. Dis. 16, e0010597. https://doi.org/10.1371/journal.pntd.0010597
Claassen, J.A.H.R., 2005. The gold standard: not a golden standard. BMJ 330, 1121. https://doi.org/10.1136/bmj.330.7500.1121
Coffeng, L.E., Vlaminck, J., Cools, P., Denwood, M., Albonico, M., Ame, S.M., Ayana, M., Dana, D., Cringoli, G., de Vlas, S.J., Fenwick, A., French, M., Kazienga, A., Keiser, J., Knopp, S., Leta, G., Matoso, L.F., Maurelli, M.P., Montresor, A., Mirams, G., Mekonnen, Z., Corrêa-Oliveira, R., Pinto, S.A., Rinaldi, L., Sayasone, S., Steinmann, P., Thomas, E., Vercruysse, J., Levecke, B., 2023. A general framework to support cost-efficient fecal egg count methods and study design choices for large-scale STH deworming programs–monitoring of therapeutic drug efficacy as a case study. PLoS Negl. Trop. Dis. 17, 1–23. https://doi.org/10.1371/journal.pntd.0011071
Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104
Coker, S.M., Box, E.K., Stilwell, N., Thiele, E.A., Cotton, J.A., Haynes, E., Yabsley, M.J., Cleveland, C.A., 2022. Development and validation of a quantitative PCR for the detection of Guinea worm (Dracunculus medinensis). PLoS Negl. Trop. Dis. 16, e0010830. https://doi.org/10.1371/journal.pntd.0010830
Cook, C., 2012. Challenges with diagnoses: sketchy reference standards. J. Man. Manip. Ther. 20, 111–112. https://doi.org/10.1179/1066981712Z.00000000025
Cooper, R.N., Dornbusch, R., Hall, R.E., 1982. The Gold Standard: Historical Facts and Future Prospects. Brook. Pap. Econ. Act. 1982, 1. https://doi.org/10.2307/2534316
Coulibaly, J.T., N'Goran, E.K., Utzinger, J., Doenhoff, M.J., Dawson, E.M., 2013. A new rapid diagnostic test for detection of anti- *Schistosoma mansoni* and anti- *Schistosoma mansoni* antibodies. Parasit. Vectors 6, 29. https://doi.org/10.1186/1756-3305-6-29
Denwood, M.J., 2016. runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. J. Stat. Softw. 71. https://doi.org/10.18637/jss.v071.i09
Eddyani, M., Sopoh, G.E., Ayelo, G., Brun, L.V.C., Roux, J.J., Barogui, Y., Affolabi, D., Faber, W.R., Boelaert, M., Van Rie, A., Portaels, F., De Jong, B.C., 2018. Diagnostic accuracy of clinical and microbiological signs in patients with skin lesions resembling buruli ulcer in an endemic region. Clin. Infect. Dis. 67, 827–834. https://doi.org/10.1093/cid/ciy197
Engelman, D., Fuller, L.C., Steer, A.C., for the International Alliance for the Control of Scabies Delphi panel, 2018. Consensus criteria for the diagnosis of scabies: A Delphi study of

international experts. PLoS Negl. Trop. Dis. 12, e0006549. https://doi.org/10.1371/journal.pntd.0006549
Engelman, D., Yoshizumi, J., Hay, R.J., Osti, M., Micali, G., Norton, S., Walton, S., Boralevi, F., Bernigaud, C., Bowen, A.C., Chang, A.Y., Chosidow, O., Estrada-Chavez, G., Feldmeier, H., Ishii, N., Lacarrubba, F., Mahé, A., Maurer, T., Mahdi, M.M.A., Murdoch, M.E., Pariser, D., Nair, P.A., Rehmus, W., Romani, L., Tilakaratne, D., Tuicakau, M., Walker, S.L., Wanat, K.A., Whitfeld, M.J., Yotsu, R.R., Steer, A.C., Fuller, L.C., 2020. The 2020 International Alliance for the Control of Scabies Consensus Criteria for the Diagnosis of Scabies. Br. J. Dermatol. 183, 808–820. https://doi.org/10.1111/bjd.18943
Everson-Stewart, S., Emerson, S.S., 2010. Bio-creep in non-inferiority clinical trials. Stat. Med. 29, 2769–2780. https://doi.org/10.1002/sim.4053
Eyayu, T., Yasin, M., Workineh, L., Tiruneh, T., Andualem, H., Sema, M., Damtie, S., Abebaw, A., Getie, B., Andargie, D., Achaw, B., Taklual, W., 2022. Evaluation of urine sample for diagnosis of visceral leishmaniasis using rK-39 immunochromatographic test in Northwest Ethiopia. PLoS ONE 17, e0263696. https://doi.org/10.1371/journal.pone.0263696
Fuss, A., Mazigo, H.D., Tappe, D., Kasang, C., Mueller, A., 2018. Comparison of sensitivity and specificity of three diagnostic tests to detect Schistosoma mansoni infections in school children in Mwanza region, Tanzania. PLoS ONE 13, e0202499. https://doi.org/10.1371/journal.pone.0202499
Gass, K., 2020. Time for a diagnostic sea-change: Rethinking neglected tropical disease diagnostics to achieve elimination. PLoS Negl. Trop. Dis. 14, e0008933. https://doi.org/10.1371/journal.pntd.0008933
Georgiadis, M.P., Johnson, W.O., Gardner, I.A., Singh, R., 2003. Correlation-Adjusted Estimation of Sensitivity and Specificity of Two Diagnostic Tests. J. R. Stat. Soc. Ser. C Appl. Stat. 52, 63–76.
Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45, 3–22. https://doi.org/10.1016/S0167-5877(00)00114-8
Gurung, P., Gomes, C.M., Vernal, S., Leeflang, M.M.G., 2019. Diagnostic accuracy of tests for leprosy: a systematic review and meta-analysis. Clin. Microbiol. Infect. 25, 1315–1327. https://doi.org/10.1016/j.cmi.2019.05.020
Harding-Esch, E.M., Holland, M.J., Schémann, J.-F., Molina, S., Sarr, I., Andreasen, A.A., Roberts, C. h, Sillah, A., Sarr, B., Harding, E.F., Edwards, T., Bailey, R.L., Mabey, D.C.W., 2011. Diagnostic Accuracy of a Prototype Point-of-Care Test for Ocular Chlamydia trachomatis under Field Conditions in The Gambia and Senegal. PLoS Negl. Trop. Dis. 5, e1234. https://doi.org/10.1371/journal.pntd.0001234
Hawkins, D.M., Garrett, J.A., Stephenson, B., 2001. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat. Med. 20, 1987–2001. https://doi.org/10.1002/sim.819
Helman, S.K., Mummah, R.O., Gostic, K.M., Buhnerkempe, M.G., Prager, K.C., Lloyd-Smith, J.O., 2020. Estimating prevalence and test accuracy in disease ecology: How Bayesian latent class analysis can boost or bias imperfect test results. Ecol. Evol. 10, 7221–7232. https://doi.org/10.1002/ece3.6448
Herbinger, K.-H., Adjei, O., Awua-Boateng, N.-Y., Nienhuis, W.A., Kunaa, L., Siegmund, V., Nitschke, J., Thompson, W., Klutse, E., Agbenorku, P., Schipf, A., Reu, S., Racz, P., Fleischer, B., Beissner, M., Fleischmann, E., Helfrich, K., van der Werf, T.S., Lüscher, T., Bretzel, G., 2009. Comparative Study of the Sensitivity of Different Diagnostic Methods for the Laboratory Diagnosis of Buruli Ulcer Disease. Clin. Infect. Dis. 48, 1055–1064. https://doi.org/10.1086/597398

Hoekstra, P.T., Madinga, J., Lutumba, P., Van Grootveld, R., Brienen, E.A.T., Corstjens, P.L.A.M., Van Dam, G.J., Polman, K., Van Lieshout, L., 2022. Diagnosis of Schistosomiasis without a Microscope: Evaluating Circulating Antigen (CCA, CAA) and DNA Detection Methods on Banked Samples of a Community-Based Survey from DR Congo. Trop. Med. Infect. Dis. 7, 315. https://doi.org/10.3390/tropicalmed7100315

Hui, S.L., Walter, S.D., 1980. Estimating the Error Rates of Diagnostic Tests. Biometrics 36, 167. https://doi.org/10.2307/2530508

Humbert, M.V., Costa, L.E., Katis, I., Fonseca Ramos, F., Sanchéz Machado, A., Sones, C., Ferraz Coelho, E.A., Christodoulides, M., 2019. A rapid diagnostic test for human Visceral Leishmaniasis using novel *Leishmania* antigens in a Laser Direct-Write Lateral Flow Device. Emerg. Microbes Infect. 8, 1178–1185. https://doi.org/10.1080/22221751.2019.1635430

Johnson, W.O., Jones, G., Gardner, I.A., 2019. Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. Prev. Vet. Med. 167, 113–127. https://doi.org/10.1016/j.prevetmed.2019.01.010

Jones, G., Johnson, W.O., Hanson, T.E., Christensen, R., 2010. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics 66, 855–863. https://doi.org/10.1111/j.1541-0420.2009.01330.x

Kazienga, A., Coffeng, L.E., De Vlas, S.J., Levecke, B., 2022. Two-stage lot quality assurance sampling framework for monitoring and evaluation of neglected tropical diseases, allowing for imperfect diagnostics and spatial heterogeneity. PLoS Negl. Trop. Dis. 16, e0010353. https://doi.org/10.1371/journal.pntd.0010353

Keddie, S.H., Baerenbold, O., Keogh, R.H., Bradley, J., 2023. Estimating sensitivity and specificity of diagnostic tests using latent class models that account for conditional dependence between tests: a simulation study. BMC Med. Res. Methodol. 23, 58. https://doi.org/10.1186/s12874-023-01873-0

Kostoulas, P., Nielsen, S.S., Branscum, A.J., Johnson, W.O., Dendukuri, N., Dhand, N.K., Toft, N., Gardner, I.A., 2017. STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models. Prev. Vet. Med. 138, 37–47. https://doi.org/10.1016/j.prevetmed.2017.01.006

Koukounari, A., Moustaki, I., Grassly, N.C., Blake, I.M., Basáñez, M.-G., Gambhir, M., Mabey, D.C.W., Bailey, R.L., Burton, M.J., Solomon, A.W., Donnelly, C.A., 2013. Using a nonparametric multilevel Latent markov model to evaluate diagnostics for Trachoma. Am. J. Epidemiol. 177, 913–922. https://doi.org/10.1093/aje/kws345

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M., 2016. **TMB** : Automatic Differentiation and Laplace Approximation. J. Stat. Softw. 70. https://doi.org/10.18637/jss.v070.i05

Lemma, G.W., Müller, O., Reñosa, M.D., Lu, G., 2020. Challenges in the last mile of the global guinea worm eradication program. Trop. Med. Int. Health 25, 1432–1440. https://doi.org/10.1111/tmi.13492

Linzer, D.A., Lewis, J.B., 2011. poLCA: An R Package for Polytomous Variable Latent Class Analysis. J. Stat. Softw. 42, 1–29.

Liu, Y., Ying, G., Quinn, G.E., Zhou, X., Chen, Y., 2022. Extending Hui-Walter framework to correlated outcomes with application to diagnosis tests of an eye disease among premature infants. Stat. Med. 41, 433–448. https://doi.org/10.1002/sim.9269

Machado de Assis, T.S., Rabello, A., Werneck, G.L., 2012. Latent class analysis of diagnostic tests for visceral leishmaniasis in Brazil. Trop. Med. Int. Health 17, 1202–1207. https://doi.org/10.1111/j.1365-3156.2012.03064.x

Macpherson, E.E., Adams, E.R., Bockarie, M.J., Hollingsworth, T.D., Kelly-Hope, L.A., Lehane, M., Kovacic, V., Harrison, R.A., Paine, M.J., Reimer, L.J., Torr, S.J., 2015. Mass Drug Administration and beyond: how can we strengthen health systems to deliver complex interventions to eliminate neglected tropical diseases? BMC Proc. 9, S7. https://doi.org/10.1186/1753-6561-9-S10-S7

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. 8, 1235–1246. https://doi.org/DOI 10.1111/j.1461-0248.2005.00826.x

Mbui, J., Wasunna, M., Balasegaram, M., Laussermayer, A., Juma, R., Njenga, S.N., Kirigi, G., Riongoita, M., de la Tour, R., van Peteghem, J., Omollo, R., Chappuis, F., 2013. Validation of two rapid diagnostic tests for visceral leishmaniasis in Kenya. PLoS Negl. Trop. Dis. 7, e2441. https://doi.org/10.1371/journal.pntd.0002441

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Medica 22, 276–282.

McKenna, S.L.B., Dohoo, I.R., 2006. Using and Interpreting Diagnostic Tests. Vet. Clin. North Am. Food Anim. Pract. 22, 195–205. https://doi.org/10.1016/j.cvfa.2005.12.006

Menten, J., Boelaert, M., Lesaffre, E., 2008a. Bayesian latent class models with conditionally dependent diagnostic tests: A case study. Stat. Med. 27, 4469–4488. https://doi.org/10.1002/sim.3317

Menten, J., Boelaert, M., Lesaffre, E., 2008b. Bayesian latent class models with conditionally dependent diagnostic tests: A case study. Stat. Med. 27, 4469–4488. https://doi.org/10.1002/sim.3317

Mesquita, S.G., Caldeira, R.L., Favre, T.C., Massara, C.L., Beck, L.C.N.H., Simões, T.C., de Carvalho, G.B.F., dos Santos Neves, F.G., de Oliveira, G., de Souza Barbosa Lacerda, L., de Almeida, M.A., dos Santos Carvalho, O., Moraes Mourão, M., Oliveira, E., Silva-Pereira, R.A., Fonseca, C.T., 2022. Assessment of the accuracy of 11 different diagnostic tests for the detection of *Schistosomiasis mansoni* in individuals from a Brazilian area of low endemicity using latent class analysis. Front. Microbiol. 13.

Meyer, T., 2016. Diagnostic procedures to detect *Chlamydia trachomatis* Infections. Microorganisms 4, 25. https://doi.org/10.3390/microorganisms4030025

Monaghan, T.F., Rahman, S.N., Agudelo, C.W., Wein, A.J., Lazar, J.M., Everaert, K., Dmochowski, R.R., 2021. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (Mex.) 57, 503. https://doi.org/10.3390/medicina57050503

Mueller, Y.K., Bastard, M., Nkemenang, P., Comte, E., Ehounou, G., Eyangoh, S., Rusch, B., Tabah, E.N., Trellu, L.T., Etard, J.-F., 2016. The "Buruli Score": Development of a multivariable prediction model for diagnosis of *Mycobacterium ulcerans* infection in individuals with ulcerative skin lesions, Akonolinga, Cameroon. PLoS Negl. Trop. Dis. 10, e0004593. https://doi.org/10.1371/journal.pntd.0004593

Mulla, S.M., Scott, I.A., Jackevicius, C.A., You, J.J., Guyatt, G.H., 2012. How to Use a Noninferiority Trial. JAMA 308, 2605. https://doi.org/10.1001/2012.jama.11235

Orish, V.N., Morhe, E.K.S., Azanu, W., Alhassan, R.K., Gyapong, M., 2022. The parasitology of female genital schistosomiasis. Curr. Res. Parasitol. Vector-Borne Dis. 2, 100093. https://doi.org/10.1016/j.crpvbd.2022.100093

Pateras, K., Kostoulas, P., 2023. PriorGen: Generates Prior Distributions for Proportions.

Pedrosa, C.M.S., Ximenes, R.A. de A., de Almeida, W.A.P., da Rocha, E.M.M., 2013. Validity of the polymerase chain reaction in the diagnosis of clinically suspected cases of American

visceral leishmaniasis. Braz. J. Infect. Dis. 17, 319–323. https://doi.org/10.1016/j.bjid.2012.10.021
Peeling, R.W., Mabey, D., 2014. Diagnostics for the control and elimination of neglected tropical diseases. Parasitology 141, 1789–1794. https://doi.org/10.1017/S0031182014000973
Plummer, M., 2003. JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling JAGS : Just Another Gibbs Sampler, in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). p. March 20-22,Vienna, Austria. ISSN 1609-395X. https://doi.org/10.1.1.13.3406
Qu, Y., Tan, M., Kutner, M.H., 1996. Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. Biometrics 52, 797. https://doi.org/10.2307/2533043
R Core Team, 2023. R: A Language and Environment for Statistical Computing.
Rezaei, Z., Pourabbas, B., Kühne, V., Pourabbas, P., Büscher, P., 2022. Diagnostic performance of three rK39 rapid diagnostic tests and two direct agglutination tests for the diagnosis of visceral Leishmaniasis in Southern Iran. J. Trop. Med. 2022, 3569704. https://doi.org/10.1155/2022/3569704
Röltgen, K., Cruz, I., Ndung'u, J.M., Pluschke, G., 2019. Laboratory diagnosis of Buruli Ulcer: challenges and future perspectives, in: Pluschke, G., Röltgen, K. (Eds.), Buruli Ulcer: Mycobacterium Ulcerans Disease. Springer, Cham (CH).
Rutjes, A., Reitsma, J., Coomarasamy, A., Khan, K., Bossuyt, P., 2007. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol. Assess. 11. https://doi.org/10.3310/hta11500
Salam, Md.A., Huda, M.M., Khan, Md.G.M., Shomik, M.S., Mondal, D., 2021. Evidence-based diagnostic algorithm for visceral leishmaniasis in Bangladesh. Parasitol. Int. 80, 102230. https://doi.org/10.1016/j.parint.2020.102230
Schiller, I., Van Smeden, M., Hadgu, A., Libman, M., Reitsma, J.B., Dendukuri, N., 2016. Bias due to composite reference standards in diagnostic accuracy studies. Stat. Med. 35, 1454–1470. https://doi.org/10.1002/sim.6803
See, C.W., Alemayehu, W., Melese, M., Zhou, Z., Porco, T.C., Shiboski, S., Gaynor, B.D., Eng, J., Keenan, J.D., Lietman, T.M., 2011. How Reliable Are Tests for Trachoma?—A Latent Class Approach. Invest. Ophthalmol. Vis. Sci. 52, 6133–6137. https://doi.org/10.1167/iovs.11-7419
Singh, H., 2014. Editorial: Helping Health Care Organizations to Define Diagnostic Errors as Missed Opportunities in Diagnosis. Jt. Comm. J. Qual. Patient Saf. 40, 99-AP1. https://doi.org/10.1016/S1553-7250(14)40012-6
Smits, N., 2010. A note on Youden's Jand its cost ratio. BMC Med. Res. Methodol. 10, 89. https://doi.org/10.1186/1471-2288-10-89
Snow, G., 2020. TeachingDemos: Demonstrations for Teaching and Learning.
Solomon, A.W., Peeling, R.W., Foster, A., Mabey, D.C.W., 2004. Diagnosis and assessment of Trachoma. Clin. Microbiol. Rev. 17, 982–1011. https://doi.org/10.1128/CMR.17.4.982-1011.2004
Stærk-Østergaard, J., Kirkeby, C., Christiansen, L.E., Andersen, M.A., Møller, C.H., Voldstedlund, M., Denwood, M.J., 2022. Evaluation of diagnostic test procedures for SARS-CoV-2 using latent class models. J. Med. Virol. 1–8. https://doi.org/10.1002/jmv.27943
Stan Development Team, 2023. RStan: the R interface to Stan.
Straily, A., Kavere, E.A., Wanja, D., Wiegand, R.E., Montgomery, S.P., Mwaki, A., Eleveld, A., Secor, W.E., Odiere, M.R., 2022. Evaluation of the point-of-care circulating Cathodic antigen assay for monitoring mass drug administration in a *Schistosoma mansoni* Control

program in Western Kenya. Am. J. Trop. Med. Hyg. 106, 303–311. https://doi.org/10.4269/ajtmh.21-0599
Tamarozzi, F., Guevara, Á.G., Anselmi, M., Vicuña, Y., Prandi, R., Marquez, M., Vivero, S., Robinzón Huerlo, F., Racines, M., Mazzi, C., Denwood, M., Buonfrate, D., 2023. Accuracy, acceptability, and feasibility of diagnostic tests for the screening of Strongyloides stercoralis in the field (ESTRELLA): a cross-sectional study in Ecuador. Lancet Glob. Health 1–9. https://doi.org/10.1016/S2214-109X(23)00108-0
Taylor, E.M., 2020. NTD diagnostics for disease elimination: A Review. Diagnostics 10, 375. https://doi.org/10.3390/diagnostics10060375
Thibodeau, L.A., 1981. Evaluating Diagnostic Tests. Biometrics 37, 801. https://doi.org/10.2307/2530161
Toft, N., Innocent, G.T., Gettinby, G., Reid, S.W.J., 2007. Assessing the convergence of Markov Chain Monte Carlo methods: an example from evaluation of diagnostic tests in absence of a gold standard. Prev Vet Med 79, 244–256. https://doi.org/10.1016/j.prevetmed.2007.01.003
Toft, N., Jørgensen, E., Højsgaard, S., 2005. Diagnosing diagnostic tests: Evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev. Vet. Med. 68, 19–33. https://doi.org/10.1016/j.prevetmed.2005.01.006
Utzinger, J., Becker, S.L., Lieshout, L., Dam, G.J., Knopp, S., 2015. New diagnostic tools in schistosomiasis. Clin. Microbiol. Infect. 21, 529–542. https://doi.org/10.1016/j.cmi.2015.03.014
van der Werf, T.S., 2018. Diagnostic tests for Buruli Ulcer: Clinical judgment revisited. Clin. Infect. Dis. 67, 835–836. https://doi.org/10.1093/cid/ciy203
Versi, E., 1992. “Gold standard” is an appropriate term. BMJ 305, 187–187. https://doi.org/10.1136/bmj.305.6846.187-b
Walker, E., Nowacki, A.S., 2011. Understanding equivalence and noninferiority testing. J. Gen. Intern. Med. 26, 192–196. https://doi.org/10.1007/s11606-010-1513-8
Walker, S.L., Collinson, S., Timothy, J., Zayzay, S.K., Kollie, K.K., Candy, N., Lebas, E., Halliday, K., Pullan, R., Fallah, M., Marks, M., 2020. A community-based validation of the International Alliance for the Control of Scabies Consensus Criteria by expert and non-expert examiners in Liberia. PLoS Negl. Trop. Dis. 14, e0008717. https://doi.org/10.1371/journal.pntd.0008717
Walter, S., 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. 41, 923–937. https://doi.org/10.1016/0895-4356(88)90110-2
Wang, Z., Dendukuri, N., Zar, H.J., Joseph, L., 2017. Modeling conditional dependence among multiple diagnostic tests. Stat. Med. 36, 4843–4859. https://doi.org/10.1002/sim.7449
Wiegand, R.E., Cooley, G., Goodhew, B., Banniettis, N., Kohlhoff, S., Gwyn, S., Martin, D.L., 2018. Latent class modeling to compare testing platforms for detection of antibodies against the Chlamydia trachomatis antigen Pgp3. Sci. Rep. 8, 4232. https://doi.org/10.1038/s41598-018-22708-9
Yang, I., Becker, M.P., 1997. Latent Variable Modeling of Diagnostic Accuracy. Biometrics 53, 948. https://doi.org/10.2307/2533555
Yang, J.L., Hong, K.C., Schachter, J., Moncada, J., Lekew, T., House, J.I., Zhou, Z., Neuwelt, M.D., Rutar, T., Halfpenny, C., Shah, N., Whitcher, J.P., Lietman, T.M., 2009. Detection of Chlamydia trachomatis Ocular Infection in Trachoma-Endemic Communities by rRNA Amplification. Invest. Ophthalmol. Vis. Sci. 50, 90–94. https://doi.org/10.1167/iovs.08-2247

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Appendix: Flow Chart