Point-of-care tests for syphilis: Meta-analysis and systematic review

Yalda Jafari, Hon. B.Sc.

Master of Science
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University
Montreal, Quebec, Canada

Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science

April 2011

ABSTRACT

Background: The World Health Organization estimates there were 12 million new cases of syphilis in 2006. . In developing countries there is often a lack of proper screening due to limited laboratory services. In contrast, in developed countries there is often limited access to care among hard-to-reach populations. In context of these healthcare system disconnects, point-of-care (POC) tests have proven to be an invaluable resource, yet in order to justify their use, their diagnostic accuracy and implementation outcomes must first be established.

Methods: We searched six electronic databases from 1 January 1980 to 24 September 2010 for articles evaluating syphilis POC tests. Data was extracted and a second reviewer independently reviewed a subset of the articles. Subgroups of studies were created according to index test, sample, and reference standard employed. Pooled sensitivity and specificity estimates were calculated using Hierarchical Summary Receiver Operating Characteristic (HSROC) curves. Adjustments were made to account for imperfect reference standards. A narrative review of implementation outcomes was undertaken.

Results: The most frequently evaluated kits were Determine® (29%), Bioline® (18%), Syphicheck® (15%), and Visitect® (14%). After adjustment for imperfect reference standard, in serum samples, using a TP (*Treponemal Pallidum*) specific reference standard (e.g. TPPA), Bioline® had the highest pooled sensitivity, 99.67% (95% credible interval 97.65, 100) , followed by Determine® , 99.14% (96.93, 100), Visitect®, 98.18% (93.53, 100) and Syphicheck®, 88.46% (73.54, 99.87). Syphicheck® had the highest

pooled specificity, 99.98% (99.64, 100), followed by Visitect®, 99.89% (99.19, 100), Determine®, 99.68% (98.70, 100) and Bioline®, 99.56% (98.55, 100). In whole blood, Bioline® had the highest pooled sensitivity, 91.47% (87.06, 96.12), followed by Determine®, 89.49% (79.88, 98.15), Visitect®, 82.93% (94.50, 100) and Syphicheck®, 81.99% (71.84, 91.99). Determine® had the highest pooled specificity, 99.91% (99.44, 100) followed by Visitect®, 99.87% (99.58, 100) followed by Syphicheck®, 99.81% (99.46, 100), and Bioline®, 99.61% (99.04, 100). Acceptability, feasibility, and impact of POC tests were demonstrated in various studies. Preference was not well established and economic evaluations were too heterogeneous to be conclusive.

Conclusion: Bioline® and Determine® had the highest estimates of pooled sensitivity and specificity respectively. Higher parameter estimates in serum warrant the use of these tests in serum, rather than whole blood where feasible. Comparing our findings to current strategies in place, it is appropriate to use POC tests to screen for syphilis where access to laboratories and laboratory based serological tests are limited or where patients do not return for results. Further research into implementation outcomes is warranted and a framework for evaluating these outcomes is urgently needed.

RÉSUMÉ ANALYTIQUE

Contexte: L'organisation mondiale de la santé (OMS) estimait à 12 millions le nombre de nouveaux cas de syphilis en 2006. Les pays en développement sont souvent confrontés à des lacunes en termes de dépistage adéquat, attribuables aux services de laboratoire limités. Pour les pays développés, ce sont les populations marginalisées qui souffrent souvent d'un accès limité aux services de santé. Dans ce contexte d'inégalités des systèmes de santé, et bien qu'on ait déjà démontré que les tests au point d'intervention représentaient une ressource de très grande valeur, leur précision diagnostique et l'analyse de résultats d'implantation (IRO) doivent d'abord être établis, afin de justifier leur utilisation.

Méthodes: Nous avons effectué une recherche d'articles traitant de l'évaluation des tests au point d'intervention pour la syphilis dans six bases de données électroniques, du 1980 au 2010. Deux évaluateurs ont analysé les données. Des sous-groupes ont été créés en fonction des types de tests, échantillons et étalons de référence. Nous avons généré des tests de sensibilité et spécificité mises en commun, à l'aide de courbes Hierarchical Summary Receiver Operating Characteristic (HSROC), et avons ajusté les valeurs pour tenir compte des étalons de référence imparfaits. Nous avons aussi synthétisé de façon narrative les analyses de résultats d'implantation (*IROs*).

Résultats: Après avoir sommairement évalué 64 articles complets, 30 (47%) articles ont été inclus dans la méta-analyse d'exactitude diagnostique, et 24 (38%) articles ont été inclus dans l'examen narratif des analyses de résultats d'implantation (*IROs*). Quatre tests ont été évalués à travers l'ensemble des études : Determine® (29%), Bioline®

(18%), Syphicheck® (15%), et Visitect® (14%). Après ajustement pour tenir compte de l'imperfection des étalons de référence, dans des échantillons de sérum, en utilisant un étalon de référence spécifique pour le TP (*treponema pallidum*) (par exemple, le TPPA), le test Bioline® s'est avéré avoir le test de sensibilité mise en commun le plus élevé, soit 99,67% (95% intervalle crédible 97,65-100), suivi de Determine® avec 99,14% (96,93 - 100), Visitect® avec 98,18% (93,53-100) et Syphicheck® avec 88,46% (73,54-99.87). Syphicheck® a obtenue test de spécificité mise en commun le plus élevé, soit 99,98% (99,64-100), suivi de Visitect® avec 99,89% (99,19 - 100), Determine® avec 99,68% (98,70-100) et Bioline® avec 99,56% (98,55-100). Dans des échantillons sanguins complets, Bioline® a obtenu le test de sensibilité mise en commun le plus élevé, soit 91,47% (87,06-96,12), suivi de Determine® avec 89,49% (79,88-98,15), Visitect® avec 82,93% (94,50-100) et Syphicheck® avec 81,99% (71,84-91,99). Determine® a obtenu le test de spécificité mise en commun le plus élevé, soit 99,91% (99,44-100), suivi de Visitect® avec 99,87% (99,58 - 100), Syphicheck® avec 99,81% (99,46 -100), et Bioline® avec 99,61% (99,04-100). L'acceptabilité, la faisabilité et l'impact des tests au point d'intervention ont aussi été démontrés dans plusieurs études. La préférence n'a pas été suffisamment bien établie, et les évaluations économiques étaient trop hétérogènes pour être concluantes.

Conclusion: Bioline® et Determine® ont obtenu respectivement les tests de sensibilité et spécificité mises en commun les plus élevés. Les tests plus élevés dans les échantillons de sérum suggèrent fortement, lorsque possible, l'utilisation de ces tests dans ce contexte, plutôt qu'avec des échantillons sanguins complets. En comparant nos résultats

aux stratégies de tests actuellement en application, on constate qu'il serait approprié de remplacer les tests sérologiques avec des tests au point d'intervention, même lorsque l'accès à un laboratoire ou à du personnel adéquatement formé n'est pas une problématique. De plus amples recherches sur les analyses de résultats d'implantation sont nécessaires, et un cadre d'étude pour l'évaluation de ces résultats est urgemment requis.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Nitika Pant Pai, for her guidance, for giving me the opportunity to grow as an academic, as well as a person, and, for believing in me. I would like to thank my co-supervisor, Dr. Lawrence Joseph, for his guidance, advice on statistical analysis, as well as for his quick feedback. I would like to thank Dr. Rosanna Peeling for her great insight and for the great dedication she showed to this project. I am grateful to Sushmita Shivkumar for her contribution as a second reviewer of this thesis as well as to the translators and authors who replied to my emails. I would also like to thank Lorie Kloda for the time she invested helping me improve my search string and search strategy. I would also like to thank Dr. Mira Johri and Denis Ako-Arrey for their guidance and introduction to economic evaluations. Finally, I would like to thank my family and friends for always being there for me.

DEDICATION

I would like to dedicate this thesis to family, my parents Maryam Rezaee Boroon and Ahmad Jafari, I will never take in vain your sacrifices and my sisters, Sara and Mitra for brining joy to my life every day. I could not be where I am without your love, support and encouragement. Thank you for believing in me and I hope I have made you proud. Doostetan daram!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANC: Antenatal clinic
FSW: Female sex workers
MSM: Men who have sex with men
POC: Point-of-care
RCT: Randomized Control Trial
TP: *Treponemal pallidum*
USA: United States of America

Chapter 1: INTRODUCTION

Humans are the sole natural host of *Treponemal pallidum* (TP), the causative agent of syphilis. Syphilis transmission occurs most commonly through sexual contact, but can also arise through blood transfusion and in utero. Syphilis manifests itself in three stages: primary, secondary and tertiary. Primary syphilis occurs after an inoculation period of approximately 21 days, when sores appear at the site of infection. Lesions last an average of 4 to 6 weeks after which they spontaneously heal. Secondary syphilis, usually resulting in skin lesions, follows 6 to 8 weeks after the end of the first stage. Within 2 to 6 weeks, these lesions also subside and the infection enters the tertiary stage. This tertiary period is also known as the latent phase because only about 30% of patients develop clinical symptoms.(1)

Lack of awareness concerning one's serostatus is one of the main driving forces of the syphilis epidemic. Approximately 90% of those infected do not know that they are infected.(2) Standard syphilis testing involves the use of non-treponemal tests such as the rapid plasma reagin (RPR) test and Venereal Disease Research Laboratory (VDRL) test followed by confirmation with a treponemal test, such as *Treponema pallidum* particle agglutination (TPPA), *Treponema pallidum* haemagglutination test (TPHA), fluorescent treponemal antibody-absorption (FTA-ABS).(3, 4) These tests require expensive equipment and electricity, and need to be run by trained staff. They usually require up to 100 samples to run, limiting the turnaround time for results. In comparison, point-of-care (POC) tests require no laboratory equipment and minimal

training is often adequate to conduct and interpret results. The test can be run one patient at a time and the results are communicated to the patient within the same visit.(5) For these reasons, POC tests have the potential to increase awareness of infection among patients and promote linkages of testing, referral and treatment, thereby improving patient outcomes and potentially diminishing syphilis transmission. These characteristics can be particularly important in difficult to reach populations and developing countries.

The diagnostic accuracy of several syphilis POC tests has been evaluated by various authors (6-36); however, there is still work to be done synthesizing the evidence in a coherent manner. While diagnostic accuracy remains a key criterion for evaluating diagnostic tests, current GRADE (Grading of Recommendations Assessment, Development and Evaluation) guidelines also emphasize the importance of a wide variety of outcomes related to patients, clinicians and the health systems in shaping recommendations concerning diagnostic tests.(37) This systematic review presents the first comprehensive portrait of evidence on implementation outcomes related to syphilis POC tests. Its aim is to assist researchers in improving their methodology and to identify any shortcomings in the evidence required to make well-informed decisions on the use of syphilis POC tests.

In light of the current situation, this thesis, through a systematic review, has the following objectives:

1. Synthesize and carry out a meta-analysis of the global evidence on the accuracy of commercially available POC diagnostic tests and

2. Summarize the global evidence on the implementation research outcomes (IROs) of commercially available POC diagnostic tests which are currently in use.

Chapter 2 explores the literature, describing syphilis in detail, including its history, epidemiology, pathogenesis, diagnosis and treatment. This chapter delves into diagnostic studies and meta-analyses and describes the characteristics of each type of study. At last, this chapter also reviews previous research in the field. Chapter 3 describes the methods used to conduct the systematic search, the data extraction and the statistical methods employed. Chapter 4 reports the results from the meta-analysis of diagnostic accuracy studies, as well as the results of a narrative review of studies that addressed research questions related to implementation. Chapter 5 is a thorough discussion of these results. The importance for research and practice is highlighted and recommendations for the future are made. Chapter 6 draws overall conclusions from the thesis.

Chapter 2: LITERATURE REVIEW

This chapter aims to provide a background for the reader on syphilis, POC tests and the methodology of diagnostic accuracy studies and meta-analyses. Section 2.1 is a historical review, section 2.2 discusses the epidemiology and observed trends of syphilis while section 2.3 describes the general pathogenesis and various modes of transmission. Section 2.4 delves into the diagnostic tests used and their implications on the patient and the health care system while section 2.5 provides options for treatment and management of the infection. Section 2.6 explores methodology in the field of diagnostic studies and meta-analyses. This section explains accepted methods in the field and the current advances in the statistical methods used to meta-analyze systematic reviews of diagnostic studies. Finally, section 2.7 is an attempt to identify prominent patterns in the conduct of POC diagnostic studies as well as an explanation of how the field has progressed through time.

2.1. Syphilis in History

There are two rival theories that attempt to explain the spread of syphilis. The New World theory states that syphilis was brought back to Europe from the Americas by Columbus and his crew in 1493. (38, 39) Some evidence to support this theory includes the absence of bones with syphilitic lesions until after return of Columbus.(38) On the other hand, the Old World theory proposes that prior to the Columbus era, syphilis was already in Europe and that it was Columbus and his crew who brought syphilis to the Caribbean.(39) Some researchers believe that syphilis has been previously described in ancient Greece and Rome, but not until after the return of Columbus and the

subsequent breakout did the various stages of the disease and the role of sexual relations in its transmission become clear.(38)

Syphilis has been long named the “great imitator” of skin diseases.(39, 40) In an attempt to fully understand the disease and its manifestations, the Tuskegee syphilis experiment was undertaken in Tuskegee, Alabama between the years 1932 to 1972.(41) A series of 399 African American men with syphilis were denied treatment and the natural progression of their disease was studied.(41) In this experiment, which was funded by the United States (US) Public Health Service, the enrolled patients were not aware that they had syphilis nor were they counselled about the disease and how to prevent its transmission to sexual partners.(42) This unethical and medical mistreatment of patients contributed to the growing mistrust of the health care system by African Americans and other minorities.(41, 42) This study is also credited as one of the primary reasons minorities are underrepresented in clinical trials due to a fear that they may be taken advantage of and treated as “guinea pigs”.(42) On a positive note, the study prompted an increased awareness of research ethics and led to the implementation of research boards at federally funded research institutions.(42)

2.2. Epidemiology of Syphilis

Syphilis is found in all parts of the world, although frequencies vary by region. Figure 1, prepared by the Sexually Transmitted Diseases Diagnostics Initiative (SDI) for the Special Programme for Research & Training in Tropical Diseases (TDR) associated with the World Health Organization (WHO), depicts the number of new cases of syphilis

estimated for the year 2006. The group estimates that the largest number of cases occurs in sub-Saharan Africa and South & Southeast Asia with approximately 4 million new cases in each region. In total, 12 million new cases occur every year.(43)

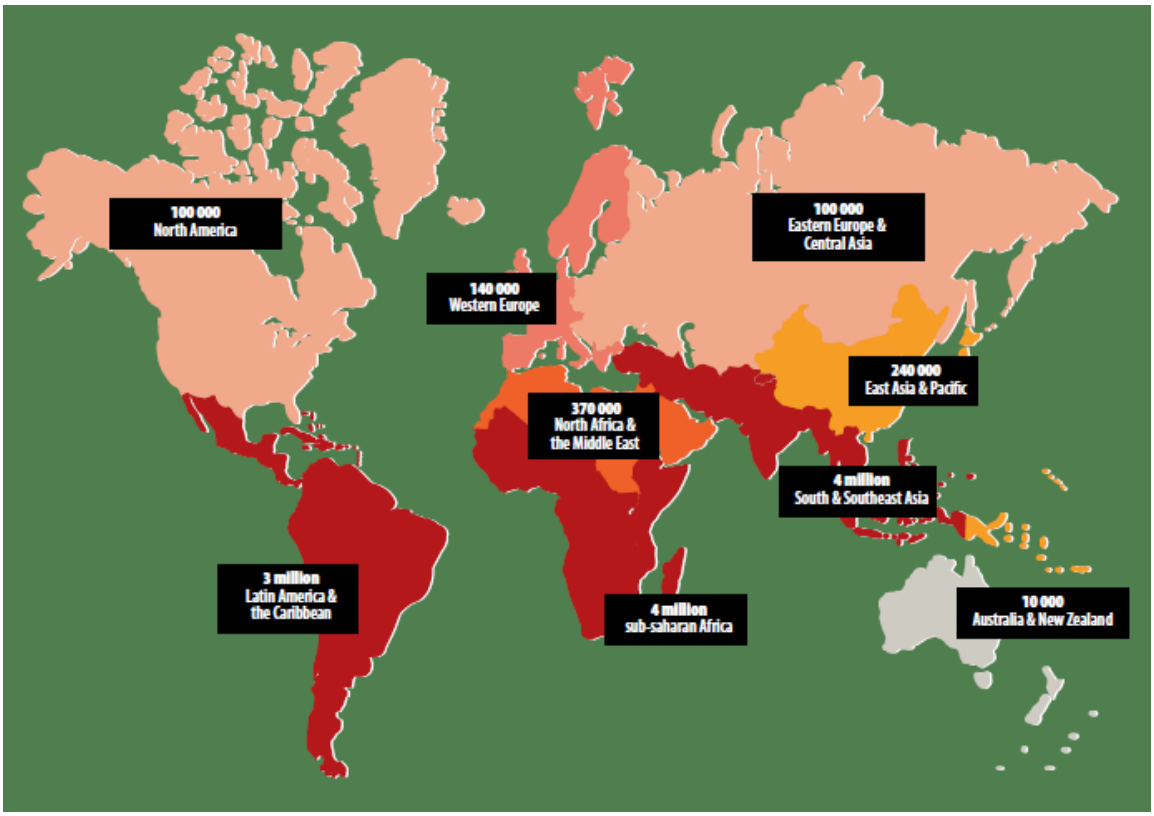

Figure 1: Global map of new cases of syphilis in 2006, as estimated by TDR/WHO

2.2.1. Syphilis in the developed world

The rate of syphilis has followed an interesting pattern in the developed world. Historically the prevalence is cyclic, and in 2000, the rate of syphilis in the US reached an all time low of 2.1 cases per 100,000 persons.(39) Unfortunately, the rates began to rise again from 2000 to 2004.(39, 44) This increase was observed mainly among men, with

men who have sex with men (MSM) contributing to about 60% of the new cases (39). Similar patterns were seen in western European countries around the same time, with the majority of cases again occurring in MSM.(39) In the most recently available estimate by the Centers for Disease Control and Prevention (CDC), as of 2006, 64% of syphilis cases were seen among MSM.(45)

As reported by Public Health Agency of Canada (PHAC), historically the rates of syphilis in men and women have been similar. However, the gap in prevalence between males and females has progressively increased from 1993 to 2008. It is estimated that in 2008, about 86% of cases of syphilis in Canada were in men. This trend is depicted in Figure 2. In Quebec in particular, the male to female rate ratio is 45.7 to 1.0, the highest amongst all provinces and territories in Canada. In areas of reported syphilis outbreak, there are also reported increased rates of congenital syphilis.(46)

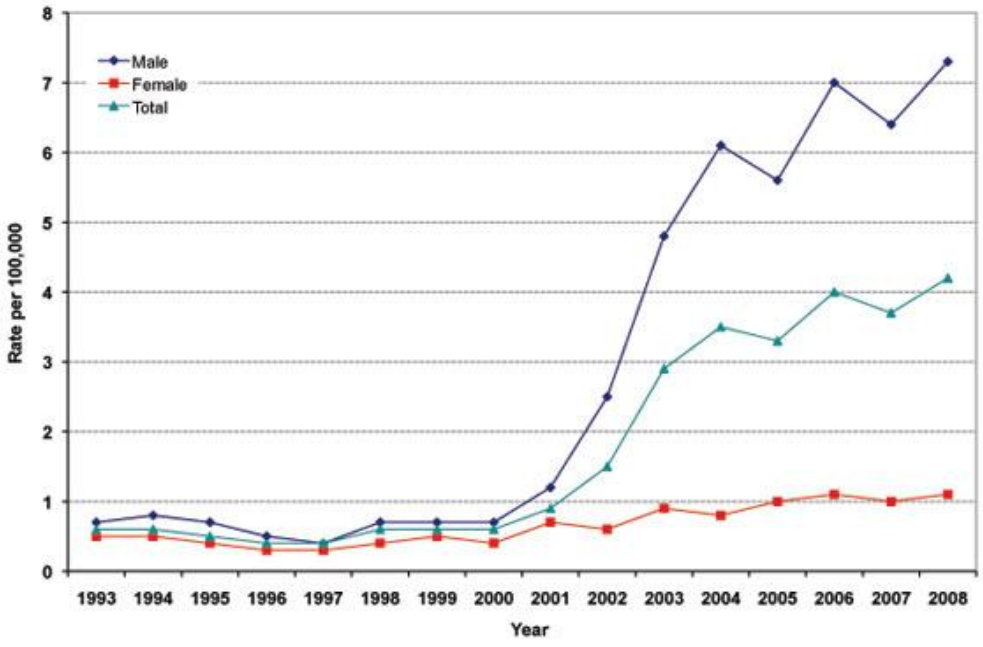

Figure 2: Reported Rates of Infectious Syphilis by Sex and Overall, 1993 to 2008, Canada (46).

2.3. Pathogenesis and Transmission

2.3.1. Pathogenesis

The genus *Treponema pallidum* (TP) belongs to the family *Spirochaetaceae* and has at least four known subspecies. These subspecies include: *pallidum*, which causes syphilis; *endemicum*, which causes bejel or endemic syphilis; *carateum,* which causes pinta; and *pertenue*, which causes yaws.(47) Humans are the sole natural host of TP *pallidum*, the causative agent of syphilis.(1) The bacteria is thin, helical, and cannot be cultivated in vitro.(44) The bacterium can equally infect any organ by entering the lymphatic system or the blood stream.(1)

Stages: Syphilis manifests itself in three stages - primary, secondary and tertiary (also known as latent). Primary syphilis is marked by lesions or sores which appear at the site of infection following an inoculation period of 21 days.(1) These lesions, which are small and painless(45), can present as either single or multiple sores lasting an average of 4 to 6 weeks, after which they spontaneously heal.(1) If the patient receives no treatment, about 6 to 8 weeks post-healing skin lesions or rashes erupt, described as secondary stage syphilis.(1) These rashes appear as red or reddish brown spots and do not cause itching. Other symptoms of secondary syphilis include fever, fatigue, muscle aches, weight loss and hair loss.(45) Within a span of 2 to 6 weeks, the lesions subside. However, without treatment, infection enters the tertiary stage. This stage is also known as latent stage because only about 30% of patients develop clinical symptoms.(1)

The latent stage can manifest itself in 10 to 20 years and can result in range of outcomes such as numbness, paralysis, blindness, auditory abnormalities, meningitis, and dementia.(45, 48) The bacterium extensively damages internal organs such as the nerves, eyes, heart, liver, joints and the brain, and can even cause death.(45)

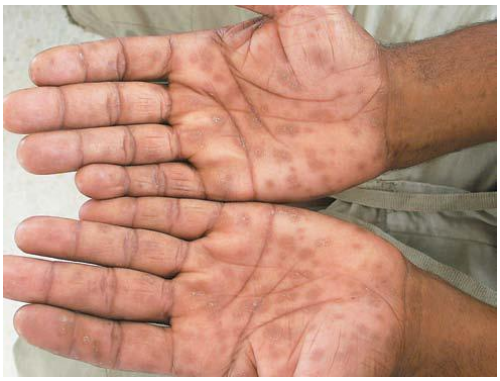

Figure 3: Secondary syphilis rash.(49)

2.3.2. Transmission

The bacterium is passed on from person to person through direct contact with an open sore, a situation which often occurs during sexual contact. The best way to prevent such transmission is to avoid sexual contact with a person who has visible sores.(45) There are other possible modes of transmission as well, including blood transfusion and in utero transmission from mother to infant. TP can cross the placenta and infect the infant throughout the entire pregnancy, but treating the mother before the 4th month usually prevents the infant from becoming infected. This is the optimal situation, as the bacterium can have dire consequences for the infected infant if timely treatment is not

provided to the mother. Depending on the stage of the disease, the risk of mother-to-child transmission can range between 10% in the latent stage up to 90% in primary and secondary syphilis.(44) There is a 40% chance of fetal loss, premature birth, neonatal death or non-fatal congenital syphilis.(1) Congenital syphilis refers to presence of syphilis in utero and birth and symptoms appear from 2 weeks up to 25 years with a wide scale of severity. There are ranges of symptoms such as rash, hepatosplenomegaly and skeletal involvement as well as neurological involvement. Untreated neurological congenital syphilis can be detrimental and result in seizure disorders, cranial nerve palsies and mental retardation. (44)

2.3.2. Syphilis and HIV co-infection

The prognosis of syphilis is altered by co-infections, particularly HIV co-infection. An HIV-mediated decrease in immunity can accelerate the progression of syphilis (39, 50) and syphilis can increase HIV viral load.(50) Additionally, the sores increase the permeability of skin and thus increase HIV transmission.(50) However, evidence of syphilis accelerating HIV progression or HIV accelerating syphilis progression is controversial and limited.

2.4. Diagnosis of Syphilis

This section discusses the various methods to detect TP. It is important to note that none of these methods can distinguish between the various subspecies of TP.(40, 47)

2.4.1. Direct detection of bacteria

Diagnosis is the first step in the control of infection.(51) The method with the highest sensitivity and specificity for detecting syphilis is to confirm the presence of TP from exudates of lesions or tissue under a dark field microscope or using direct fluorescent antibody tests.(44, 48, 52) Dark field microscopy is advantageous because TP can be identified several weeks earlier compared to serological tests.(52) Dark field microscopy should be conducted immediately after sample collection because the sample needs to be observed while still alive in order to distinguish between the different pathogens that may be present in the sample.(52) However, this method is not very practical because it is time consuming, and, in order to achieve the highest benefit of the technique, trained and experienced personnel need to conduct the test.(52) Additionally, the test cannot be performed if no exudates can be collected.(52, 53) Hence, serological tests are of higher practical value.

2.4.2. Serological tests

There are two kinds of serological tests for diagnosis of syphilis: non-treponemal and treponemal tests. Current serological tests indirectly test for the presence of syphilis by testing for antibodies to specific components.(48)

2.4.2.1. Non-treponemal tests

The two most common non-treponemal tests are the Rapid Plasma Reagin test (RPR) and the Venereal Disease Research Laboratory test (VDRL) (3-5). Non-treponemal tests measure levels of Immunoglobulin G (IgG) and Immunoglobulin M (IgM) which are directed against a cardiolipin-lecithin-cholesterol antigen complex. Other bacteria and

medical conditions such as rheumatoid arthritis and malaria can result in increase of anti-cardiolipin antibodies, thus contributing to the probability of a false positive test result for a non-treponemal test. (3-5) Non-treponemal tests exist in both qualitative and quantitative format. Qualitative non-trepnemal tests are used in many settings as a first-line screening strategy. However, their sensitivity is low in early primary and latent syphilis, which can result in a missed diagnosis for roughly one third of patients.(40) Meanwhile, quantitative non-treponemal tests are important in tracking disease progression and treatment response. In order to demonstrate a positive response to treatment, there needs to be a fourfold decrease in the dilution of the titre. Titres from RPR and VDRL tests are not directly comparable and hence, progression should be followed using the same test, and ideally, the same laboratory.(47, 48)

2.4.2.2. Treponemal tests

A treponemal test is typically used for confirmation of non-treponemal tests. (1, 3-5) Some examples of treponemal tests are the FTA-ABS test (Fluorescent Treponemal Antibody-Absorbed), the TP-PA test (*T. pallidum* Particle Agglutination), and the TPHA test (T. pallidum Hemoagglutination). (1, 3-5, 54) These tests are expensive and need to be conducted by laboratory based personnel. They also have a long turnaround time, typically 24 hours to one week depending on the setting, which often provides an opportunity for loss to follow-up.(47, 55)

2.4.3. POC tests

2.4.3.1. Justification for use

In resource limited settings, there is a lack of proper laboratory infrastructure for disease diagnosis: there is limited access to equipment and even if the equipment exists, there is often lack of adequate electricity to run the machines. There are also few to none trained laboratory technicians available to run the tests and interpret results. In such settings, health care professionals are not sufficient in number for the population they serve and are often overburdened with patients. As a result, they usually do not collect all the samples needed to obtain a proper diagnosis, and even if the required samples are collected, transporting the samples to a laboratory facility presents its own unique challenges, and transport is not usually guaranteed.(56) From the point of view of the patient, there are long distances to health clinics, and long waits to see a care provider. Once they have provided their samples, subjects need to return in about 7-10 days for their results, if available.(56) All this translates into a very inefficient system in which the progression of disease and transmission to others is guaranteed for the vast majority of cases. The need for a complete infrastructure makeover to the healthcare system in such settings is undeniable. However, for the short term, point-of-care (POC) diagnostics provide a very promising solution.

Not knowing one's serostatus is one of the main driving forces of the syphilis epidemic worldwide. It is estimated that as high as 90% of those infected are unaware of their status.(2) High-risk populations are often of lower socioeconomic status and thus limited access to health care, as well as high costs may deter individuals from testing. By linking testing, referral and treatment, there is greater opportunity to prevent the

transmission of Sexually Transmitted Infections (STIs) from high-risk populations to the general population.(51) New POC tests which can reach remote populations and provide results within one visit are extremely important to maintaining the momentum in this critical linkage. Most resource-limited settings lack access to laboratories, and when these are available, patients may find it difficult to return for their results due to real life constraints, thus breaking the chain.(57)

In a study conducted by Gift et al in 1999, the treatment rate of patients tested using a POC with sensitivity of 63% and PCR with sensitivity of 94% was evaluated. The authors found that if the return rate is as low as 65%, more patients receive treatment using POC versus PCR tests despite the higher sensitivity of PCR.(58) This explains the impact POC tests can have by eliminating potential loss to follow-up. Also, in order to perform laboratory tests, blood needs to be drawn. This can be uncomfortable for some and also more costly as phlebotomists and lab technicians are required.(5) In contrast, POC tests need only a single drop of blood obtained using a simple finger prick, a process which can be done quickly with almost no pain and which requires minimal training of staff and clinicians.(5) This illustrates just a few of the advantages of POC tests in settings where follow-up is an issue.

2.4.3.2. Role of the World Health Organization

The Sexually Transmitted Diseases Diagnostics Initiative (SDI) of the WHO has set out criteria in evaluating POC tests. The tests have to follow the ASSURED guidelines.

ASSURED is an acronym for: Affordable, Sensitive, Specific, User-friendly (can be performed in a few simple steps with minimal training), Rapid and robust, Equipment-free and Deliverable to developing countries.(3, 5) The SDI has stressed the importance for novel POC tests to address the gap in coverage as developing countries often lack the infrastructure to conduct laboratory tests. In places where such facilities exist, it is often too expensive to be accessible by those most in need. A World Bank report published in 2004, titled World Development Report, attributed the failing healthcare services in the developing world to lack of access and high costs.(5) POC tests have the potential to eliminate these impedances.

2.4.3.3. POC tests for syphilis

The POC tests which are currently in use for syphilis detect treponemal antibodies**.** POC tests use either a lateral-flow or flow-through technology. The majority of tests are lateral flow tests, also known as immune-chromatographic tests (ICT), which have a strip format.(27) Using capillary action, antibodies from an infected person flow over the strip and bind to the treponemal antigens on the strip, inducing a visual change which allows for detection of infection.(54) Meanwhile, flow-through technology utilizes a device with multiple membrane layers which allow the sample and reagent to flow vertically. Using the same concept as ICT, the antibody binds to the antigen on the membrane. The visual signal, however, is a spot instead of a line.(10) It is important to note that these treponemal tests remain active even after treatment because the antibody to the bacterium remains in the body, whereas non-treponemal tests are nonreactive

following treatment. Therefore, treponemal tests are not useful in distinguishing between treated and untreated infections that require treatment. This proves problematic in regions with a high prevalence of syphilis as those who have been previously treated will be falsely classified as falsely positive if tested using POC tests.(5)

2.5. Treatment and Management of Syphilis

2.5.1. Resource Limited Settings

The WHO has recommended syndromic management of symptoms in areas where there is no access to reliable diagnostic services. While this method is advantageous in diagnosing STIs in men, it is both non-sensitive and non-specific in women.(5) As a result, there is a possibility of overtreatment, which can lead to resistance. Additionally, stigma, and particularly for women, domestic violence, can often follow a positive diagnosis. (28) POC tests can help improve detection and thus reduce both the overtreatment and the stigmatization associated with being falsely diagnosed. (57) Errors in diagnosis and treatment can also result in an infection going unchecked, potentially resulting in morbidity and mortality.(56) A false negative diagnosis is especially detrimental in pregnant women, for a misdiagnosis or unknown status in a pregnant woman can have repercussions long past the pregnancy itself, forever affecting the life of the infant to come.

2.5.2. Available Treatment

Penicillin G administered parenterally is the standard treatment for syphilis at any stage as well as in pregnant women.(44, 48, 59) The appropriate dose, duration and type of

penicillin is determined by the stage of the disease and clinical symptoms.(39, 48) A single dose of benzathine Pencillin G administered directly to the muscle is recommended for primary, secondary, and early latent syphilis cases (39, 44). For late latent cases, defined as cases of syphilis with a duration of more than one year, the recommended treatment regimen is three doses administered intramuscularly within a three week period.(44) If a patient has an allergy to penicillin, CDC recommends desensitization of allergy followed by adequate treatment.(39) Non-treponemal tests should then be used to monitor treatment response.(39) Fortunately, there have been no documented cases of resistance to penicillin. (50) This situation should be taken advantage of and systematic testing and treatment should be employed in order to eventually eradicate syphilis.

2.6. Methods of Diagnostic Literature and Meta-analyses

2.6.1. Diagnostic studies

2.6.1.1. Basic design

Epidemiologic studies aimed at evaluating diagnostic assays and outcomes associated with their use and implementation fall under the single umbrella term “diagnostic studies”. The first level of evaluation of an assay in a real life setting is accuracy. Accuracy studies aim to determine how good a test, also known as an index test, performs compared to an already established test, known as a gold or reference standard. Accuracy is determined through two measures: sensitivity and specificity. Sensitivity is the proportion, expressed as a percentage, of those with a disease who test positive. The formula to calculate sensitivity is the number of true positives (TP) over the

total number of true positives and false negatives (FN) (i.e., TP/ (TP+FN)). Specificity on the other hand, is the proportion, expressed as a percentage, of those without a disease who test negative. The formula to calculate specificity is the number of true negatives (TN) over the number of true negatives and false positives (FP), (i.e., TN/ (TN+FP)).(60) In this manner, sensitivity and specificity theoretically do not depend on prevalence.(61)

Other important measures that are particularly helpful to clinicians are positive predictive value and negative predictive value. Positive predictive value (PPV) is the proportion of those who tested positive with the index test who truly have the disease, and negative predictive value (NPV) is the proportion of individuals who tested negative with the index test who are truly negative. The respective formulas are: PPV= (TP/(TP+FP)) and NPV=(TN/(TN+FN)). The predictive values depend on both the sensitivity and specificity of the test, as well as the prevalence of the disease in the population.(61)

Accuracy studies can also evaluate repeatability, concordance and inter-rater reliability. Typically, these studies are cross-sectional in nature, conducted as cross-sectional representative spectrum studies or as part of surveys. The optimal method to evaluate a diagnostic test is to conduct the study on patients whose disease severity is representative of those that the test is intended for in a cross-sectional design.(60, 62) Failure to follow this results in spectrum bias (60) which in turn affects the

generalizability of the results. A minority of studies are case control in nature but these are biased because they sample individuals from extreme ends of the spectrum.

2.6.1.2. Outcomes beyond accuracy

Studies beyond accuracy measure patient-centered outcomes, or implementation outcomes. Although there is no consensus in the diagnostics literature on which study design should be used for this, the simplest design is a randomized controlled trial (RCT). RCTs are also utilized in evaluating diagnostics for impact, but this is rare. Other potential study designs include prospective cohorts, which involve testing at one point in time, and historical control and intervention studies. However, in settings with low prevalence, a large number of individuals must be sampled or randomized in order to obtain a sufficient sample size for generating data on outcomes. Therefore, clinic-based cross-sectional study designs with follow-up on a cohort, as well as the use of clinic-based data, continue to be popular.(63) This inconsistency, combined with a lack of information beyond accuracy to assist in recommendation making, led to the development of GRADE in 2000, and, more specifically, a set of guidelines for diagnostic studies in order to improve on conduct and reporting of studies focused on patient-centered outcomes. Employing such a framework helps clinicians and scientists more fully understand the consequences of diagnosis on the patient.(37)

Cornell et al (64), in an editorial in the Annals of Internal Medicine, elude to the importance of research beyond the diagnostic accuracy of a test, asserting that diagnostic tests should not be solely evaluated on their accuracy but also on their

implementation research outcomes. These outcomes need to be assessed and the harms and benefits of the test for patients, clinicians, and the healthcare system need to be evaluated.(64) The authors also advise authors of systematic reviews to consider the effect of a test in clinical decision making. They also correctly indicate that a systematic review of all studies which evaluate the downstream effects of diagnostic tests will be most useful.(64)

In addition to the call from Cornell et al, there is also a current movement which stresses the importance of research when evaluating the effects of a new diagnostic test when incorporated into practice. Such studies have collectively become known as implementation research in the literature.(65) This movement has been generated in response to the American Agency for Healthcare Research and Quality's (AHRQ) declaration to study implementation in the context of comparative effectiveness research.(65) The authors of this declaration describe the five main steps which need to be undertaken in research in order to allow for the implementation of a desired strategy. The most important of these steps is the need for researchers to fully realize the aim of their implementation research so they may capture what the stakeholder's priority is and then correctly determine what is needed for further exploration. The study should then be designed to capture this outcome. In order to best compare new and old interventions, both qualitative and quantitative methods should be implemented in both the early and later stages of a study so that effect of the new interventions on the populations involved may be assessed. Researchers must also

determine whether the optimal delivery of the new intervention in specific settings is feasible. (65)

2.6.1.2. Biases in diagnostic studies

Diagnostic studies, like all research studies, are prone to bias.(60, 63) The most common biases are: spectrum bias, inappropriate reference standard bias, verification bias, incorporation bias and reviewer bias.(60) Spectrum bias refers to the presence of other conditions and demographics independent of the target disease in a chosen patient sample which are not reflective of those in the intended population. Spectrum biases are present in case-control studies since this study design requires the researcher to actively recruit patients with known serostatus.(66) In order to prevent inappropriate reference standard bias, the reference standard chosen should correctly identify the disease status of the patients and be the best currently available. (67) It is also pertinent that the same reference and index test be utilized on all patients as differential testing based on the result of the index test generates differential verification bias. (66, 67) The reference test and index test should also be independent in order to avoid incorporation bias. Finally, the clinicians analyzing the test results should be blinded to the results of both the index and the reference test when interpreting results in order to prevent reviewer bias. (66)

2.6.2. Methods for Meta-Analysis of Diagnostic Studies

2.6.2.1. Meta-analysis

Meta-analysis is the process of systematically reviewing prior research in a field and statistically combining the results to obtain an overall meaningful measure in order to answer a hypothesis.(60, 61) Studies are identified by means of a thorough systematic review, using keywords to retrieve information from various databases. These keywords are chosen based on the PICO principle (Population, Intervention, Control, and Outcome). (68) The retrieved studies are analyzed and included in the final analysis if they meet the preset inclusion criteria.(60, 61) The measure effects, depending on the study design, are combined using either a risk ratio or odds ratio (60, 61). In the case of diagnostic studies, a simple weighted average of the results is unwarranted and statistical pooling is more complex because there are two measures which are correlated: sensitivity and specificity.(69)

There are currently two state-of–the-art methods in use, both of which use a hierarchical approach. Hierarchical models are advantageous over a binomial regression model as they allow for variability between as well as within studies.(70) These two methods are the Bivariate model and the Hierarchical Summary Receiver Operating Characteristic (HSROC) model.(69, 70) The two models are comparable if there are no covariates added to the model.

In the Bivariate model, sensitivity and specificity are transformed to a logit scale which follows a bivariate normal distribution across studies.(69) The trade-off between the sensitivity and specificity is permitted by including a correlation term. In the HSROC

model, the sensitivity and specificity also follow a binomial distribution and vary across studies. The positivity threshold – that is the threshold used by different tests to denote positive status – and accuracy parameters follow a normal distribution with the mean derived from study level information.(70)

2.6.2.2. Biases in meta-analyses

In addition to the limitations and biases resulting from the individual studies included in the meta-analysis, there are also biases that need to be addressed at the meta-analytic level. For example, there is a possibility of publication bias. Publication bias is a type of study-selection bias and refers to the systematic failure to report or publish studies with certain outcomes.(60) This presents a certain limitation to the conduct of diagnostic meta-analyses as there are currently no statistical methods which assess publication bias in diagnostic studies. While there are current methods with respect to observational studies, these methods cannot be applied to diagnostic studies as they presuppose that studies classify their results based on significance or non-significance as indicated by p-values. Unfortunately, there is no such cut-off values in diagnostic studies and hence, these methods cannot be applied.(68)

Meta-analyses are also prone to review bias. In order to avoid this, two independent reviewers should blindly conduct the literature review, select studies and extract data. The reviewers should also hide the name of the authors and authors' affiliation when possible. This will prevent the author or their affiliation to influence the reviewer's objectivity.(71, 72) Finally, there is the possibility of language bias when only studies in

one language are selected. This typically occurs when English language analyses exclude non-English publications. This results in the exclusion of data from certain regions of the world, thus eliminating any potential contribution these results could have in the decision making process which typically results from meta-analyses. (71, 72)

2.6.3. Quality Evaluation

2.6.3.1. Quality evaluation of individual diagnostic studies

Like other epidemiological methods, there are many criteria which meta-analyses need to follow in order to obtain valid results. In order to facilitate this process for both researchers and reviewers, various tools have been established in order to assess the quality of diagnostic studies, both in their methodology and reporting. The most notable of these tools are the QUADAS (66, 73) and STARD (62) checklists. QUADAS (QUality Assessment of studies of Diagnostic Accuracy included in Systematic reviews) was developed by an expert panel in diagnostic studies. Fourteen items are included in the checklist which addresses sources of bias and variation. Each item is either checked off as yes, no or unclear (66). A second major limitation in assessing the quality of diagnostic studies is reporting. In order to allow for more transparent reporting and in order to better assess the validity and generalizability of the results of diagnostic accuracy studies, STARD (STAndard for Reporting of Diagnostic Accuracy) was developed (62). The STARD statement has 25 items, each of which seeks to determine the presence of bias. Each item is either checked off as reported or not reported. For example, one item enquires as to whether the distribution of disease severity for the disease in question, along with that of other diseases, is reported. The purpose of this item is to

assess for the presence of spectrum bias. In a similar manner, other items are designed to assess the presence of other shortcomings in the diagnostic study such as reviewer and verification bias.(62) It must be noted that the overall score of these checklists is not an incredibly useful tool. Rather, the results should be reported by item as each item has a different level of impact depending on the diagnostic test. Moreover, each item also has a differential effect on the direction of the result, rendering an overall score insensible.(66) As such, equal weighting of all items has been recently discredited by the Cochrane group, a group which instead highly recommends a visual assessment of these tools.(74) It is also important to note that reviewers are encouraged to modify the checklists according to their review.(73)

QUADAS and STARD checklists have been demonstrated to show what their authors claim. Three reviews evaluated 30 articles using the QUADAS checklist in a study by Whiting et al. (73) The study found that the reviewers generally were concordant in evaluating studies using the checklist. STARD checklist has also been evaluated by 2 reviewers using 32 studies.(75) After discussing the disagreements, the disagreements between the reviewers were found to be result of unclear reporting within the studies rather than variations in interpretation of the results and use of checklist.

2.6.3.2. Quality evaluation of meta-analysis

Recently, to improve the quality of meta-analyses and to evaluate their methodology and reporting, and, in order to improve the transparency of systematic reviews, an additional checklist has been developed. This checklist, referred to as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), is being increasingly adopted by top tier journals which publish systematic reviews. (71) This checklist is available in the appendix as Table 1. Developed in 2005 by a panel of experts and consumers, this checklist includes 27 necessary items which assist reviewers in assessing the quality of the systematic reviews and meta-analyses. This checklist can also serve as a guideline for authors conducting a review, allowing them to ensure their methodology is up-to-date and that their methodology and results are clearly reported to interested parties.

2.7. Previous Research

2.7.1. Previous diagnostic studies on syphilis POC tests

In the literature, the first reference to POC tests dates to 1996.(23) From the description of the tests provided, it can be confirmed that they follow the ASSURED criteria of the WHO (3, 5) and therefore can be correctly labelled as a POC test. Their accuracy is contested as these studies were conducted in what can be assumed to be well-controlled and sterile laboratories using serum samples with well-controlled levels of antibodies. Moreover, the studies were case-control, which is not ideal for evaluating diagnostic accuracy. However, given that these were the first tests in the field, it was logical to begin with this choice of study design as it is intended to obtain and optimize accuracy parameters. The optimistic estimates generated using this study design are

diluted downstream with the introduction of representative patient population samples, cross-sectional designs and variable reference standards. Earlier studies like Oshiro et al (22), Sano et al (25) and Qiajiao (23), and Zarakolu et al (36) indeed obtained very high accuracies, reporting sensitivity and specificities of 100%, with the exception of one experiment in Qiajiao (23) which recorded a specificity of 99.3%. In 2008, a Russian study by Rotanov et al (24) evaluated a popular POC test, Determine®, using serum, again in a case-control study. Using RPGA as reference standard, a sensitivity and specificity of 100% were once again obtained. In 2009, Van Dommelen (30) evaluated Biorapid Syphilis® in the Netherlands in a case-control design in serum from patients with known serostatus. The results were compared to TPPA, and a sensitivity and specificity of 92% and 79% were determined respectively. In 2010 study in China, Yang et al(35) evaluated two new tests they developed in serum, one based on quantum dots and one based on a colloidal gold-based lateral flow, and found a specificity of 100% for both tests and a sensitivity of 100% and 82% respectively.

In the year 2000, the first cross-sectional study evaluating a POC syphilis diagnostic test was published by Lien et al. (16) Unfortunately, while this study represented a step-up in study design from earlier articles, it was limited by the use of a reference standard which did not measure the same antibody as the index test: Lien et al used a non-TP-specific reference standard. To our knowledge the next study to be published was the first to correctly evaluate a syphilis POC diagnostic test. In 2002, West et al (34), evaluated the RST (Rapid Syphilis Test) in a cross-sectional design using the proper TP-

specific reference standard test. Following this breakthrough, the employed methodologies became more comprehensive.

In 2004, Siedner et al (76) evaluated Determine , Phoenix Biotech Trep-Strip IV(35) and Guardian One Step(35) in a city clinic in San Francisco, USA . Whole blood specimens were used and the results of index test were compared to a TPPA reference standard. All tests had a specificity of 100%. Phoenix Biotech had the lowest sensitivity while Determine had the highest sensitivity. Comparing these three tests, the authors recommended implementation of Determine.

In 2006, Juarez-Figueroa (14) evaluated Determine in female sex workers (FSWs) and pregnant women in Mexico. The study was conducted on serum specimens and the results were compared to a combination of VDRL and FTA-Abs reference standard. The results were read by three reviewers and agreement between reviewers was calculated to be 99.3%. In the same year, Hernandez-Trejo (13) also evaluated Determine in Mexico amongst pregnant women but instead, these authors used whole blood. The Determine test, compared to a combination of VDRL and FTA-Abs as reference standard, was found to have a sensitivity and specificity of 100%. The strength of the study is that the authors used whole blood but, unfortunately, they used an improper reference standard. In a Bolivian study in the same year, Tinajeros et al (28) evaluated the Determine test in pregnant women, using RPR and TPPA as a reference standard. Determine was found to be more accurate than the routine RPR performed and easier

to conduct. The authors explored the performance of POC tests using variations in specimens, patient populations, index tests and reference tests.

Subsequently, Campos et al (9) evaluated the accuracy of one popular test, Determine, in FSWs in Lima, Peru using a combination of reference standards (RPR and TPHA). This was the first study conducted in FSWs using whole blood. The authors conducted the test in whole blood and found that the sensitivity and specificity of the test was lower than previously reported in serum studies. The authors reported that incorporating this test into already established testing programs is very feasible. Subsequent studies followed in pregnant women, a population which benefit greatly from POC testing given the downstream effects on children.

In 2006, Montoya et al (19) evaluated Bioline in pregnant women in Mozambique. The authors compared results in both whole blood and serum. Additionally, the results for each specimen were compared to both a TP-specific reference standard (TPHA) and a combination of TP specific and non-TP specific reference standards (RPR and TPHA). This study was unique in that comparisons of specimens were undertaken, and the results were promising. Since then, studies which compare more than one test head-to-head and which evaluate at multiple sites have become more common. Some examples of such studies are Herring et al (27) and Mabey et al (17), both in 2006. Herring et al (27) conducted a study sponsored by the Sexually Transmitted Diseases Diagnostics Initiative (SDI) of the WHO/TDR. The study compared nine POC tests in different geographical

locations using archived sera. The authors found that the tests had a range of sensitivity of 84.5% to 97.7% and a range of specificity of 84.5% to 98%. The WHO conducted a quality assessment of each laboratory in order to ensure that the results from each site were directly comparable. The tests were also rated on their test characteristics. The tests which were found to be user-friendly and which were compatible with whole blood samples were further evaluated by Mabey et al (17) in another study sponsored by the SDI of the WHO/TDR. Four tests were selected and evaluated in four geographical locations. The rationale of selecting four tests was not known. Tests were conducted in clinic attendees using whole blood in the clinic as well a whole blood and s serum in the laboratory. The range of sensitivities found was to be 64% to 100% and the specificities were greater than 95% in all cases. These studies both touched on the limitation of using POC tests with whole blood as the sensitivity parameter was lower and more variable when using whole blood rather than serum.

In 2007, Wang et al (32) evaluated Determine, Standard Bioline, Qualpro Syphicheck® and Omega Visitect® in whole blood as well as serum. Using TPHA as the reference standard, sensitivity and specificity were higher in serum than in whole blood. The Determine test had the best performance. This was most likely the first study to establish Determine's performance.

In 2007 and 2008, Benzaken et al (6, 7) published two studies evaluating POC tests. The first study conducted in 2007 (6) evaluated SD Bioline and Qualpro Syphicheck,

Determine and Visitect. The study was conducted in STD clinic attendees in the red light district of Manaus, Brazil using both serum and whole blood. The tests were directly compared in groups of two: SD Bioline and Syphicheck were compared head-to-head and Determine and Visitect were compared head-to-head. The authors used an appropriate TP -specific reference standard, FTA-ABS. It was unclear whether the results were read in a blinded fashion. In addition, the authors were able to collect data on the acceptability of these POC tests by clinicians and patients.

The study conducted in 2008(7) was also conducted in STD clinic attendees in the red light district of Manaus, Brazil but this time only Visitect was used. The study investigated the sensitivity and specificity of the test in detecting syphilis as well as the performance of the test in detecting active syphilis cases. The results were read in a blinded fashion, in order to avoid reviewer bias. Importantly, the authors, in line with their 2007 study, continued their research in outcomes other than accuracy, and were able to define preference for venous blood and pain from finger prick as barriers. In the context of the Brazilian healthcare system, the results of both studies are promising because the specificity of all four tests was similar to VDRL, the routine test currently used. The sensitivity of all four tests was actually higher than VDRL.

In 2007, Bronzan et al (8) evaluated Determine in antenatal clinic (ANC) attendees in Cape Town, South Africa using whole blood. Unfortunately, the authors used RPR and TPHA as their reference standard. The authors conducted an interim analysis of the data

and found that the accuracy results were systematically more variable for patients with low syphilis titres. After investigating this, they found this was because not enough blood was drawn and hence the authors retrained the clinicians conducting the test. This level of quality control is praiseworthy and is highly recommended to future researchers.

In 2008, Nessa et al (20) evaluated ICS-ACON® and RTD-ACON® in FSWs in Bangladesh. The results of the index tests were compared to a combination of reference standards (RPR and TPHA). In 2009, Villazaon-Vargas(31) evaluated Determine in Bolivian pregnant women using whole blood. The results were compared to FTA-Abs and a sensitivity of 98% and a specificity of 99.8% were found. The authors also found that the POC results achieved using Determine were in strong contrast to the 60% of patients who historically received their results when administered the conventional test. This comparison of implementation outcomes to historical controls was a particular strength of this study. It is unclear whether the readers of the index test and reference standard were blinded to the results of the other test.

In 2009, Li et al (15) evaluated Determine, Visitect, SD Bioline and Syphicheck in STD clinic attendees in China. The tests were done on both whole blood and serum and compared to a TP-specific reference standard (TPHA). Determine and Visitect were compared head-to-head, as were SD Bioline and Syphicheck. In all tests except

Syphichek, sensitivity was lower in whole blood than in serum. In all cases, specificity was more than 95%. All tests were read independently by two reviewers.

In 2009, Nyamwamu et al (21) evaluated Accurate Ultra Rapid® in serum samples in a Kenyan ANC attendee population. The POC test was compared to VDRL and TPHA reference standards separately. The authors took care to investigate the cause of disagreement between the index test and reference standard. They found that often the reason for disagreement was that the individual had a previous case of treated syphilis. Additionally, the authors identified that 56% of pregnant women tested were already in their third trimester. Unfortunately, by this point there is a high probability that the disease has already been transmitted to the fetus. Lamentably, if screening is not provided early in the pregnancy, even treatment at the same visit is not useful.

Mishra et al(18) evaluated Syphicheck in Bangalore, India, in 2009. The test was done in whole blood samples obtained from FSWs. The results were compared to a combination of reference standards (RPR and TPHA) and a sensitivity and specificity of 70.8% and 97.8% were found respectively. However, again, these results are not reliable as the reference standard used was not the gold standard. Use of POC testing resulted in an increase in the number of individuals who received treatment. The authors were limited by the number of participants and hence no subgroup analysis on stage of disease could be conducted. The study was conducted in four clinic groups. Sensitivity among these

four groups was found to have high variability. Further investigation into why this may have occurred would have improved on our understanding of the POC test.

In 2010, the first study evaluating a POC test with both a TP and non-TP specific component was published by Castro et al (10) who developed a syphilis POC test which contains both TP specific and non-TP specific components. Unfortunately, similar to the earlier studies evaluating TP-specific syphilis POC tests, the study was conducted in a case-control manner in order to optimize test performance. The test was conducted in serum samples from Georgia Public Health Laboratory. The results from earlier studies were not, however, echoed in Castro et al, for even though they obtained relatively high sensitivity and specificity, they were not 100%: Castro et al reported a sensitivity of 96.5% and specificity of 97.7% in one experiment and a sensitivity of 97.4% and specificity of 99.1% in another experiment. This difference could potentially be linked to the non-TP component of the test.

In conclusion, the majority of studies were conducted in ANC and STD clinic attendees. There were up to 17 kits investigated using both whole blood and serum specimens. Unfortunately, there was great variability in the reference standard used, despite the fact the basic property of each POC test remains the same. On this point, it should be noted that even though a meta-analysis of the available data is highly relevant, efforts are greatly hampered by this reference standard variability.

Papers which evaluated implementation outcomes can also be found throughout, starting in 1996 with Qiajio.(23) Similar to diagnostic accuracy studies, the rigour of the studies improved over time. A list of these publications is found in Table 8.

2.7.2. Previous meta-analysis and systematic review

Last year, the first systematic review in this rapidly evolving field was published by Tucker et al.(29) The objectives of this study, which focused exclusively on accuracy of syphilis POC tests in STD users and pregnant women, were to assess the characteristics of ICS tests, and conduct sub-group analyses of these characteristics by population, specimen, non-treponemal syphilis titre, HIV co-infection and test manufacturer.

The statistically pooled median ICS sensitivity was reported to be 0.86 (IQR 0.75-0.94) and the median specificity was reported to be 0.99 (IQR 0.98-0.99). The median sensitivity in ANC attendees was 0.86 (IQR 0.75-0.91) and the specificity was 0.99 (IQR 0.97-1.00). The median sensitivity in STD clinic attendees was 0.86 (IQR 0.73-0.94) and the specificity was 0.99 (IQR 0.98-0.99).

This systematic review has strengths worth mentioning. The authors followed the PRISMA guidelines which allowed their review to be more transparent to the reader. The assessment of papers included in the systematic review and data extraction was conducted by two parties in a blinded fashion. Papers with partial verification bias were excluded. The authors also conducted a χ^2 test for heterogeneity of sensitivity and specificity, although these tests are not usually thought to be worthwhile because of

power issues and the poor decision making which can follow decisions based on p-values. Publication bias was explored using forest plots, Begg's test (tests for rank correlation) and Egger's test (test for regression asymmetry).

There are, however, limitations with this study. This review did not compare current tests head-to-head and did not evaluate implementation outcomes. The authors assumed the reference standard used was perfect. They did, however, account for the type of reference standard used and attempted to adjust for this using a Wilcoxon rank-sum test. This systematic review took a frequentist approach and hence there are limitations with the methods they chose. Tests of heterogeneity are not very practical as they test for the presence of perfect homogeneity when the probability that this is in fact true is very small. Hence, having a result which states that the sensitivity and specificity are heterogeneous does not add more information. The methods used to explore publication bias are also inappropriate, as these methods are generally reserved for meta-analyses of RCTs and observational studies. To elaborate, these methods are not applicable to meta-analyses of diagnostic studies as there is no cut-off value of significance in diagnostic studies (i.e., p-values) which separates studies into those deemed significant versus those deemed insignificant.(68) As this point, there are still no methods developed to assess publication bias in the context of diagnostic studies. Finally, Tucker et al(29) only evaluated the diagnostic accuracy of the tests and did not further explore the literature to evaluate the IROs which are associated with the implementation of syphilis POC interventions.

This thesis attempts to provide an improvement over the work by Tucker et al(29). A comprehensive search of six databases was done and all languages were included. Each included study was also critiqued on quality. Tucker et al(29) looked at the performance of ICS within ANC and STD clinic groups, but since Tucker et al(29) found no difference in ANC and STD clinic attendees in sensitivity and specificity, this author decided to explore different stratification strategies. We evaluated each kit separately, classified by the type of sample used, as well as by the type of reference standard used. We used a more sophisticated analytical model compared to Tucker et al(29) to evaluate the role of using TP, non-TP and combinations reference standards. Most importantly, we used a random effects Bayesian hierarchical model in comparison to Tucker et al(29) who used a fixed effects binomial regression model. We also evaluated the role of accuracy parameter variability within reference standards on the overall accuracy after adjusting for the fact that reference standards do not identify the correct disease status in 100% of instances. In addition, we analysed articles reporting outcomes relevant to the implementation of syphilis POC test, IROs, in order to provide a complete picture of the relevance of syphilis POC tests given various contexts. As a result, this thesis is more useful a meta-analysis for it allows for more specific, and hence practical, conclusions to be drawn, conclusions which can benefit clinicians, community leaders, and policy makers.

Chapter 3: METHODS

The aim of this chapter is to explain the methodology employed in completing this thesis. This chapter will cover the literature search process, the acquisition of appropriate papers, the data extraction process and the statistical analysis.

3.1. Protocol

We conducted this systematic review in several phases, and assigned primary, secondary and tertiary reviewers (Yalda Jafari (YJ), Sushmita Shivkumar (SS), and Nitika Pant Pai (NPP)). YJ conducted the literature search, obtained the final subset of studies, extracted the data from all studies, conducted statistical analysis and analyzed results. SS also independently conducted the literature search and arrived at final subset of papers, and extracted data from 34% of studies. NPP was consulted when consensus could not be reached between YJ and SS when comparing the included studies and extracted data. The author followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) check list for conducting and reporting the review (71). Critical appraisal for quality was undertaken using STAndards for the Reporting of Diagnostic accuracy studies (STARD) (62) and Quality Assessment of Diagnostic Accuracy Studies (QUADAS).(66)

3.2. Eligibility Criteria

Types of studies: Studies conducted within the last three decades (1980-2010) in any language were considered. The study must have been done on humans or human samples. No restriction on study type was imposed, but the type was recorded.

Abstracts, letters and brief reports, editorials, perspectives, opinion pieces and manufacturer reports were excluded from the meta-analysis.

Types of participants: Participants of any age or risk group were considered. Studies on live participants using whole blood as well as studies on serum specimens either from participants or serum panels were considered.

Types of interventions: Studies documenting the use of POC tests were used. If the study referred to the test as POC or rapid but upon further reading, it was found that the device did not follow the ASSURED criteria of the WHO (3, 5), the study was not included. The intervention must have been completed in parallel with a reference standard test to be eligible for a diagnostic accuracy study. If the reference standard was not utilized to verify outcomes of POC tests in all samples, a situation which results in partial verification bias (68, 77), then the study was excluded.

3.3. Types of outcome measures

We divided our outcomes into two parts.

Part I: Outcome measures were defined as sensitivity and specificity of the index test as measured against a reference standard. The raw cell values, true positive, false positive, true negative, and false negative cell counts must have been either available in the study or have been calculable from the sensitivity or specificity values and sample sizes provided in order to be included in the final meta-analysis.

Part II: Outcomes measures were defined as IROs and included acceptability, preference, feasibility, impact, prevalence, barriers and challenges and economic evaluations. In order for an article to be included under a category, the article had to mention a relevant key word under results.

3.4. Information Sources

Our search strategy covered the time period from 1 January 1980 to 24 September 2010. We chose this time period in order to ensure that no POC tests were missed. We searched the following electronic databases: MEDLINE, EMBASE, GLOBAL HEALTH, CINAHL, Web of Science and SCOPUS. All languages were considered and non-English articles were translated using translators. We also screened bibliographies of primary studies and review articles for additional articles of potential interest.

3.5. Literature Search

Lorie Kloda, a librarian at the McGill University Life Sciences library was consulted regarding the appropriate search strategy. With her help, a preliminary search was conducted in PubMed using indexed MESH headings for syphilis, *Treponema pallidum*, sensitivity and specificity as well as a filter for diagnostic studies. Due to the limited number of articles following the preliminary search, the final search string chosen was very broad. We searched using the following terms: (syphilis OR *Treponema pallidum*) AND (point-of-Care OR rapid test OR rapid assays). Keywords rather than MESH terms were used in order to capture more recent papers not yet indexed and in order to

permit searching multiple databases at once. No filters for diagnostic studies were used as these have been shown to remove possibly relevant studies (77).

3.6. Study Selection

The search string was run and eligibility assessment was performed independently by two reviewers, YJ and SS. The reviewers followed PRISMA guidelines for recording the papers screened, deleted and included. After a reaching a final set, the reviewers convened and compared the papers they found. Disagreements were discussed between the reviewers and were resolved by consensus. If consensus regarding the inclusion of a paper could not be reached, a third reviewer, NPP, was consulted.

3.7. Data collection process

The data collection process was divided into two parts:

Part I: We developed an extraction sheet, Table 2, that was pilot tested on five studies and was refined accordingly. After the final set of studies was collected and agreed upon by all reviewers, one reviewer, YJ, extracted data on all papers and the second reviewer, SS, extracted data on 34% of randomly selected papers. The abstraction was done independently and blindly. The two reviewers came together and compared entries and disagreements were resolved by consensus. If consensus could not be reached, a third reviewer, NPP, was consulted.

Part II: An extraction sheet specific to IROs, Table 2, was test piloted on two studies and was refined accordingly. After the final set of studies were collected, one reviewer, YJ, extracted data on all papers while a second reviewer, SS, extracted data on 25% of randomly selected papers. The two reviewers compared entries and disagreements were resolved by consensus. If consensus could not be reached, a third reviewer, NPP, was consulted. If papers contained missing information, the authors of the work were contacted.

3.8. Data Items

Information from each study was extracted, including: (1) setting (including type of hospital or clinic, city, country, and World Bank economic ranking); (2) characteristics of study participants (including whether they belonged to a high-risk group or had other underlying diseases or conditions); (3) inclusion and exclusion criteria; (4) index test (including name of test, name of company, and type of technology the test used (e.g., immuno-chromatographic or flow-through)); (5) reference test (including name of test, name of company, and whether the test detects TP or non-TP specific antibodies or both); and, (6) Outcome measures (including cell counts from the two by two table leading to estimates of sensitivity and specificity, acceptance, preference, feasibility, prevalence, impact, barriers and challenges and economic evaluations).

3.9. Risk of Bias in Individual Studies

The methodological quality of studies was assessed using QUADAS. QUADAS is a validated quality checklist composed of 14 items which encompass the most import

sources of bias and variation observed in diagnostic accuracy studies. The two reviewers used the QUADAS checklist and scored each item as “Yes”, “No” or “Unclear,” as per the recommendations of the QUADAS checklist authors.(66, 73)

The quality of reporting was evaluated using the STARD criteria. STARD consists of a list of 25 items which assesses the completeness of reporting in diagnostic studies, potential sources of bias and generalizability. The two reviewers used the STARD checklist, scoring each item as “Reported” or “Not Reported” as per the recommendations of the authors of the checklist.(62)

3.10. Summary Measures

Summary measures are divided into two parts:

Part I: Pooled sensitivity and specificity were the primary measures of accuracy for POC tests.

Part II: Acceptability, preference, feasibility, impact, prevalence, challenges and barriers and economic evaluations were summarized and tabulated. However, based on the previous experience of the research team, it was highly anticipated that a meta-analysis of the results would not be possible due to inconsistent definitions of the terms and methods of calculation across studies. Therefore, a narrative review approach was adopted.

3.11. Definition of Summary Measures

Sensitivity was defined as the proportion of those with positive disease status who test positive with the index test.(60)

Specificity was defined as the proportion of those with negative status who test negative with index test. (60)

Acceptability was defined as the proportion of patients and providers who accept the POC test and the associated testing strategy when compared to conventional tests and other testing methods.

Preference was defined as the proportion of patients, doctors, nurses and lab technicians who prefer the test and testing strategy compared to conventional tests and other testing methods.

Feasibility was defined as how feasible it is to conduct the POC test as well as to conduct the test strategy. Feasibility of study conduct was captured by recruitment rate and feasibility of study procedure was captured by consent rate and completion rate.

Prevalence was defined as the proportion of individuals who tested positive when tested with POC test. Studies that solely used POC tests to determine prevalence of syphilis were included in this category.

Impact refers to outcomes such as: proportional increase in cases with new strategy versus old strategy, proportional reduction in infant transmission, reduction in infant morbidity and mortality due to congenital syphilis, proportional reduction in time to confirmatory testing and treatment and referral linkages compared to conventional strategy, proportional increases in treatment initiation compared to conventional strategy, and increase in partner notification with the POC strategy compared to conventional strategy.

3.12. Planned Methods of Analysis

Part I: We did not conduct a formal test of heterogeneity. The null is that the studies are homogenous. Since this null is highly improbable regardless of data collected, we will not gain any knowledge by conducting a heterogeneity test. It is highly implausible that all studies will have exactly the same outcomes regardless of setting. Therefore, we employed a random effects model throughout.

Subgroups were created based on the index test, sample type (serum or whole blood) and reference standards used (TP, non-TP specific, or TP and non-TP specific). Studies were grouped based on index test in order to make the available information most useful. Subgroups were created based on sample type (serum or whole blood) because biologically it is known that results in serum are found to have higher sensitivity and specificity. The reference tests used, whether TP or non-TP specific antibody or both, is also important because the results indicate different biological status and this affects

the accuracy results of the index test. Therefore, it is important to keep a distinction between types of reference tests. This is summarized in the figure below.

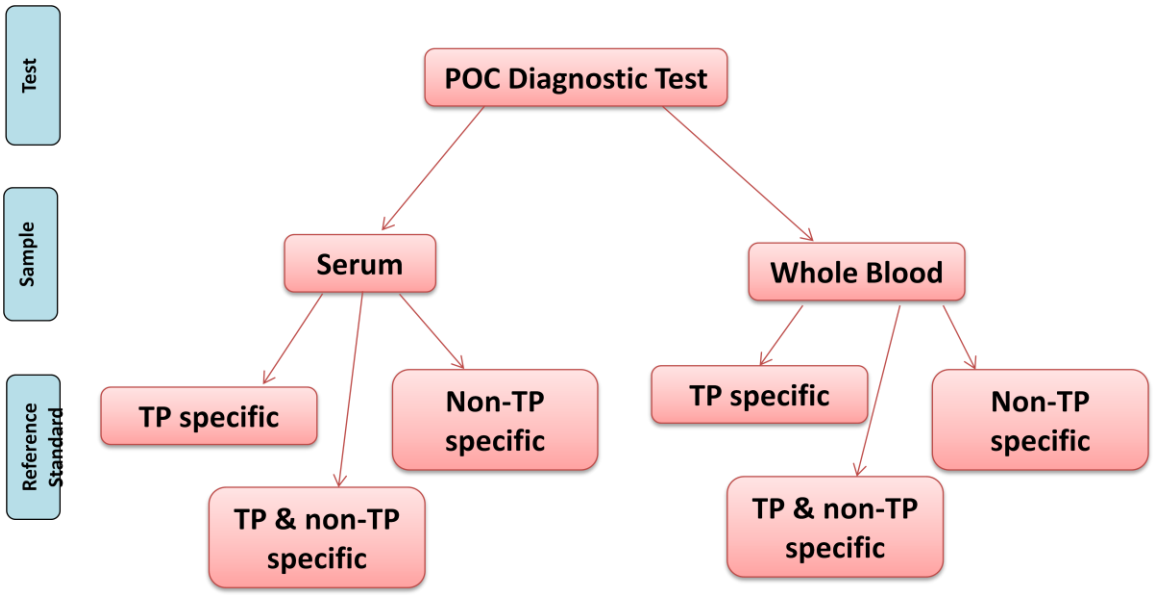

Figure 4: Stratification strategy employed in order to make subgroups fit for pooling.

Each stratum formed needed at least four entries in order to be statistically pooled. Forest plots for each strata displaying the sensitivity and specificity of each study were created using Meta-Disc software (version 1.5). We also conducted a Hierarchical Summary Receiver Operating Characteristic Curve (HSROC) analysis using an add-on package(version 0.9.0, 2011, Montreal) to R software (version 2.12.1, 2010, The R Foundation for Statistical Computing) developed by researchers at McGill University and reported pooled sensitivities and specificities.(78) The HSROC model is based on the assumption that each study has its own individual positivity threshold, represented by an accuracy parameter which is independent between studies. In each study, the

positive results of the index test follow a binomial distribution, with parameters for positivity threshold (θ_i) and accuracy (α_i), as represented in Equation 1, each of which follows a hierarchical normal distribution amongst studies, as represented by Equation 2. (70)

Equation 1: Within study variation (70)

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i X_{ij})e^{-\beta X_{ij}}$$

Equation 2: Between study variation (70)

$$\left.\begin{aligned} \theta_i|\Theta,\gamma,Z_i,\sigma_\theta^2 &\sim \text{N}(\Theta+\gamma Z_i,\sigma_\theta^2) \\ \alpha_i|\Lambda,\lambda,Z_i,\sigma_\alpha^2 &\sim \text{N}(\Lambda+\lambda Z_i,\sigma_\alpha^2) \end{aligned}\right\} \text{conditionally independent}$$

Adopting this method, analysis using gold reference standard was carried out. The HSROC package also allows for adjustment of imperfect reference standards, for in reality the reference standards used to confirm the results of an index test are not 100% sensitive and specific. This adjustment was made using the HSROC package and the results were compared to results generated using the gold standard to see the effect of improper reference standards. To do this, we used prior distributions that were extracted from the literature. A paper by Peeling et al (3) reviewed the available testing

protocols for syphilis and provided estimates of sensitivity and specificity for each test. TPHA and TPPA are estimated to have a sensitivity of 85% to 100% and specificity of 98% to 100%. RPGA, the reference standard used by Rotanov et al (24), was not discussed by Peeling et al (3), but from the description of the test, it was classified as being similar to TPHA and TPPA, and thus the same accuracy parameters were used. FTA-ABS is estimated to have a sensitivity of 70% to 100% and a specificity of 94% to 100%. For combination TP and non-TP tests, an estimated range of 90% to 100% for both sensitivity and specificity were used because no available estimate of this combination testing was found. This was considered as the most plausible range.

Chapter 4: RESULTS

This chapter reports the results of the meta-analysis and systematic review in two separate parts. Part I explains the results of the meta-analysis of diagnostic accuracy for POC syphilis tests, while Part II explains the results of the systematic review of IROs of these tests.

4.1. Part I

4.1.1. Study Selection

A total of 30 articles evaluating the diagnostic accuracy of POC tests for syphilis were identified for inclusion into the meta-analysis. A list of included studies can be found in Table 3. Our search string was run in MEDLINE, EMBASE, GLOBAL HEALTH, CINAHL, Web of Science and SCOPUS, with results limited to the years 1 January 1980 to 24 September 2010.Through snow ball searching, which refers to searching through the listed reference of included studies, one additional article was found: Sano et al(25) was found in the reference list of Oshiro et al.(22)

After removing duplicates, the remaining 647 articles were screened using titles and abstracts. Sixty-four full text articles were assessed for eligibility. Reasons for exclusions were: article was not actually evaluating a POC test (n=9), the study was either a bulletin report, review, a letter or an abstract (n=7), and one article, Herring et al(27), despite meeting all the inclusion criteria, had to be excluded because there were no raw cell values and they could not be calculated based on the information provided, and no response was received from the authors when contacted. An article by Wang et al(33)

was excluded because the information was already captured in the paper by Mabey et al(17) as a part of a multicentre evaluation of syphilis POC tests conducted by the Sexually Transmitted Diseases Diagnostic Initiative(STDDI) of WHO. It must be noted that the raw cell values of the paper by Mabey et al (17) were not provided and thus had to be calculated from the sensitivity and specificity estimates in the paper. We were not able calculate the values at one instant and thus, the information from Wang et al(33)was used to fill in the missing information. Overall, five authors were contacted and three responded. Figure 6 depicts the study selection process. Some articles evaluated more than one test, in more than one sample or more than one population, hence, the overall number of data point entries was 127.

4.1.2. Study Characteristics:

The complete list of study characteristics can be found in Table 3.

4.1.3. Participants and samples

Sixteen of 30, or 53% of studies, used whole blood, for a total of 55% (70/127) of data entries. In the whole blood group, 44% (7/16) tested STI clinic attendees, 19% (3/16) tested female sex workers (FSW), 38% (6/16) tested pregnant women and 6% (1/16) used blood supplied by hospital. No studies were conducted on children.

4.1.4. Intervention

Through this review, 17 unique POC tests were identified. Only one test was a filtration assay, the rest were immune-chromatographic strips (ICS). The most commonly used

tests were Determine (Abbott Diagnostics, UK) at 29% (37/127), Bioline Syphilis (Standard, South Korea) at 18% (23/127), Syphicheck test (Qualpro, India) at 15% (19/127) and Visitect test (Omega Diagnostics, UK) at 14% (18/127). One study did not mention the name of the test or the manufacturing company and simply referred to the device as an ICS.

4.1.5. Outcomes

In all studies, the primary outcomes, sensitivity and specificity, and raw cell counts for the two by two tables were tabulated. The raw cell values were derived from comparing the results of the POC test with a reference standard. The reference standards were TP specific, non-TP specific or a combination of both TP and non-TP specific. Two percent (3/127) of studies used non-TP tests, 79% (100/127) used TP specific reference standard and 18% (23/127) used a combination reference standard.

4.1.6. Risk of Bias within Studies

The quality of methodology and reporting of studies was assessed using QUADAS and STARD checklists.

4.1.6.1 QUADAS

Results from the QUADAS evaluation of included articles are summarized in Table 4.

All articles had absence of incorporation bias (100%), adequate description of index test execution (100%), and adequate description of reference test execution (100%). An overwhelming majority of articles had an adequate reference standard (97%), absence

of disease progression bias (97%), absence of partial verification (93%), absence of differential verification bias (90%), and clear description of selection criteria (87%). Six quality items were addressed in 60% of papers or less with following breakdown: adequate spectrum composition (60%), absence of clinical review bias (43%), report of uninterpretable results (33%), absence of index test review bias (37%), absence of reference test review bias (20%), and description of withdrawals (17%).

A limiting factor in assessing methodological quality is unclear reporting. Absence of reference and index test review bias was unclear in 80% and 63% of articles respectively. Absence of clinical review bias was unclear in 57% of articles.

4.1.6.2. STARD

Results from the STARD evaluation of included articles are summarized in Table 5.

Four of the 25 items were reported by all the articles. Three of these items (items 7-9) were under methods: reference standard used and its rational (100%), technical specifications of materials and methods (100%), and definition of and rationale for units. The fourth item was item 25 under discussion which details whether the article discussed the clinical applicability of findings (100%).

Items 5, 6, 11, 13, 17,20, 22, 23, and 24 were reported by 50% of articles or less, with Item 20 (report of any adverse effects from use of index test or reference standard) was the least reported item (3%).

4.1.6.3. Other

Only 37% (11/30) of papers reported on conflicts of interest, with one study, Castro et al(10), indicating conflict of interest. In three papers, Castro et al(10), Yang e al(35), Zarakolu et al(36), authors evaluated diagnostic tests they had manufactured themselves, which may introduce bias as the evaluation was not done independently. The relevant information is compiled in Table 6.

4.1.7. Synthesis of Results

Data were pooled using an HSROC model, assuming perfect and imperfect reference standard.

4.1.7.1 Meta-analysis assuming perfect reference standard

In order to minimize heterogeneity, the data were stratified as previously described in the methods. Stratification was used as a tool to minimize variability of sensitivity and specificity by combining data with similar characteristics in the following order of importance: POC test used, sample used and reference standard used. In order for the HSROC models to run, there needed to be a minimum of four studies per stratum. The groups with sufficient data to be meta-analyzed are displayed as forest plots (Figures 7-25) as well as in Table 7, which presents the results of our meta-analysis.

Using a TP specific reference standard, the Determine test in serum had a sensitivity of 98.81% (95% confidence interval: 96.52, 99.98) and a specificity of 97.94% (96.30, 98.48). In comparison, using a TP specific reference standard, the Determine test

in whole blood had a sensitivity of 85.55% (76.15, 94.49), and a specificity 99.50% (98.95, 99.93).

Using a TP specific reference standard, Bioline in serum had a sensitivity of 93.99% (88.79, 98.45) and a specificity of 98.58% (97.54, 99.31), and in whole blood had a sensitivity of 87.70% (84.78, 90.58) and a specificity of 99.07% (98.50, 99.59).

Using a TP specific reference standard, Syphi-check in serum had a sensitivity of 81.91% (69.10, 93.19) and a specificity of 99.05% (98.35, 99.62), and in whole blood had a sensitivity of 76.78% (69.36, 84.03) and a specificity of 99.44% (99.06, 99.76).

Using a TP specific reference standard, Visitect in serum had a sensitivity of 93.96% (89.24, 98.01) and specificity of 98.51% (97.70, 99.19), and in whole blood had a sensitivity of 78.05% (70.34, 85.02) and specificity of 99.52% (99.16, 99.82).

4.1.7.2 Meta-analysis adjusted for imperfect reference standards

Results from each stratum were adjusted for imperfect reference standards used using HSROC modeling. The results are found in Table 9.

Using a TP specific reference standard, the Determine test in serum had a sensitivity of 99.17% (96.56, 100) and a specificity of 99.28% (98.15, 100), while in whole blood it had a sensitivity of 89.49% (79.88, 98.15), and a specificity 99.91% (99.44, 100).

Using a TP specific reference standard, Bioline in serum had a sensitivity of 99.67% (97.65, 100) and a specificity of 99.56% (98.55, 100), and in whole blood had a sensitivity of 91.47% (87.06, 96.12) and a specificity of 99.61% (99.04, 100).

Using a TP specific reference standard, Syphicheck in serum had a sensitivity of 88.46% (73.54, 99.87) and specificity 99.98% (99.64, 100), and in whole blood had a sensitivity of 81.99% (71.84, 91.99) and a specificity of 99.81% (99.46, 100).

Using a TP specific reference standard, Visitect in serum had a sensitivity of 98.18% (93.53, 100) and a specificity of 99.89% (99.19, 100), and in whole blood had a sensitivity of 82.93% (94.50, 100) and a specificity of 99.87% (99.58, 100).

Forest plots for all studies, as well as forest plots for each stratum that could be meta-analyzed are provided. In all instances, sensitivity is more variable than specificity.

4.2. Part II

From 64 full text articles assed, 25 (39%) articles were identified which addressed issues beyond the diagnostic accuracy of POC tests. Each study was categorized depending on the IRO investigated. Keywords mentioned in the articles were used to accomplish this as explained in the methods. Three articles reported on acceptability, 4 on preference, 8 on feasibility, 7 on impact, 7 on prevalence and 7 on barriers and challenges and 7 on economic evaluations. Figure 5 depicts the classification of studies. Sixteen percent

(n=4) were in a language other than English: 2 articles were in Spanish, 1 article was in Portuguese and 1 article was in Russian. Thirteen percent (n=3) of studies were fully or partially conducted in a high income country. In the group of articles evaluating outcomes other than economic evaluations, 19, the majority of the studies, 89% (n=17), employed a cross-sectional methodology, while 10% (n=2) used a case-control design, and 5.3% (n=1) used a clustered randomized trial. Table 8 describes the characteristics of studies included in part II, excluding economic evaluation studies.

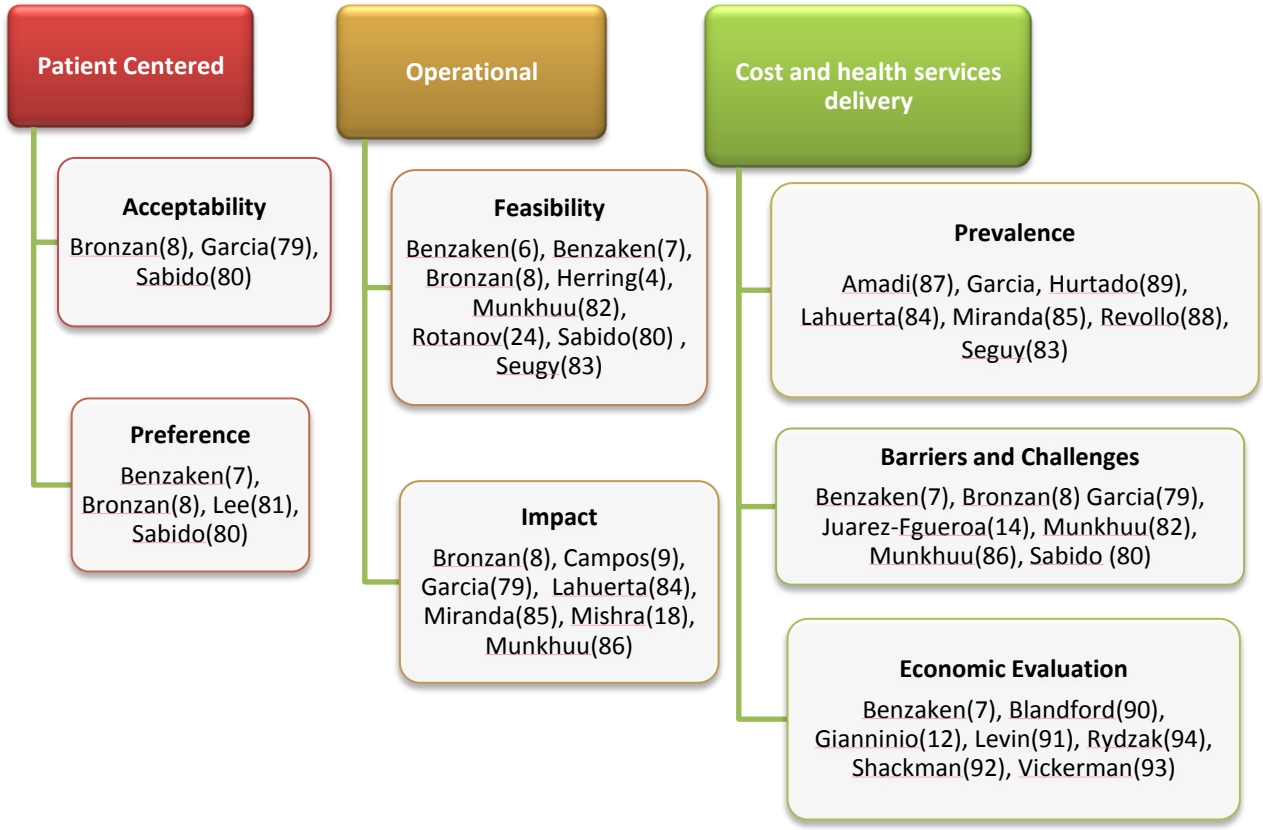

Figure 5 : Categorization of studies included in Part II

4.2.1. Acceptability

Through a questionnaire, Garcia et al (79) were able to conclude that the Determine test was highly acceptable by participants, clinicians, and laboratory technicians. Ninety-five percent of patients in the study by Sabido et al (80) would recommend Visitect syphilis test to their friends. In a study by Bronzan et al (8), onsite ICS testing strategy was highly acceptable to 100% of the nurses. Overall, POC tests were acceptable.

4.2.2. Preference

Sixty percent of clinicians in Sabido et al (80) preferred conventional testing over the Visitect syphilis test because they noted that conventional tests provide better information for staging and treatment. Forty-eight percent of patients in the same study preferred the Visitect syphilis test over conventional testing, citing reasons such as knowing their status quickly and having a fear of needles. Sixty eight percent of nurses in Bronzan et al (8) preferred the onsite ICS test over offsite testing because onsite testing allowed for rapid diagnosis and the possibility of treatment at the same visit. In a study by Benzaken et al (7), Visitect syphilis was evaluated in Brazil where 62% of participants preferred the rapid test.

The study by Lee et al (81) was the only study which investigated the effect of self-testing using a POC test by the patient in the privacy of their home. In this study, the Determine Abbott test was used to detect syphilis in MSM attending gay community events. The conduct of the test was explained to participants but a clinician conducted the test. Having seen how the test works, the participants were asked in a questionnaire

if they would prefer conducting the test at home by themselves versus having the test conducted by a clinician. While 21% were unsure whether they would home test if the option was available, a majority, 54%, stated that they would conduct the test at home by themselves. Nineteen percent of those who said they would self-test mentioned convenience and ease of testing while 4% mentioned confidentiality and privacy as reasons for choosing the self-testing option.

4.2.3. Feasibility

A qualitative study was conducted by Munkhuu et al (82) to evaluate the feasibility of a one-stop anetenatal service providing POC testing for syphilis using Bioline syphilis 3.0. Patients and providers were asked their opinions and their level of satisfaction with the service provided. All patients were satisfied with the rapid test. Some of the cited reasons for the high level of satisfaction were savings on time and expenses related to transportation (87.8%) and painless testing and rapidly available results (76.7%). Also, all providers supported the rapid test to prevent an additional journey and mentioned that the testing procedure was straightforward and easy to conduct. Overall, the authors found the one-stop antenatal service to be feasible in rural Mongolia.

In an article by Seguy et al (83), uptake rate was used as a proxy to measure feasibility of testing. The Determine test was offered to miners and up to 80% agreed to be tested. This high rate shows the feasibility of POC testing offered to Guayanese gold and diamond miners.

Rotanov et al (24) evaluated 10 test kits based on characteristics such as conciseness of instructions provide and ease of use. Under the category “ease of interpretation” and “simplicity,” highest scores were given to the Determine, Syphicheck-WB, Bioline Syphilis and Trepeonema-Express, while the lowest score was given to Smart Strip syphilis. In the study by Sabido et al (80), Visitect syphilis was offered to patients at an STD clinic in the red light district of Brazil. 75% of staff found the test instructions to be easy or very easy. Twenty perfect of clinicians found the interpretation of test results was not very easy or even complex because the line was blurry. Also, negative results became positive if left for more than 15 minutes. Sixty-nine percent of participants found that the timing for the test was short: the average time spent at the clinic was 51 minutes with a standard deviation of 32 minutes. Forty-eight percent reported no discomfort as a result of the rapid test.

In a study by Benzaken et al (6), 4 tests, Bioline, Syphicheck, Visitect and Determine, were evaluated. One hundred percent of patients were willing to wait up 30 minutes and 59.1% up to an hour. Health professionals rated instruction comprehension, and result interpretation as 100% easy or very easy for all tests. Results were obtained in under 15 minutes for 100% of samples tested with Bioline, 75% tested with Syphicheck-WB, 89% tested with Visitect syphilis, and 78% tested with Determine.

In Bronzan et al (8), 100% of nurses found the onsite ICS test fast, reliable and easy to perform. Almost all, 95%, also found the onsite RPR test time consuming, unreliable and

difficult to perform. Similarly, in Benzaken et al (7), 75% of clinic staff found the POC test easy to use and 67% found it easy to interpret. However, pain by finger prick in 57% of the participants decreased feasibility.

Herring et al (4) conducted a multi-site evaluation of multiple POC tests. Case-control evaluations were done in South Africa, Gambia, Tanzania, China, Sri Lanka, Haiti, USA and Russia. Six kits (Determine, Syphilis Fast, Espline TP, Syphicheck-WB, SD Bioline and Visitect) were evaluated on the instructions of the kit, ease of use and interpretation. The tests were scored on each category and the results added. Determine had the highest score while Syphilis Fast had the lowest score.

4.2.4. Impact

Garcia et al (79) reported that 93.2% of individuals categorized as positive for TP antibodies using the Determine test received a single dose of penicillin at the same visit. Eighty one and a half percent received the 3 recommended doses of penicillin, indicative of a high follow-up rate. Eighty-six point three percent of positive individuals identified their partners and as a results, 76.9% of identified partners showed up for treatment.

In the study by Lahuerta et al (84), the Determine test was offered in mobile van clinics as well as traditional clinics. Comparing the two settings, a vast majority of MSM and transgender (TG) individuals who showed up to be tested showed up at the mobile van clinics. Treatment was offered to all found positive with the POC test.

In the study by Miranda et al (85), the Determine test was offered to women attending antenatal clinic in Brazil. Five point one percent of the patients had not received prenatal care previously and thus had no information about their status. Among these, one case of syphilis was found using the POC test.

In a study by Bronzan et al (8), an onsite ICS was compared to onsite RPR and offsite RPR. Onsite ICS resulted in the greatest proportion of women who were correctly diagnosed and subsequently treated. Eighty-nine percent of patients with high titer received treatment versus 60% of women who received treatment using offsite testing method.

In the study by Campos et al (9), the Determine test was used to test FSWs in Peru. Eighty-seven percent of those who tested positive with Determine visited the local health center for treatment and 64% completed the three-dose treatment regime.

Mishra et al(18) evaluated the Qualpro Syphicheck-WB among FSWs in India. There was a substantial difference in the proportions of infected individuals who received treatment using POC, 68%, versus the historical control, 45%. Time from testing to treatment also decreased from 11 days (range 0-317) using conventional testing versus same day treatment in POC strategy.

The cluster randomized trial conducted by Munkhuu et al (86) demonstrated the positive effect Bioline syphilis had in rural Mongolia in detecting syphilis cases and the

number of patients and their partners who have been treated as a result of POC testing. Using Bioline syphilis in a one-stop service versus conventional testing, there was statistically significant increased detection rate of syphilis, 1.9% versus 0.9% , increased treatment partners, 94.6% versus 55.2%, and decreased congenital syphilis, 0.03% versus 0.42%. Also, more syphilis infected women, 98.9% versus 89.6%, were treated using one-stop service, the difference was not statistically significant.

4.2.5. Prevalence

Prevalence was defined as the proportion of individuals found to have a positive result using the POC test under investigation. In the study by Amadi et al (87), dental clinic attendees had a prevalence of 1% using syphilis Ultra Rapid test (ACON, USA).

The Determine test was used by Miranda et al (85), Garcia et al (79), Revollo et al (88), Seguy et al (83), Hurtado et al (89), and Lahuerta et al (84) to determine prevalence of TP antibodies. In the study by Hurtado et al (89), the prevalence was 5% in MSM tested in saunas and 2.3% in MSM tested in flats where prostitution is practiced. In studies by Miranda et al (85), Garcia et al (79), and Seguy et al (83), the prevalence of syphilis among ANC attendees in Brazil, Bolivia, and Guyanese gold and diamond miners was 0.4%, 5% and 6.4% respectively. Lahuerta et al (84) tested non-risk populations, MSM/TG and FSWs in mobile van clinics as well as traditional clinics. The prevalence of syphilis ranged from 0% in MSM/TG presenting to mobile van clinics to a high of 20% among FSWS presenting to moving van clinics.

4.2.6. Barriers and Challenges

Garcia et al (79) noted a lack of political will from leaders as required to assure success of antenatal programs to prevent congenital syphilis.

In the study by Munkhuu et al (82), providers indicated that they were worried about where to get the testing material once the study was over. The majority stated that one-stop service is time consuming and requires good management for smooth operation. Other barriers cited were a lack of strategies for women to inform their husbands of their status without leading to partner violence. As a result, providers recommended presumptive treatment of partners.

In Juarez-Figueroa(14), participants cited difficulty with reading the results as the lines on the test strip were faint.

In the study by Sabido et al (80), 60% of clinicians lacked confidence in the results because they correctly identified that POC test could not differentiate between past and present infections. Munkhuu et al (86), Benzaken et al (7) and Bronzan et al (8) also noted unnecessary treatment.

4.2.7. Economic Evaluations

Studies which conducted economic evaluation of syphilis POC tests were Benzaken et al (7), Blandford et al (90), Gianino et al (12), Levin et al (91), Schackman et al (92), Vickerman et al (93), and Rydzak et al (94). Evaluation of the methods employed to

conduct economic evaluations is beyond the expertise of this author. However, a summary of the conclusions reached by each author and the apparent limitations of each article will be discussed.

Benzaken et al (7), in an STD clinic in Brazil, found that that Visitect was less cost effective than VDRL, which is the conventional test at the setting. In an STD clinic in Italy, Ginanino et al(12) found opposite results. The authors found that the Determine test was actually more cost effective but this was compared to ELISA, a different reference standard than that used in Benzaken et al (7). Blandford et al (90) and Rydzak et al(94) conducted their studies in ANC clinic attendees in South Africa while Schackman et al(92) conducted their studies in ANC clinic attendees in Haiti. Benzaken et al (7) and Blandford et al (90) compared the POC test to other screening options available which included screening with RPR and TPHA, and onsite RPR. Rydzak et al (94) compared 3 POC tests to syndromic management, universal treatment, RPR with return for results, and POC test in combination with confirmatory RPR. These three studies found that POC tests are more cost-effective than other strategies. A study by Vickerman et al (93) in ANC clinic attendees in Tanzania compared 4 POC tests and RPR. The authors found one of the tests, Bioline, was just as cost effective as RPR while the rest were less cost-effective. In order to increase the cost-effectiveness of POC testing strategies, the authors concluded that the POC test itself needs to be less expensive. Levin et al(91) conducted their study in ANC clinic attendees in Bolivia and Mozambique. The authors adopted two different strategies for rural and urban settings. In rural

settings they compared the cost of POC test with no screening, and, in urban settings they compared the cost of POC to RPR. The authors conducted a simple cost comparison analysis and found that POC testing cost more than both RPR and no screening strategy.

Chapter 5: DISCUSSION

The discussion chapter is divided into two parts, corresponding to the two main objectives.

5.1. Part I

To our knowledge, this is the first meta-analysis of its kind reviewing the diagnostic accuracy of POC tests of syphilis at the global level and adjusting for imperfect reference standards using state-of-the-art HSROC Bayesian method.

5.1.2. Summary of main findings

In 29 of the 30 articles of diagnostic accuracy identified, POC tests determined the presence of antibodies specific to TP. Only one test detected the presence of specific as well as non-specific antibodies to TP.

Based on the stratification described in Figure 4 and assuming a perfect reference standard, in serum samples, using a TP specific reference standard, Determine had the highest pooled sensitivity, 98.81% (96.52 , 99.98), followed by Bioline, 93.99 (88.79, 98.45), Visitect, 93.96% (89.24, 98.01) and Syphicheck, 81.91% (69.10, 93.19). Syphicheck had the highest pooled specificity, 99.05% (98.35, 99.62), followed by Bioline, 98.58 (97.54, 99.31), Visitect, 98.51% (97.70, 99.19), and Determine, 97.94% (96.30, 98.48). In whole blood samples, using a TP specific reference standard, Bioline had the highest pooled sensitivity, 87.70% (84.78, 90.58), followed by Determine, 85.55% (76.15, 94.49), Visitect, 78.05% (70.34, 85.02), and Syphicheck, 76.88 %(68.97,

84.71). Visitect had the highest pooled specificity, 99.52% (99.16, 99.82), followed by Determine, 99.50% (98.95, 99.93), Syphicheck, 99.41%(99.05, 99.73), and Bioline, 99.07% (98.50 , 99.59).

Assuming an imperfect reference standard, in serum samples, using a TP specific reference standard, Bioline had the highest pooled sensitivity, 99.67% (95% credible interval 97.65, 100) , followed by Determine , 99.14% (96.93, 100), Visitect, 98.18% (93.53, 100) and Syphicheck, 88.46% (73.54, 99.87). Syphicheck had the highest pooled specificity, 99.98% (99.64, 100), followed by Visitect, 99.89% (99.19, 100), Determine, 99.68% (98.70, 100) and Bioline, 99.56% (98.55, 100). In whole blood, Bioline had the highest pooled sensitivity, 91.47% (87.06, 96.12), followed by Determine, 89.49% (79.88, 98.15), Visitect, 82.93% (94.50, 100) and Syphicheck, 81.99% (71.84, 91.99). Determine had the highest pooled specificity, 99.91% (99.44, 1) followed by Visitect, 99.87% (99.58, 100) followed by Syphicheck, 99.81% (99.46, 100), and Bioline, 99.61% (99.04, 100).

Overall, sensitivity and specificity estimates were higher in serum than in whole blood samples. After adjusting for imperfect reference standard, new sensitivity and specificity of POC tests improved.

5.1.3. Applicability of Findings

Ideally, if a TP specific POC test is as accurate as current TP specific serological tests, it can be used in resource limited settings and hard to reach populations. A study by

Peeling et al, in 2004, (95) estimated the sensitivity and specificity of serological TP tests. Enzyme Immunoassay (EIA) has a sensitivity of 82-100% and a specificity of 97-100%. TPHA and TPPA have a sensitivity of 85-100% and specificity of 98-100%. Finally, FTA-ABS has a sensitivity of 70-100% and a specificity of 94-100%. From these figures it can be seen that the four diagnostic tests meta-analyzed easily fall within these ranges. It is therefore appropriate to use POC tests to screen for syphilis, particularly in settings where access to laboratories is limited or if patients do not return for results. POC tests had higher accuracy parameters when used with serum rather than whole blood. Based on this, to fully benefit from the potential of POC tests, testing using serum samples is endorsed where facilities exist to obtain serum from whole blood.

There are limitations in the use of POC syphilis tests that arise from the property of the test itself. With the current technology that is widely used, POC tests can only detect treponemal antibodies and hence provide only a fraction of information required by clinicians to choose appropriate course of treatment. The treatment administered is dependent upon the stage of infection. At the latent stage, three doses of penicillin over a three week period is required while for all the other stages, a single dose is appropriate.(44) Information regarding staging of disease can only be inferred from results of non-treponemal tests. Non-treponemal tests are also important in tracking the progress in treatment. Treponemal tests cannot provide information on course of treatment because they remain positive even if the patient is treated. (5) With these limitations, POC tests are only useful in detecting infection in undiagnosed or those

diagnosed but not treated patients. In short, POC tests should be recommended as screening tools only, pending confirmation with conventional laboratory based tests.

5.2. Part II

5.2.1. Summary of findings

Acceptability of POC test were high and is expected by clinicians and patients is expected. Establishing POC testing and POC testing strategies is feasible, particularly in low income settings. Results on preference favouring POC were not conclusive. While some studies found high preference, others were unclear. Prevalence was variable depending on the population tested. Impact was demonstrated, especially so by Munkhuu et al (21) using a clustered randomized trial. Due to the heterogenity of methodology employed, no conclusive statements regarding economoic evaluation of syphilis POC tests can be drawn.

5.2.2. Applicability of findings

The findings in this part were very heterogeneous, both in methodology and in results obtained and reported. To improve research, a set of standards which explicitly defines each outcome needs to be established. Only in this way can various different studies and POC tests be compared fairly.

A recent development is the creation of an IAF (Impact Assessment Framework) by researchers at the Liverpool School of Tropical Medicine which can be used in order to systematically asssess new diagnostic tools for tuberculosis (TB), which is explained in

detail in a paper by Mann et al. (96) The IAF evolved in order to address the goals of the Millenium Development Goals and the Global Plan to Stop TB.(96) This framework has five layers of analysis, evaluating effectiveness, equity, health systems, scale-up and policy. Such a framework is needed for syphilis. This tool would allow for a transparent method of evaluation of tests and most importantly, use of a consistent framework by researchers would allow for direct comparisons.

Finally, it must be noted that despite the fact that the outcomes in part II are not defined in a systematic manner and inconsistent definitions are used by various authors, these outcomes remain extremely important in evaluting POC tests. As such, their inclusion in this thesis is warranted for it furthers the knowledge of the implementation research outcomes required to ease the transition of POC tests into the healthcare system for the benefit of those who need them most.

5.3. Strengths and weaknesses

To our knowledge, this is the first meta-analysis reviewing the diagnostic accuracy of POC tests of syphilis at the global level that adjusted for imperfect reference standards using state-of-the-art HSROC Bayesian method. A recent publication by Tucket et al (29) conducted a meta-analysis of POC ICS syphilis tests. The authors considered the performance of ICS tests in STI and ANC clinic attendees while this study used a different stratification strategy. As well, Tucker et al(29) used a frequentist approach to pooling their results while this study used a Bayesian approach, allowing for adjustment of prior

distributions, specifically adjusting for the accuracy of reference standards used. Tucker et al(29) only explored English studies while we had no such limitations. Another improvement over Tucker's study is that we critiqued the included papers on their quality. Additionally, we evaluated outcomes beyond accuracy known as implementation research outcomes.

There were some limitations to our review. All eligible studies may not have been captured by our search strategy. Although not including abstracts, reports and bulletins raised the quality of the review, we may have missed legitimate information. As well, during the time period captured, there may have been change in guidelines in testing such as changes in testing algorithms that were not captured in this meta-analysis. The reviewers were not blinded to the study authors and institution which can introduce reviewer bias.

There are also limitations of the statistical analysis conducted. Prior distributions used for the HSROC analysis with imperfect reference standard were obtained from literature. However, there is no guarantee that these ranges are correct. Hence, interpretations of Bayesian inferences must be made conditional on choice of priors.

There were limitations associated with the studies included in the systematic review of diagnostic accuracy, Part I. First, there was improper evaluation of POC test by using an inappropriate reference standard. Syphilis POC tests detect TP specific antibodies and therefore reference standards that detect TP specific antibodies should be used.

However, some authors used reference standards that detect non-TP specific antibodies or combination of reference standards that detect TP and non-TP specific antibodies. Despite rigorous methodological and reporting quality of some studies, results from the index test and reference test are not directly comparable as they detect presence of different biological specimen in the sample. Second, there is a possibility of publication bias, with studies that showed higher accuracy being published. There is not, however, an appropriate method to investigate publication bias as explained by Leeflang et al (68). Since there is no set level of significance for diagnostic tests to be considered "significantly" accurate, the funnel plot, current method employed in randomized controlled trials and observational studies, is not a proper investigative tool for diagnostic studies.(68) Further work in this area would highly improve the quality of meta-analyses. Third, the results of studies may be influenced by industry participation. This is demonstrated through the reporting of conflict of interest by the authors. Only 37% of the included studies reported conflict of interest, leaving the reviewer unclear about the position of the other authors. A greater reporting of conflicts of interest should be encouraged by journals in order to make research more transparent. Also, some evaluations were not done independently as some authors evaluated the test they had manufactured themselves, introducing bias. Since only 23% of papers reported whether they had conducted their study blindly, there is a great possibility of reviewer bias. Lack of blinding bias occurs when the person reading the POC result knows the disease status of the patient, which often leads to over estimation of sensitivity and specificity. (62)

Similarly, there were limitations of studies included in the implementation research outcomes systematic review, Part II. There were limitation due to the studies included in the systematic review. There were inconsistent definitions of acceptance, preference, feasibility, impact and economic evaluations. The majority of articles did not explain in their methodology how they defined each term.

Chapter 6: CONCLUSION

6.1. Implications for practice

Comparison of global performance of major POC syphilis kits are very useful for policy recommendations by local, domestic, and global stakeholders . This analysis answered questions regarding accuracy and implementation outcomes of POC syphilis tests. . This study examined the global evidence on the accuracy of POC tests, analyzing results by test kit, sample and reference standard used. From the meta-analysis results of this review, syphilis POC tests have accepable sensitivity and specificty and are on par or better than the laboratory-based treponemal serological tests currently used. Using the tests in serum resulted in higher sensitivity and speificity than in whole blood. Use of POC tests for detection of syphilis is recommended for screening pending confirmation by conventional testing.Testing in serum, if the resources permit, is preferred. Analysis of non-accuracy outomes confirmed that the testing was acceptable and feasibility and impact was demonstrated.

6.2. Implications for research

Overall, there needs to be an improvement in diagnostic accuracy studies, especially with regards to using correct reference standard tests. Steps to improve quality control and quality assurance also need to be taken, particularly when testing is performed outside of laboratories. For non-accuracy outcomes, better definitions and documentation of outcomes is warranted.

A great step forward is the availability POC tests which simulataneoulsy detect TP and non-TP antibodies. An example is the test manufactured by Zarakolu et al (36). These tests provide a great advantage over current POC tests which only detect TP antibodies, as they allow past and present infections to be disintguished. These tests have not yet been extensively evaluated in field settings. With the accumulation of enough evidence, in the near future, a new meta-analysis of the results of such tests is useful.

6.3. Concluding Remarks

Without political will and consistent championing by researchers and advocates for appropriate technologies, it is difficult to improve health of hard to reach populations, allowing continued disease transmission and death. In areas where individuals currently do not have access to any type of testing, introducing current POC testing and treatmentwill save lives even despite the lack of 100% sensitive and specific POC tests.

BIBLIOGRAPHY

1. Kasper DL, Harrison TR. Harrison's principles of internal medicine. 16th ed. New York: McGraw-Hill, Medical Pub. Division; 2005.
2. Steinbrook R. HIV in India--a complex epidemic. N Engl J Med. 2007 Mar 15;356(11):1089-93.
3. Peeling RW, Mabey D, Herring A, Hook EW, 3rd. Why do we need quality-assured diagnostic tests for sexually transmitted infections? Nat Rev Microbiol. 2006 Dec;4(12):909-21.
4. Herring A, Ballard R, Mabey D, Peeling RW. Evaluation of rapid diagnostic tests: syphilis. Nat Rev Microbiol. 2006 Dec;4(12 Suppl):S33-40.
5. Peeling RW, Ye H. Diagnostic tools for preventing and managing maternal and congenital syphilis: an overview. Bull World Health Organ. 2004 Jun;82(6):439-46.
6. Benzaken AS, Garcia EG, Sardinha JCG, Dutra JC, Peeling R. Rapid tests for diagnosing syphilis: validation in an STD clinic in the Amazon Region, Brazil. Cad Saude Publica. [Article]. 2007;23:S456-S64.
7. Benzaken AS, Sabido M, Galban EG, Pedroza V, Vasquez F, Araujo A, et al. Field evaluation of the performance and testing costs of a rapid point-of-care test for syphilis in a red-light district of Manaus, Brazil. Sex Transm Infect. [Evaluation Studies Research Support, Non-U.S. Gov't]. 2008 Aug;84(4):297-302.
8. Bronzan RN, Mwesigwa-Kayongo DC, Narkunas D, Schmid GP, Neilsen GA, Ballard RC, et al. Onsite rapid antenatal syphilis screening with an immunochromatographic Strip improves case detection and trip treatment in rural south African clinics. Sex Transm Dis. [Article]. 2007 Jul;34(7):S55-S60.
9. Campos PE, Buffardi AL, Chiappe M, Buendia C, Garcia PJ, Carcamo CP, et al. Utility of the Determine Syphilis TP rapid test in commercial sex venues in Peru. Sex Transm Infect. [Multicenter Study
Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. 2006 Dec;82 Suppl 5:v22-5.
10. Castro AR, Mody HC, Parab SY, Patel MT, Kikkert SE, Park MM, et al. An immunofiltration device for the simultaneous detection of non-treponemal and treponemal antibodies in patients with syphilis. Sex Transm Infect. 2010 Dec;86(7):532-6.
11. Diaz T, Almeida MDB, Georg I, Maia SD, de Souza RV, Markowitz LE. Evaluation of the determine rapid syphilis TP assay using sera. Clin Diagn Lab Immunol. [Article]. 2004 Jan;11(1):98-101.
12. Gianino MM, Dal Conte I, Sciole K, Galzerano M, Castelli L, Zerbi R, et al. Performance and costs of a rapid syphilis test in an urban population at high risk for sexually transmitted infections. Journal of Preventive Medicine & Hygiene. 2007;48(4):118-22.
13. Hernandez-Trejo M, Hernandez-Prado B, Uribe-Salas F, Juarez-Figueroa L, Conde-Gonzalez CJ. [Maternal and congenital syphilis in two Mexican hospitals: evaluation of a rapid diagnostic test]. Revista de Investigacion Clinica. 2006;58(2):119-25.
14. Juarez-Figueroa L, Uribe-Salas F, Garcia-Cisneros S, Olamendi-Portugal M, Conde-Glez CJ. Evaluation of a rapid strip and a particle agglutination tests for syphilis diagnosis. Diagn Microbiol Infect Dis. [Article]. 2007 Oct;59(2):123-6.
15. Li J, Zheng HY, Wang LN, Liu YX, Wang XF, Liu XR. Clinical evaluation of four recombinant Treponema pallidum antigen-based rapid diagnostic tests for syphilis. Journal of the European Academy of Dermatology & Venereology. 2009;23(6):648-50.
16. Lien TX, Tien NT, Chanpong GF, Cuc CT, Yen VT, Soderquist R, et al. Evaluation of rapid diagnostic tests for the detection of human immunodeficiency virus types 1 and 2, hepatitis B surface antigen, and syphilis in Ho Chi Minh City, Vietnam. American Journal of Tropical Medicine & Hygiene. 2000;62(2):301-9.
17. Mabey D, Peeling RW, Ballard R, Benzaken AS, Galban E, Changalucha J, et al. Prospective, multi-centre clinic-based evaluation of four rapid diagnostic tests for syphilis. Sex Transm Infect. 2006;13(82).
18. Mishra S, Naik B, Venugopal B, Kudur P, Washington R, Becker M, et al. Syphilis screening among female sex workers in Bangalore, India: comparison of point-of-care testing and traditional serological approaches. Sex Transm Infect. [Research Support, Non-U.S. Gov't]. 2010 Jun;86(3):193-8.
19. Montoya PJ, Lukehart SA, Brentlinger PE, Blanco AJ, Floriano F, Sairosse J, et al. Comparison of the diagnostic accuracy of a rapid immunochromatographic test and the rapid plasma reagin test for antenatal syphilis screening in Mozambique. Bull World Health Organ. [Article]. 2006 Feb;84(2):97-104.
20. Nessa K, Alam A, Chawdhury FA, Huq M, Nahar S, Salauddin G, et al. Field evaluation of simple rapid tests in the diagnosis of syphilis. International Journal of STD & AIDS. 2008;19(5):316-20.
21. Nyamwamu L, Gicheru M, Sharma R, Kimutai A, Tonui W, Ngure P. Evaluation of the immunochromatographic strip test for the rapid diagnosis of antenatal syphilis in women in Eldoret, Kenya. Journal of Nanjing Medical University. 2009;23(5):317-21.
22. Oshiro M, Taira R, Kyan T, Yamane N. Laboratory-based evaluation of DainaScreen TPAb to detect specific antibodies against Treponema pallidum. Rinsho Biseibutsu Jinsoku Shindan Kenkyukai shi = JARMAM : Journal of the Association for Rapid Method and Automation in Microbiology. 1999;10(1):27-32.
23. Qiajiao Hh, Lan X, Tong T, Wu X, Chen M, Feng X, et al. Dot-immunogold filtration assay as a screening test for syphilis. J Clin Microbiol. 1996 Aug;34(8):2011-3.
24. Rotanov SV, Frigo NV, Kliueva VI. [Comparative study of immunochromatographic sets for rapid diagnosis of syphilis]. Klin Lab Diagn. 2008 Feb(2):42-5.
25. Sano J, Iwakabe A, Ohno H, Aoki Y. Study of anti-Treponema pallidum antibody detection kit <DAINASCREEN-TPAb> employing immunochromatography assay. Journal of Analytival Bio-Science. 1999;22(4):351-4.
26. Sato NS, de Melo CS, Zerbini LC, Silveira EP, Fagundes LJ, Ueda M. Assessment of the rapid test based on an immunochromatography technique for detecting anti-Treponema pallidum antibodies. Rev Inst Med Trop Sao Paulo. 2003 Nov-Dec;45(6):319-22.
27. Herring AJ, Ballard RC, Pope V, Adegbola RA, Changalucha J, Fitzgerald DW, et al. A multi-centre evaluation of nine rapid, point-of-care syphilis tests using archived sera. Sex Transm Infect. 2006 Dec;82 Suppl 5:v7-12.

28. Tinajeros F, Grossman D, Richmond K, Steele M, Garcia SG, Zegarra L, et al. Diagnostic accuracy of a point-of-care syphilis test when used among pregnant women in Bolivia. Sex Transm Infect. 2006 Dec;82 Suppl 5:v17-21.
29. Tucker JD, Bu J, Brown LB, Yin YP, Chen XS, Cohen MS. Accelerating worldwide syphilis screening through rapid testing: a systematic review. Lancet Infect Dis. 2010 Jun;10(6):381-6.
30. van Dommelen L, Smismans A, Goossens VJ, Damoiseaux J, Bruggeman CA, van Tiel FH, et al. Evaluation of a rapid one-step immunochromatographic test and two immunoenzymatic assays for the detection of anti-Treponema pallidum antibodies. Sex Transm Infect. 2008 Aug;84(4):292-6.
31. Villazon-Vargas N, Conde-Glez CJ, Juarez-Figueroa L, Uribe-Salas F. [Evaluation of a rapid diagnostic test to assess the prevalence of maternal syphilis in Bolivia]. Rev Med Chil. 2009 Apr;137(4):515-21.
32. Wang LN, Yang L, Zheng HY. Clinical evaluation of four recombinant Treponema pallidum antigen-based rapid tests in the diagnosis of syphilis. Chin Med Sci J. 2007 Dec;22(4):250-3.
33. Wang L-n, Zheng H-y, Li J, Wang X-f, Shao Y-l, Wang Y-q, et al. Clinical evaluation of 4 recombinant *Treponema pallidum* antigen based rapid tests in the diagnosis of syphilis. . Journal of Clinical Dermatology. 2005;34(11):726-8.
34. West B, Walraven G, Morison L, Brouwers J, Bailey R. Performance of the rapid plasma reagin and the rapid syphilis screening tests in the diagnosis of syphilis in field conditions in rural Africa. Sex Transm Infect. 2002 Aug;78(4):282-5.
35. Yang H, Li D, He R, Guo Q, Wang K, Zhang X, et al. A Novel Quantum Dots-Based Point of Care Test for Syphilis. Nanoscale Res Lett. 2010;5(5):875-81.
36. Zarakolu P, Buchanan I, Tam M, Smith K, Hook EW. Preliminary evaluation of an immunochromatographic strip test for specific Treponema pallidum antibodies. J Clin Microbiol. [Article]. 2002 Aug;40(8):3064-5.
37. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008 May 17;336(7653):1106-10.
38. Hackett CJ. On the Origin of the Human Treponematoses (Pinta, Yaws, Endemic Syphilis and Venereal Syphilis). Bull World Health Organ. 1963;29:7-41.
39. Kent ME, Romanelli F. Reexamining syphilis: an update on epidemiology, clinical manifestations, and management. Ann Pharmacother. 2008 Feb;42(2):226-36.
40. Brown DL, Frank JE. Diagnosis and management of syphilis. Am Fam Physician. 2003 Jul 15;68(2):283-90.
41. Gamble VN. Under the shadow of Tuskegee: African Americans and health care. Am J Public Health. 1997 Nov;87(11):1773-8.
42. Corbie-Smith G. The continuing legacy of the Tuskegee Syphilis Study: considerations for clinical investigation. Am J Med Sci. 1999 Jan;317(1):5-8.
43. SDI. The uses of Rapid Syphilis Tests. Geneva: TDR/WHO2006.
44. Khayat A, Rathore M. Treponema pallidum (Neurosyphilis). In: Barton L, Friedman N, editors. The Neurological Manifestations of Pediatric Infectious Diseases and Immunodeficiency Syndromes: Humana Press.; 2008. p. 287-95.
45. CDC. CDC Fact Sheet: Syphilis & MSM: CDC2007.
46. PHAC. Report on Sexually Transmitted Infections in Canada: 20082008.
47. Egglestone SI, Turner AJ. Serological diagnosis of syphilis. PHLS Syphilis Serology Working Group. Commun Dis Public Health. 2000 Sep;3(3):158-62.
48. Workowski KA, Berman S. Sexually transmitted diseases treatment guidelines, 2010. MMWR Recomm Rep. 2010 Dec 17;59(RR-12):1-110.
49. Badri T, Ben Jennet S. Rash Associated with Secondary Syphilis. New Engl J Med. 2011;364(1):71-.
50. Douglas JM, Jr. Penicillin treatment of syphilis: clearing away the shadow on the land. JAMA. 2009 Feb 18;301(7):769-71.
51. De Cock KM, Bunnell R, Mermin J. Unfinished business--expanding HIV testing in developing countries. N Engl J Med. 2006 Feb 2;354(5):440-2.
52. Larsen SA, Steiner BM, Rudolph AH. Laboratory diagnosis and interpretation of tests for syphilis. Clin Microbiol Rev. 1995 Jan;8(1):1-21.
53. McMillan A, Young H. Qualitative and quantitative aspects of the serological diagnosis of early syphilis. Int J STD AIDS. 2008 Sep;19(9):620-4.
54. Greer L, Wendel GD, Jr. Rapid diagnostic methods in sexually transmitted infections. Infect Dis Clin North Am. 2008 Dec;22(4):601-17, v.
55. Peeling RW, Mabey DC. Syphilis. Nat Rev Microbiol. 2004 Jun;2(6):448-9.
56. Petti CA, Polage CR, Quinn TC, Ronald AR, Sande MA. Laboratory medicine in Africa: a barrier to effective health care. Clin Infect Dis. 2006 Feb 1;42(3):377-82.
57. Peeling RW, Mabey D, Herring A, Hook EW. Why do we need quality-assured diagnostic tests for sexually transmitted infections? Nature Reviews Microbiology. 2006 Dec:S7-S19.
58. Gift TL, Pate MS, Hook EW, 3rd, Kassler WJ. The rapid test paradox: when fewer cases detected lead to more cases treated: a decision analysis of tests for Chlamydia trachomatis. Sex Transm Dis. 1999 Apr;26(4):232-40.
59. WHO. DoMPS. Prevention of mother-to-child transmission of syphilis. Geneva: WHO; 2006.
60. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
61. Szklo M, Nieto FJ. Epidemiology : beyond the basics. 2nd ed. Sudbury, Mass.: Jones and Bartlett Publishers; 2007.
62. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003 Jan 7;138(1):W1-12.
63. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol. 2003 Nov;56(11):1118-28.
64. Cornell J, Mulrow CD, Localio AR. Diagnostic test accuracy and clinical decision making. Ann Intern Med. 2008 Dec 16;149(12):904-6.

65. Atkins D, Kupersmith J. Implementation research: a critical component of realizing the benefits of comparative effectiveness research. Am J Med. 2010 Dec;123(12 Suppl 1):e38-45.
66. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 Nov 10;3:25.
67. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004 Feb 3;140(3):189-202.
68. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM, Accuracy CDT. Systematic Reviews of Diagnostic Test Accuracy. Ann Intern Med. 2008 Dec 16;149(12):889-+.
69. Roger MH, Penny W. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. Stata Journal. 2009;9(2):211-29.
70. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001 Oct 15;20(19):2865-84.
71. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009 Jul 21;6(7):e1000100.
72. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009 Jul 21;6(7):e1000097.
73. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.
74. Cochrane C. Handbook for DTA Reviews. 2011.
75. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. BMC medical research methodology. [Research Support, Non-U.S. Gov't
Validation Studies]. 2006;6:12.
76. Siedner M, Zapitz V, Ishida M, De La Roca R, Klausner JD. Performance of rapid syphilis tests in venous and fingerstick whole blood specimens. Sex Transm Dis. 2004 Sep;31(9):557-60.
77. Leeflang MMG, Debets-Ossenkopp YJ, Visser CE, Bossuyt PMM. Meta-analysis of diagnostic test accuracy. Clin Infect Dis. 2006 Nov 1;43(9):1220-.
78. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold-standard reference. Biometrics (submitted).
79. Garcia SG, Tinajeros F, Revollo R, Yam EA, Richmond K, Diaz-Olavarrieta C, et al. Demonstrating public health at work a demonstration project of congenital syphilis prevention efforts in Bolivia. Sex Transm Dis. [Article]. 2007 Jul;34(7):S37-S41.
80. Sabidó M, Benzaken AS, de-Andrade-Rodrigues EJ, Mayaud P. Rapid Point-of-Care Diagnostic Test for Syphilis in High-Risk Populations, Manaus, Brazil. Emerging Infectious Diseases. 2009;15(4):647-9.
81. Lee D, Fairley C, Cummings R, Bush M, Read T, Chen M. Men who have sex with men prefer rapid testing for syphilis and may test more frequently using it. Sex Transm Dis. 2010;37(9):557-8.
82. Munkhuu B, Liabsuetrakul T, McNeil E, Janchiv R. Feasibility of one-stop antenatal syphilis screening in Ulaanbaatar, Mongolia: women and providers perspectives. Southeast Asian J Trop Med Public Health. [Research Support, Non-U.S. Gov't]. 2009 Jul;40(4):861-70.
83. Seguy N, Denniston M, Hladik W, Edwards M, Lafleur C, Singh-Anthony S, et al. HIV and Syphilis Infection among Gold and Diamond Miners - Guyana, 2004. West Ind Med J. [Article]. 2008 Nov;57(5):444-9.
84. Lahuerta M, Sabidó M, Giardina F, Hernández G, Palacios JF, Ortiz R, et al. Comparison of users of an HIV/syphilis screening community-based mobile van and traditional voluntary counselling and testing sites in Guatemala. Sexually Transmitted Infections.
85. Miranda AE, Rosetti E, Trindade CR, Gouvea GM, Costa DM, Oliveira TG, et al. Prevalence of syphilis and HIV using rapid tests among parturients attended in public maternity hospitals in Vitoria, State of Espirito Santo. Rev Soc Bras Med Trop. [Article]. 2009 Jul-Aug;42(4):386-91.
86. Munkhuu B, Liabsuetrakul T, Chongsuvivatwong V, McNeil E, Janchiv R. One-stop service for antenatal syphilis screening and prevention of congenital syphilis in Ulaanbaatar, Mongolia: a cluster randomized trial. Sex Transm Dis. 2009 Nov;36(11):714-20.
87. Amadi ES, Ononiwu CE, Aballa N, Oladimeji SA, Aneke FA, Ugwuoru DC. PREVALENCE OF SYPHILIS AMONG PATIENTS IN A DENTAL CLINIC IN ENUGU, NIGERIA. Pak J Med Sci. [Article]. 2010 Jan-Mar;26(1):59-61.
88. Revollo R, Tinajeros F, Hilari C, Garcia SG, Zegarra L, Diaz-Olavarrieta C, et al. [Maternal and congenital syphilis in four provinces in Bolivia]. Salud Publica Mex. [English Abstract]. 2007 Nov-Dec;49(6):422-8.
89. Hurtado I, Alastrue I, de Olalla PG, Albiach D, Martin M, Perez-Hoyos S. Preventive intervention in venues for interaction used by men who have sex with men. Gac Sanit. [Article]. 2010 Jan;24(1):78-80.
90. Blandford JM, Gift TL, Vasaikar S, Mwesigwa-Kayongo D, Dlali P, Bronzan RN. Cost-effectiveness of on-site antenatal screening to prevent congenital syphilis in rural eastern Cape Province, Republic of South Africa. Sex Transm Dis. [Evaluation Studies]. 2007 Jul;34(7 Suppl):S61-6.
91. Levin CE, Steele M, Atherly D, Garcia SG, Tinajeros F, Revollo R, et al. Analysis of the operational costs of using rapid syphilis tests for the detection of matemal syphilis in Bolivia and Mozambique. Sex Transm Dis. [Article]. 2007 Jul;34(7):S47-S54.
92. Schackman BR, Neukermans CP, Fontain SN, Nolte C, Joseph P, Pape JW, et al. Cost-effectiveness of rapid syphilis screening in prenatal HIV testing programs in Haiti. PLoS Med. 2007 May;4(5):e183.
93. Vickerman P, Peeling RW, Terris-Prestholt F, Changalucha J, Mabey D, Watson-Jones D, et al. Modelling the cost-effectiveness of introducing rapid syphilis tests into an antenatal syphilis screening programme in Mwanza, Tanzania. Sex Transm Infect. [Research Support, Non-U.S. Gov't]. 2006 Dec;82 Suppl 5:v38-43.

94. Rydzak CE, Goldie SJ. Cost-effectiveness of rapid point-of-care prenatal syphilis screening in sub-Saharan Africa. Sex Transm Dis. 2008 Sep;35(9):775-84.
95. Peeling RW, Ye H. Diagnostic took for preventing and managing maternal and congenital syphilis: an overview. Bulletin of the World Health Organization. [Article]. 2004 Jun;82(6):439-46.
96. Mann G, Squire SB, Bissell K, Eliseev P, Du Toit E, Hesseling A, et al. Beyond accuracy: creating a comprehensive evidence base for TB diagnostic tools. Int J Tuberc Lung Dis. 2010 Dec;14(12):1518-24.
97. WorldBank. 2008 Country Classification Tables. Washington, D.C.: The World Bank; 2009; Available from: http://go.worldbank.org/K2CKM78CC0.

APPENDICES

Tables:

Table 1: PRISMA checklist (72)
*This table is split over 3 pages.

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	

Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	

DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

Table 2: Extraction sheet used for Part I and Part II

Table x: Data Abstraction Form Part I and II
Study ID
Author
Date of study
Country that study was performed in
Study Population
Vitro/Vivo
Type of Test
Original Sample size
Inclusion criteria
Exclusion Criteria
Blinding
Study Design
Index Test
performing Index Test
Reference test
Index Indeterminate results
Reference Indeterminant
Final sample size
TP
FN
FP
TN
Sensitivity
Specificity

Table x: Data abstraction form Part II only
Acceptability
Preference
Feasibility
Impact
Barriers/Challenges

Table 3: Characteristics and results of studies included in Part I.

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
1	Benzaken(61)	2007	Manaus, Brazil (UMI)	Cross-sectional	FTA-Abs	Standard Bioline	Whole Blood	45	6	3	487	STD clinic attendees
1	Benzaken(61)	2007	Manaus, Brazil (UMI)	Cross-sectional	FTA-Abs	Standard Bioline	Serum	46	5	3	487	STD clinic attendees
1	Benzaken(61)	2007	Manaus, Brazil (UMI)	Cross-sectional	FTA-Abs	Qualpro Syphicheck	Whole Blood	43	8	2	488	STD clinic attendees
1	Benzaken(61)	2007	Manaus, Brazil (UMI)	Cross-sectional	FTA-Abs	Qualpro Syphicheck	Serum	45	6	2	488	STD clinic attendees
2	Benzaken(62)	2008	Manaus, Brazil (UMI)	Cross-sectional	FTA-Abs	Omega Visitect	Whole Blood	52	40	5	409	Outreach clinic
3	Bronzan(63)	2007	Cape town, South Africa (UMI)	Cross-sectional	RPR + TPHA	Determine	Whole Blood	20	2	29	290	ANC attendees
4	Campos(a)(57)	2006	Lima, Peru (UMI)	Cross-sectional	RPR reactivity at any dilution + TPHA	Determine	Whole Blood	70	108	28	3277	FSWs in commercial sex venues
4	Campos(b)(40)	2006	Lima, Peru (UMI)	Cross-sectional	RPR >=1:8 + TPHA	Determine	Whole Blood	16	9	82	3376	FSWs in commercial sex venues
4	Campos(c)(40)	2006	Lima, Peru (UMI)	Cross-sectional	RPR>=1:16 + TPHA	Determine	Whole Blood	7	3	91	3382	FSWs in commercial sex venues
5	Castro(a)(27)	2010	Atlanta, United States (HI)	Case-Control	RPR	Span Diagnostics	Serum	111	4	6	255	Clinical samples

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
5	Castro(b)(27)	2010	Atlanta, United States (HI)	Case-Control	TPPA	Span Diagnostics	Serum	147	4	2	223	Clinical samples
6	Diaz(a)(95)	2003	Rio de Janeiro, Brazil (UMI)	Cross-sectional	TPHA	Determine	Serum	244	6	8	292	Infectious disease research attendees
6	Diaz(b)(95)	2003	Rio de Janeiro, Brazil (UMI)	Cross-sectional	TPHA	Determine	Serum	246	4	13	287	Infectious disease research attendees
6	Diaz(c)(95)	2003	Rio de Janeiro, Brazil (UMI)	Cross-sectional	TPHA	Determine	Serum	241	9	11	289	Infectious disease research attendees
7	Gianino(87)	2007	Turin, Italy (HI)	Cross-sectional	Clinical symptoms + ELISA + RPR or TP-PA or ELISA IgM	Determine	Whole Blood	94	5	5	212	High Risk
8	Hernandez-Trejo(56)	2006	Cuernavaca City and Mexico City, Mexico ((UMI)	Cross-sectional	VDRL + FTA-Abs	Determine	Whole Blood	4	0	0	1318	Pregnant Women
9	Juarez-Figueroa(a)(55)	2006	Cuernavaca City, Mexico (UMI),	Cross-sectional	VDRL + FTA-Abs	Determine	Serum	57	1	4	78	FSWs attending urban clinic

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
9	Juarez-Figueroa(b) (55)	2006	Mexico City, Mexico (UMI)	Cross-sectional	VDRL + FTA-Abs	Determine	Whole Blood	30	1	0	167	FSWs attending STI clinic
9	Juarez-Figueroa(c) (55)	2006	Cuernavaca City, Mexico (UMI)	Cross-sectional	VDRL + FTA-Abs	Determine	Whole Blood	3	0	1	196	Pregnant Women attending urban clinic
10	Li(a)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Determine	Whole Blood	68	15	2	360	STD clinic attendees
10	Li(b)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Determine	Whole Blood	64	19	0	362	STD clinic attendees
10	Li(c)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Determine	Serum	83	0	4	358	STD clinic attendees
10	Li(d)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	61	22	1	361	STD clinic attendees
10	Li(e)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	64	19	0	362	STD clinic attendees
10	Li(f)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Omega Visitect	Serum	78	5	7	362	STD clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
10	Li(g)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	57	32	1	325	STD clinic attendees
10	Li(h)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	63	26	1	325	STD clinic attendees
10	Li(i)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Serum	60	29	4	322	STD clinic attendees
10	Li(j)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Standard Bioloine	Whole Blood	78	11	1	325	STD clinic attendees
10	Li(k)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	78	11	1	325	STD clinic attendees
10	Li(l)(66)	2009	Beijing, China (LMI)	Cross-sectional	TPHA	Standard Bioline	Serum	85	4	7	319	STD clinic attendees
11	Lien(52)	2000	Ho Chi Minh City, Vietnam (LMI)	Cross-sectional	VDRL Carbon + Antigen RPR	Determine	Serum	72	0	3	216	potential cross reactives
12	Mabey(a)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Determine	Whole Blood	29	11	11	710	STI clinic attendees
12	Mabey(b)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Determine	Whole Blood	40	0	31	690	STI clinic attendees
12	Mabey(c)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Determine	Serum	40	0	31	690	STI clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
12	Mabey(d)(59)	2006	China (LMI)	Cross-sectional	TPHA	Determine	Whole Blood	68	15	2	360	STI clinic attendees
12	Mabey(e)(59)	2006	China (LMI)	Cross-sectional	TPHA	Determine	Whole Blood	64	19	0	362	STI clinic attendees
12	Mabey(f)(59)	2006	China (LMI)	Cross-sectional	TPHA	Determine	Serum	83	0	4	358	STI clinic attendees
12	Mabey(g)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Determine	Whole Blood	34	23	3	468	ANC attendees
12	Mabey(h)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Determine	Whole Blood	46	11	3	468	ANC attendees
12	Mabey(i)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Determine	Serum	52	5	10	461	ANC attendees
12	Mabey(j)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Determine	Whole Blood	46	8	4	191	STI clinic attendees
12	Mabey(k)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Determine	Serum	46	8	4	191	STI clinic attendees
12	Mabey(l)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	40	15	4	457	STI clinic attendees
12	Mabey(m)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	54	4	6	455	STI clinic attendees
12	Mabey(n)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Omega Visitect	Serum	54	1	6	455	STI clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
12	Mabey(o)(59)	2006	China(LMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	61	22	1	361	STI clinic attendees
12	Mabey(p)(59)	2006	China(LMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	63	20	0	362	STI clinic attendees
12	Mabey(q)(59)	2006	China (LMI)	Cross-sectional	TPHA	Omega Visitect	Serum	78	5	7	355	STI clinic attendees
12	Mabey(r)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	43	12	5	522	ANC attendees
12	Mabey(s)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	46	11	2	469	ANC attendees
12	Mabey(t)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Omega Visitect	Serum	48	9	4	467	ANC attendees
12	Mabey(u)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	49	2	3	190	STI clinic attendees
12	Mabey(v)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Omega Visitect	Serum	49	2	3	190	STI clinic attendees
12	Mabey(w)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	33	8	11	491	STI clinic attendees
12	Mabey(x)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	40	1	7	495	STI clinic attendees
12	Mabey(y)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Qualpro Syphicheck	Serum	40	1	8	494	STI clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
12	Mabey(z)(59)	2006	China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	57	32	1	325	STI clinic attendees
12	Mabey(aa)(59)	2006	China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	63	26	1	325	STI clinic attendees
12	Mabey(bb)(59)	2006	China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Serum	60	29	4	322	STI clinic attendees
12	Mabey(cc)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	43	12	5	522	ANC attendees
12	Mabey(dd)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	47	8	5	522	ANC attendees
12	Mabey(ee)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Qualpro Syphicheck	Serum	48	7	6	521	ANC attendees
12	Mabey(ff)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	42	8	2	490	FSWs
12	Mabey(gg)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Serum	44	6	2	490	STI clinic attendees
12	Mabey(hh)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	30	0	8	477	STI clinic attendees
12	Mabey(ii)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	29	1	7	478	STI clinic attendees
12	Mabey(jj)(59)	2006	Haiti (LI)	Cross-sectional	TPHA	Standard Bioline	Serum	30	0	7	478	STI clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
12	Mabey(kk)(59)	2006	China (LMI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	78	11	2	324	STI clinic attendees
12	Mabey(ll)(59)	2006	China (LMI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	78	11	2	324	STI clinic attendees
12	Mabey(mm)(59)	2006	China (LMI)	Cross-sectional	TPHA	Standard Bioline	Serum	85	4	7	319	STI clinic attendees
12	Mabey(nn)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	57	9	10	506	ANC attendees
12	Mabey(oo)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	60	6	20	496	ANC attendees
12	Mabey(pp)(59)	2006	Tanzania (LI)	Cross-sectional	TPHA	Standard Bioline	Serum	60	6	23	493	ANC attendees
12	Mabey(qq)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	44	6	3	489	STI clinic attendees
12	Mabey(rr)(59)	2006	Brazil (UMI)	Cross-sectional	TPHA	Standard Bioline	Serum	45	5	3	489	STI clinic attendees
13	Miranda(80)	2009	Brazil (UMI)	Cross-Sectional	TPHA	Determine Abbott	Whole Blood	4	1	2	1373	Antenatal clinic attendants in Labor
14	Mishra(a)(81)	2010	Bangalore, India (LMI)	Cross-sectional	TPHA	Biorapid Syphilis	Whole Blood	114	128	4	1371	FSWs
14	Mishra(b)(81)	2010	Bangalore, India (LMI)	Cross-sectional	RPR + TPHA	Biorapid Syphilis	Whole Blood	85	35	33	1464	FSWs

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
14	Mishra(c)(81)	2010	Bangalore, India (LMI)	Cross-sectional	RPR (titre>/1:8) + TPHA	Biorapid Syphilis	Whole Blood	44	16	74	1483	FSWs
14	Mishra(d) 81)	2010	Bangalore, India (LMI)	Cross-sectional	TPHA	Biorapid Syphilis	Serum	193	49	19	1356	FSWs
14	Mishra(e)(81)	2010	Bangalore, India (LMI)	Cross-sectional	RPR + TPHA	Biorapid Syphilis	Serum	117	3	95	1402	FSWs
14	Mishra(f)(81)	2010	Bangalore, India (LMI)	Cross-sectional	RPR (titre>/1:8) + TPHA	Biorapid Syphilis	Serum	60	0	152	1405	FSWs
15	Montoya(a) (58)	2006	Sofala province, Mozambique (LI)	Cross-sectional	TPHA	Standard Bioline	Serum	488	44	25	3912	Pregnant Women
15	Montoya(b) (58)	2006	Sofala province, Mozambique (LI)	Cross-sectional	TPHA + RPR	Standard Bioline	Serum	367	15	146	3941	Pregnant Women
15	Montoya(c) (58)	2006	Sofala province, Mozambique (LI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	420	111	39	3892	Pregnant Women
15	Montoya(d) (58)	2006	Sofala province, Mozambique (LI)	Cross-sectional	TPHA + RPR	Standard Bioline	Whole Blood	327	54	132	3949	Pregnant Women
16	Nessa(a)(64)	2008	Mirpur, Dhaka, Bangladesh (LI)	Cross-sectional	RPR+TPHA	ICS-ACON	Whole Blood	138	4	43	499	FSW
16	Nessa(b)(64)	2008	Mirpur, Dhaka, Bangladesh (LI)	Cross-sectional	RPR+TPHA	ICS-ACON	Whole Blood	134	8	40	502	FSW
16	Nessa(c)(64)	2008	Mirpur, Dhaka, Bangladesh (LI)	Cross-sectional	RPR+TPHA	RTD-ACON	Whole Blood	135	7	38	504	FSW

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
16	Nessa(d)(64)	2008	Mirpur, Dhaka, Bangladesh (LI)	Cross-sectional	RPR+TPHA	RTD-ACON	Whole Blood	123	19	38	504	FSW
17	Nyamwamu(a) (67)	2009	Eldoret, Kenya (LI)	Cross-sectional	VDRL	Accurate Ultra Rapid	Serum	4	1	2	143	ANC attendees
17	Nyamwamu(b) (67)	2009	Eldoret, Kenya (LI)	Cross-sectional	TPHA	Accurate Ultra Rapid	Serum	6	0	0	144	ANC attendees
18	Oshiro(a)(67)	1999	Japan (HI)	Case-Control	Mediace TPLA	DainaScreen	Serum	67	0	0	69	Commercial panel and clinical Samples
18	Oshiro(b)(67)	1999	Japan (HI)	Case-Control	FTA-ABS	DainaScreen	Serum	34	0	0	66	Commercial panel and clinical Samples
18	Oshiro(c)(67)	1999	Japan (HI)	Case-Control	FTA-ABS	DainaScreen	Whole Blood	34	0	0	66	Commercial panel and clinical Samples
19	Qiaojia(a)(48)	1996	China (LMI)	Case-Control	RPR	Dot-Immunogold Filtration Assay (DIGFA)	Serum	50	0	0	300	Clinical Samples
19	Qiaojia(b)(48)	1996	China (LMI)	Case-Control	FTA-ABS	Dot-Immunogold Filtration Assay (DIGFA)	Serum	48	0	2	300	Clinical Samples
20	Rotanov(a)(50)	2008	Russia (UMI)	Case-Control	RPGA	Determine	serum	50	0	0	50	Archived serum panels

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
20	Rotanov(b)(50)	2008	Russia (UMI)	Case-Control	RPGA	Bioline Syphilis anti-TP Test Card	serum	48	1	2	49	Archived serum panels
20	Rotanov(c)(50)	2008	Russia (UMI)	Case-Control	RPGA	Treponema-Express	serum	50	5	0	45	Archived serum panels
21	Sano(47)	1999	Japan (HI)	Case-Control	TPPA	DainaScreen	Serum	0	0	3	997	Clinical Samples
22	Sato(96)	2003	Sao Paulo, SP, Brazil (UMI)	Case-Control	Symptomatic diagnosis	Determine	Serum	59	4	3	59	STD clinic attendees
23	Siedner(a)(54)	2004	San Francisco, USA (HI)	Crpss-Sectional	TP-PA	Determine	Whole Blood	52	0	0	47	City Clinic
23	Siedner(b)(54)	2004	San Francisco, USA (HI)	Cross-Sectional	TPPA	Determine	Whole Blood	60	8	0	59	City Clinic
23	Siedner(c)(54)	2004	San Francisco, USA (HI)	Cross-Sectional	TPPA	Phoenix Biotech Trep-Strip IV	Whole Blood	23	10	0	38	City Clinic
23	Siedner(d)(54)	2004	San Francisco, USA (HI)	Cross-Sectional	TPPA	Guardian One Step	Whole Blood	41	16	0	59	City Clinic
24	Tinajeros(29)	2006	Bolivia (LMI)	Cross-sectional	RPR + TPPA	Determine	Whole Blood	314	28	128	8422	Pregnant Women

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
25	van Dommelen(51)	2008	Maastricht, the Netherlands (HI)	Case-Control	TPPA	Biorapid Syphilis	Serum	133	12	39	146	University Hospital
26	Villazon-Vargas(65)	2009	Cochabamba, Bolivia (LMI)	Cross-sectional	FTA-ABS	Determine	Whole Blood	50	1	1	437	Pregnant Women
27	Wang(a)(60)	2007	China (LMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	82	30	1	388	STD clinic attendees
27	Wang(b)(60)	2007	China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	75	41	1	352	STD clinic attendees
27	Wang(c)(60)	2007	China (LMI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	102	14	2	351	STD clinic attendees
27	Wang(d)(60)	2007	China (LMI)	Cross-sectional	TPHA	Determine	Whole Blood	83	29	2	387	STD clinic attendees
27	Wang(e)(60)	2007	China (LMI)	Cross-sectional	TPHA	Omega Visitect	Whole Blood	84	28	0	389	STD clinic attendees
27	Wang(f)(60)	2007	China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Whole Blood	82	34	1	352	STD clinic attendees
27	Wang(g)(60)	2007	China (LMI)	Cross-sectional	TPHA	Standard Bioline	Whole Blood	102	14	2	351	STD clinic attendees
27	Wang(h)(60)	2007	China (LMI)	Cross-sectional	TPHA	Determine	Whole Blood	86	26	0	389	STD clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 3: Characteristics and results of studies included in Part I. (Cont'd)

Study ID	Author	Year	Location	Study Design	Reference Standard	Index Test	Sample	TP	FN	FP	TN	Population
27	Wang(i)(60)	2007	China (LMI)	Cross-sectional	TPHA	Omega Visitect	Serum	106	6	7	382	STD clinic attendees
27	Wang(j)(60)	2007	China (LMI)	Cross-sectional	TPHA	Qualpro Syphicheck	Serum	79	37	4	349	STD clinic attendees
27	Wang(k)(60)	2007	China (LMI)	Cross-sectional	TPHA	Standard Bioline	Serum	111	5	7	346	STD clinic attendees
27	Wang(l)(60)	2007	China (LMI)	Cross-sectional	TPHA	Determine	Serum	112	0	4	385	STD clinic attendees
28	West(53)	2002	Farafenni area of Gambia (LI)	Cross-sectional	TPHA	RST, Quorum Diagnostics, Vancouver, BC	Serum	51	29	41	1204	Women of reproductive age
28	West(53)	2002	Farafenni area of Gambia (LI)	Cross-sectional	RPR + TPHA	RST, Quorum Diagnostics, Vancouver, BC	Serum	30	10	62	1223	Women of reproductive age
29	Yang(97)	2010	China (LMI)	Case-Control	TPPA titer of 1:80 or above FTA-Abs	Quantum Dots-Based	Serum	50	0	0	50	lab and healthy blood donors
29	Yang(97)	2010	China (LMI)	Case-Control	TPPA titer of 1:80 or above FTA-Abs	colloidal gold-based lateral flow test	Serum	41	9	0	50	lab and healthy blood donors
30	Zarakolu(49)	2002	Turkey (UMI)	Case-Control	FTA-ABS	ICS	Serum	13	0	0	124	STD clinic attendees

*HI: High Income, UMI: Upper-Middle Income, LMI: Lower-Middle Income, LI: Low Income.

Table 4: QUADAS checklist results

Item #	Item	Description	Yes	No	Unclear
1	Adequate Spectrum composition	Was the spectrum of patients representative of the patients who will receive the test in practice?	60% (n=18)	20% (n=6)	20% (n=6)
2	Clear description of selection criteria	Were selection criteria clearly described?	87% (n=26)	13% (n=4)	0% (n=0)
3	Adequate reference standard	Is the reference standard likely to correctly classify the target condition?	97% (n=29)	3% (n=1)	0% (n=0)
4	Absence of disease progression bias	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	97% (n=29)	0% (n=0)	3% (n=1)
5	Absence of partial verification	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	93% (n=28)	3% (n=1)	3% (n=1)
6	Absence of differential verification bias	Did patients receive the same reference standard regardless of the index test result?	90% (n=27)	7% (n=2)	3% (n=1)
7	Absence of incorporation bias	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	100% (n=30)	0% (n=0)	0% (n=0)
8	Adequate description of the index test execution	Was the execution of the index test described in sufficient detail to permit replication of the test?	100% (n=30)	0% (n=0)	0% (n=0)
9	Adequate description of the reference test execution	Was the execution of the reference standard described in sufficient detail to permit its replication?	100% (n=30)	0% (n=0)	0% (n=0)
10	Absence of index test review bias	Were the index test results interpreted without knowledge of the results of the reference standard?	37% (n=11)	0% (n=0)	63% (n=19)
11	Absence of reference test review bias	Were the reference standard results interpreted without knowledge of the results of the index test?	20% (n=6)	0% (n=0)	80% (n=24)
12	Absence of clinical review bias	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	43% (n=13)	0% (n=0)	57% (n=17)
13	Report of uninterpretable results	Were uninterpretable/ intermediate test results reported?	33% (n=10)	57% (n=17)	10% (n=3)
14	Description of withdrawals	Were withdrawals from the study explained?	17% (n=5)	73% (n=22)	10% (n=3)

Table 5: STARD checklist results

Section and Topic	Item #	Item	R	NR
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	70% (n=21)	30% (n=9)
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	67% (n=20)	33% (n=10)
METHODS				
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	93% (n=28)	7% (n=2)
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	97% (n=29)	3% (n=1)
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	50% (n=15)	50% (n=15)
Test methods	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	47% (n=14)	53% (n=16)
	7	The reference standard and its rationale.	100% (n=30)	0% (n=0)
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	100% (n=30)	0% (n=0)
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	100% (n=30)	0% (n=0)
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	60% (n=18)	40% (n=12)
Statistical methods	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	23% (n=7)	77% (n=23)
	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	63% (n=19)	37% (n=11)
	13	Methods for calculating test reproducibility, if done.	13% (n=4)	87% (n=26)

Table 5: STARD checklist results (Cont'd)

Section and Topic	Item #	Item	R	NR
RESULTS				
Participants	14	When study was performed, including beginning and end dates of recruitment.	67% (n=20)	33% (n=10)
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	67% (n=20)	33% (n=10)
Test results	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	80% (n=24)	20% (n=6)
	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	50% (n=15)	50% (n=15)
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	67% (n=20)	33% (n=10)
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	93% (n=28)	7% (n=2)
Estimates	20	Any adverse events from performing the index tests or the reference standard.	3% (n=1)	97% (n=29)
	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	90% (n=27)	10% (n=3)
	22	How indeterminate results, missing data and outliers of the index tests were handled.	37% (n=11)	63% (n=19)
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	33% (n=10)	67% (n=20)
DISCUSSION	24	Estimates of test reproducibility, if done.	20% (n=6)	80% (n=24)
	25	Discuss the clinical applicability of the study findings.	100% (n=30)	0% (n=0)

Table 6: Conflicts of Interest

Study ID	Author	Address conflict of interest (COI)?	If addressed, did the authors have COI?
1	Benzaken (61)	No	NA
2	Benzaken (62)	Yes	No
3	Bronzan (63)	No	NA
4	Campos (57)	Yes	No
5	Castro (27)	Yes	Yes
6	Diaz (96)	No	NA
7	Gianino (88)	No	NA
8	Hernandez-Trejo (56)	Yes	No
9	Juarez-Figueroa (55)	No	NA
10	Li (66)	Yes	No
11	Lien (52)	Yes	No
12	Mabey (59)	Yes	No
13	Mirand(81)	No	NA
14	Mishra (82)	Yes	No
15	Montoya (58)	Yes	No
16	Nessa (64)	No	NA
17	Nyamwamu (67)	No	NA
18	Oshiro (46)	No	NA
19	Qiaojia (48)	No	NA
20	Rotanov (50)	No	NA
21	Sano (47)	No	NA
22	Sato (97)	No	NA
23	Siedner (54)	No	NA
24	Tinajeros (29)	Yes	No
25	van Dommelen (51)	Yes	No
26	Villazon-Vargas (65)	No	NA
27	Wang (60)	No	NA
28	West (53)	No	NA
29	Yang (73)	No	NA
30	Zarakolu (49)	No	NA

Table 7: Results of pooled sensitivity and specificity, before and after adjustment for imperfect reference standard

POC Test	Sample	Reference Standard	Pooled Parameters	Assuming Perfect Reference Standard	Assuming Imperfect Reference Standard	Manufacturer Reported Parameter
Determine	Serum	TP Specific (n=10)	Sensitivity	98.81% (96.52 , 99.98)	99.17% (96.56, 100)	100%(96)
			Specificity	97.94% (96.30, 98.48)	99.28% (98.15, 100)	100%(96)
	Whole Blood	TP Specific (n=14)	Sensitivity	85.55% (76.15, 94.49)	89.49% (79.88, 98.15)	92.30%(96)
			Specificity	99.50% (98.95, 99.93)	99.91% (99.44, 1)	100%(96)
		TP & non-TP Specific (n=8)	Sensitivity	83.60 (64.85, 97.81)	90.65% (68.68, 99.99)	----------
			Specificity	98.59% (94.91, 100)	98.93% (96.49, 99.97)	----------
Bioline	Serum	TP Specific (n=8)	Sensitivity	93.99 (88.79, 98.45)	99.67% (97.65, 100)	99.3%(97)
			Specificity	98.58 (97.54, 99.31)	99.56% (98.55, 100)	99.5%(97)
	Whole Blood	TP Specific (n=13)	Sensitivity	87.70% (84.78, 90.58)	91.47% (87.06, 96.12)	----------
			Specificity	99.07% (98.50 , 99.59)	99.61% (99.04, 100)	----------
Syphicheck	Serum	TP Specific (n=7)	Sensitivity	81.91% (69.10, 93.19)	88.46% (73.54, 99.87)	100%(98)
			Specificity	99.05% (98.35, 99.62)	99.98% (99.64, 100)	----------
	Whole Blood	TP Specific (n=12)	Sensitivity	76.88 %(68.97, 84.71)	81.99% (71.84, 91.99)	----------
			Specificity	99.41%(99.05, 99.73)	99.81% (99.46, 100)	----------
Visitect	Serum	TP Specific (n=6)	Sensitivity	93.96% (89.24, 98.01)	98.18% (93.53, 100)	----------
			Specificity	98.51% (97.70, 99.19)	99.89% (99.19, 100)	----------
	Whole Blood	TP Specific (n=12)	Sensitivity	78.05% (70.34, 85.02)	82.93% (94.50, 100)	----------
			Specificity	99.52% (99.16, 99.82)	99.87% (99.58, 100)	----------

Table 8: Characteristics of included studies for non-accuracy outcomes

Author	Year	Country (Economic Classification)	Language	Population	Sample	Index Test	Study Design
Amadi (39)	2010	Nigeria(LMI)	English	Dental clinic patients	Plasma	Syphilis Ultra Rapid Test Strip	Cross-Sectional
Benzaken (27)	2007	Brazi(UMI)	English	Men and Women with no history	Whole Blood Serum	SD Bioline Syphilis Syphicheck-WB VisiTect Syphlis Determine	Cross-Sectional
Benzaken (12)	2008	Brazil(UMI)	English	Outreach clinic for high risk populations	Whole Blood	Visitect	Cross-Sectional
Bronazn (22)	2007	South Africa(UMI)	English	ANC attendees	Whole Blood	Determine	Cross-Sectional
Campos (36)	2006	Peru(UMI)	English	FSWs in commercial sex venues	Whole Blood	Determine	Cross-Sectional
Garcia (20)	2006	Bolivia(LMI)	English	ANC attendees	Whole Blood	Determine	Cross-Sectional
Herring	2006	South Africa(UMI) The Gambia(LI Tanzania(LI) China(LMI) Sri Lanka(LMI) Haiti(LI) USA(HI) Russian Federation(UMI)	English	Evaluation panel from archived specimen	Serum	Determine Syphilis Fast Espline TP Syphicheck WB SD Bioline Visitect Syphilis	Case-Control
Hurtado (41)	2009	Spain(HI)	Spanish	MSM with high risk behaviour	Whole Blood	Determine	Cross-Sectional
Juarez-Figueroa (30)	2007	Mexico(UMI)	English	FSWs ANC attendees	Whole Blood Serum	Determine	Cross-Sectional

*HI=High Income, UMI=Upper Middle Income, LM=Low-Middle Income, LI=Low Income as defined by World Bank (97)

Table 8: Characteristics of included studies for non-accuracy outcomes (Cont'd)

Author	Year	Country (Economic Classification)	Language	Population	Sample	Index Test	Study Design
Lahuerta (34)	2010	Guatemala(LMI)	English	People not at risk, MSM, Transgender	Whole Blood	Determine	Cross-Sectional
Lee (23)	2010	Australia(HI)	English	MSM	Whole Blood	Determine	Cross-Sectional
Miranda (35)	2009	Brazil(UMI)	Portuguese	ANC attendees in labor	Whole Blood	Determine	Cross-Sectional
Mishra (37)	2010	India(LMI)	English	FSWs attending STI clinic	Whole Blood Serum	Qualpro Syphicheck	Cross-Sectional
Munkhuu (38)	2009	Mongolia(LMI)	English	ANC attendees	Whole Blood	SD Bioline Syphilis	Cluster Randomized Trial
Munkhuu (24)	2009	Mongolia(LMI)	English	ANC attendees	Whole Blood	SD Bioline Syphilis	Cross-Sectional
Revollo (40)	2007	Bolivia(LMI)	Spanish	Postnatal women in hospital	Serum	Determine	Cross-Sectional

*HI=High Income, UMI=Upper Middle Income, LM=Low-Middle Income, LI=Low Income as defined by World Bank (97)

Table 8: Characteristics of included studies for non-accuracy outcomes (Cont'd)

Author	Year	Country (Economic Classification)	Language	Population	Sample	Index Test	Study Design
Rotanov (26)	2008	Russian Federation (UMI)	Russian	Known and unknown cases	Serum	Determine Abbott Expline TP Syphicheck-WB SD Bio-Line Syphilis Visi Test Syhilis Syphilis Rapid Screening Test Bioline Syphilis anti-TP Test Card Syphicheck WB Rapid Test for Syphilis Smart Strip SyphilisWB Serum Treponema-Express	Case-Control
Sabido (21)	2009	Brazil(UMI)	English	High risk populations	Whole Blood	Visitect Syphilis test	Cross-Sectional
Seguy (25)	2008	Guyana(LMI)	English	Miners	Whole Blood	Determine	Cross-Sectional

*HI=High Income, UMI=Upper Middle Income, LM=Low-Middle Income, LI=Low Income as defined by World Bank (97)

Figures:

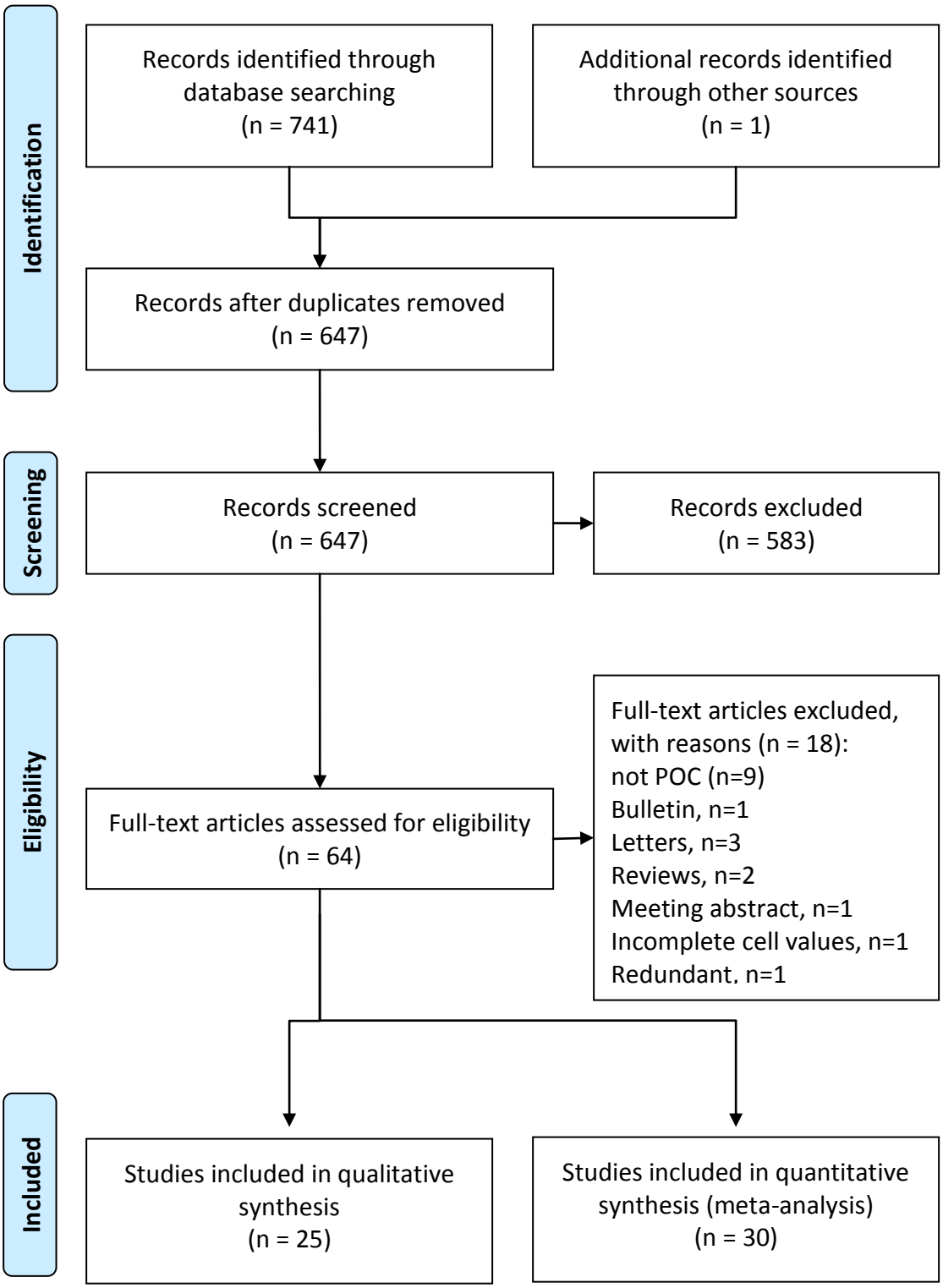

Figure 6: Study selection

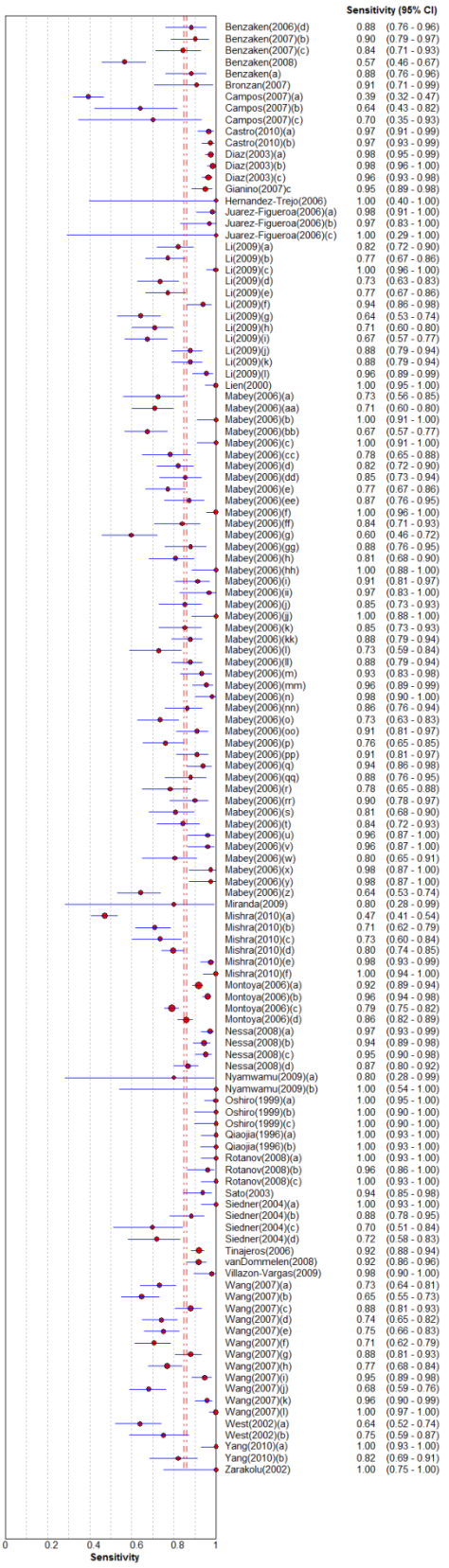

Figure 7: Forest plot of sensitivity of studies included in Part I (n=127)

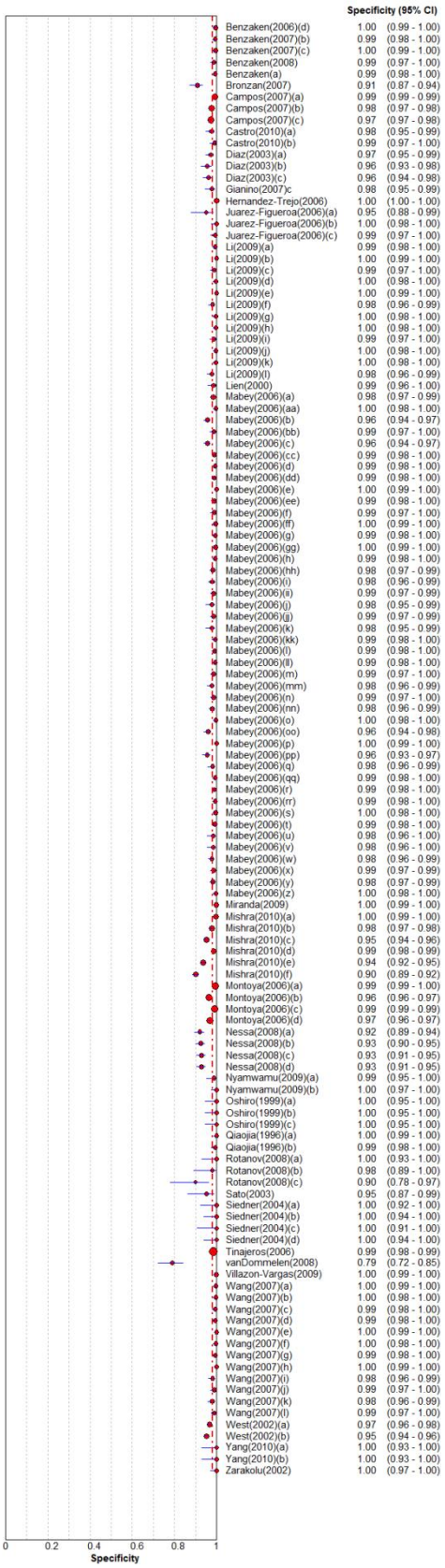

Figure 8: Forest plot of specificity of studies included in Part I (n=127)

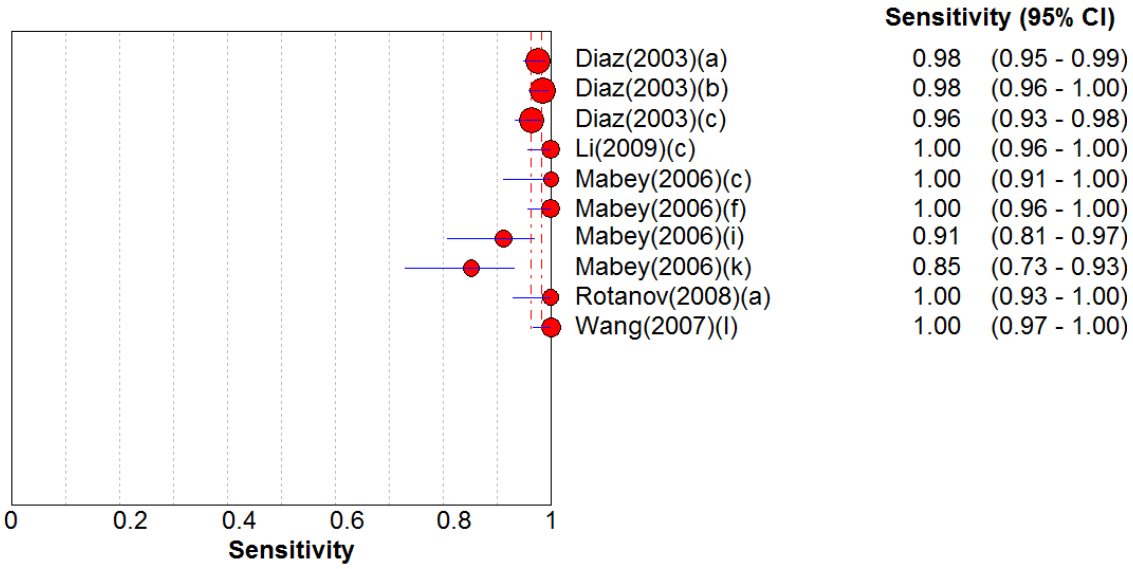

Figure 9: Forest plot of sensitivity for subgroup: Determine, Serum, TP specific reference standard (n=10)

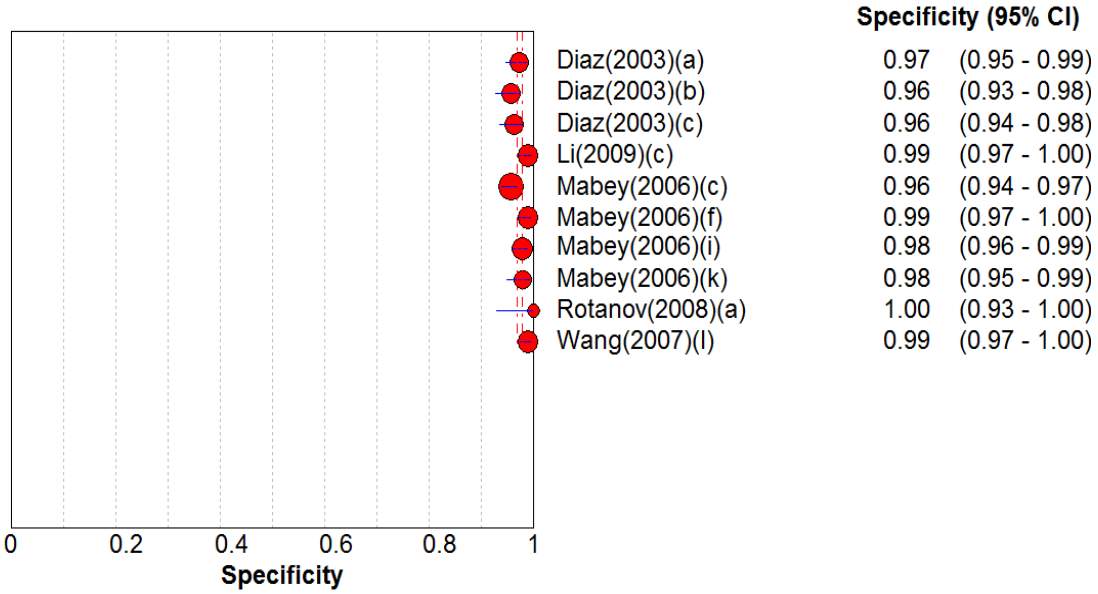

Figure 10: Forest plot of specificity for subgroup: Determine, Serum, TP specific reference standard (n=10)

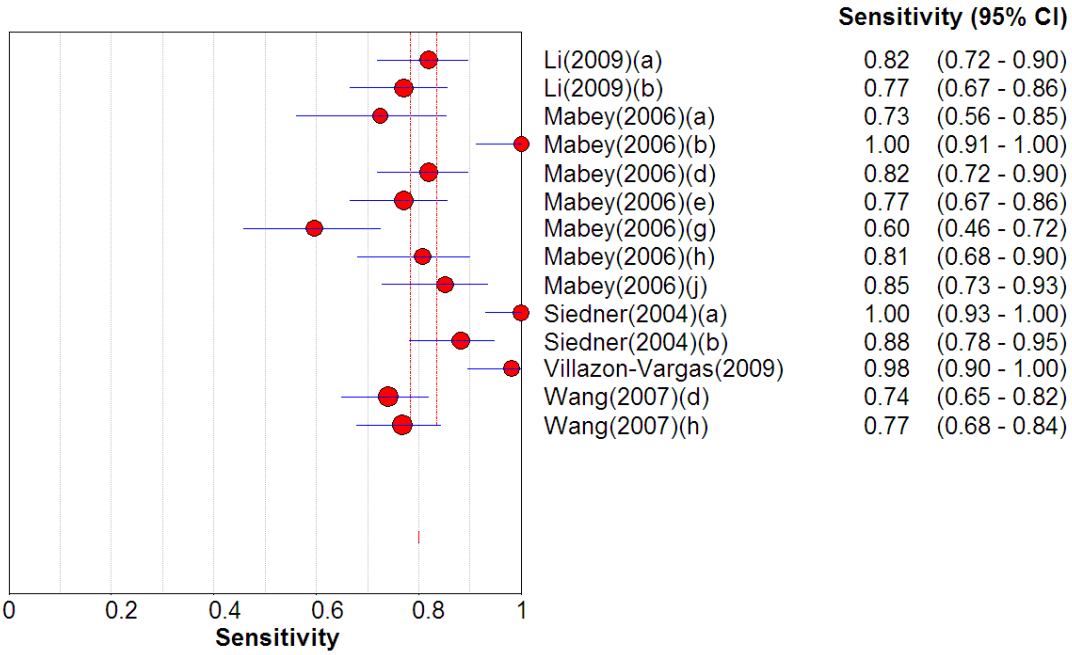

Figure 11: Forest plot of sensitivity for subgroup: Determine, Whole Blood, TP specific reference standard (n=14)

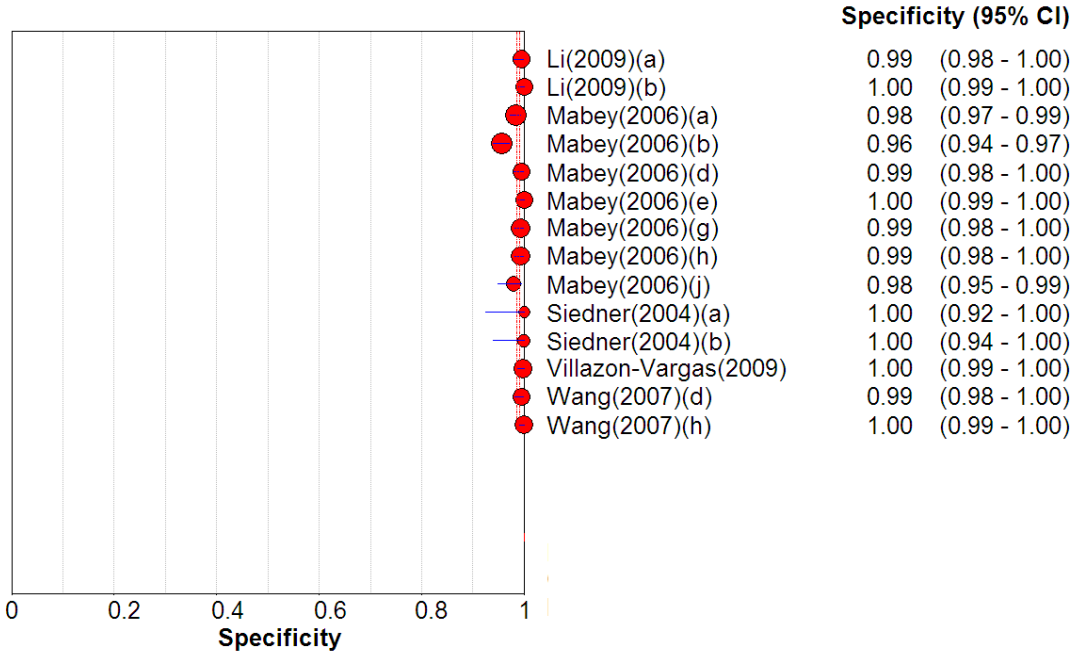

Figure 12: Forest plot of specificity for subgroup: Determine, Whole Blood, TP specific reference standard (n=14)

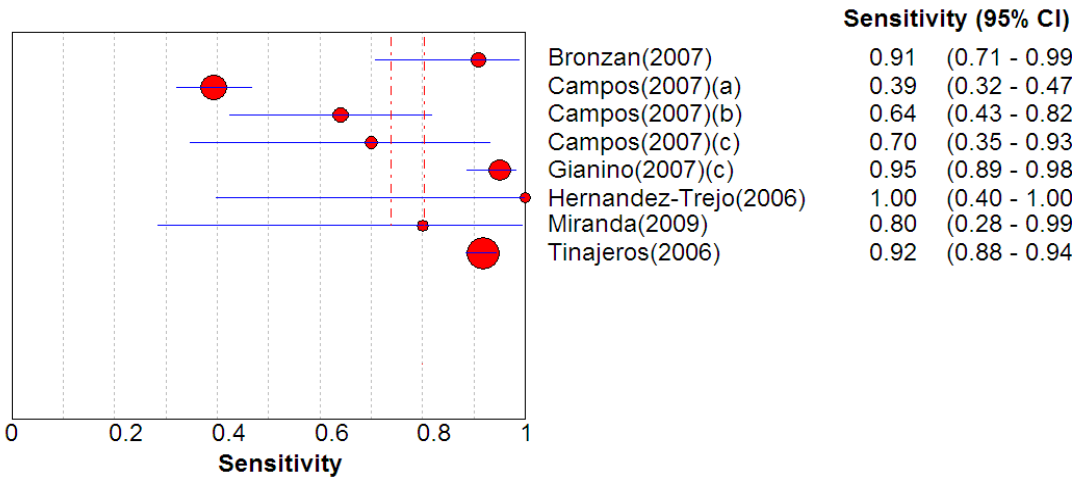

Figure 13: Forest plot of sensitivity for subgroup: Determine, Whole Blood, TP and non-TP specific reference standard (n=8)

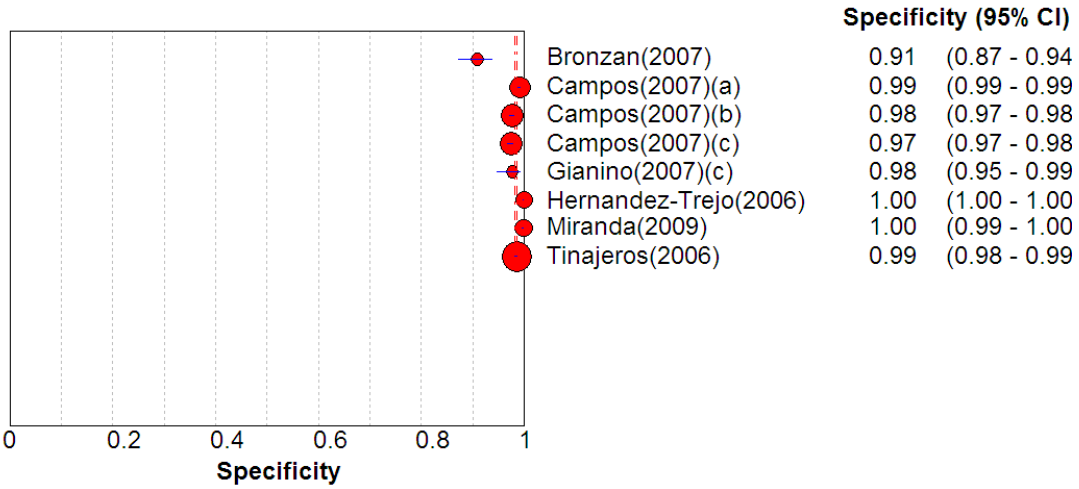

Figure 14: Forest plot of specificity for subgroup: Determine, Whole Blood, TP and non-TP specific reference standard (n=8)

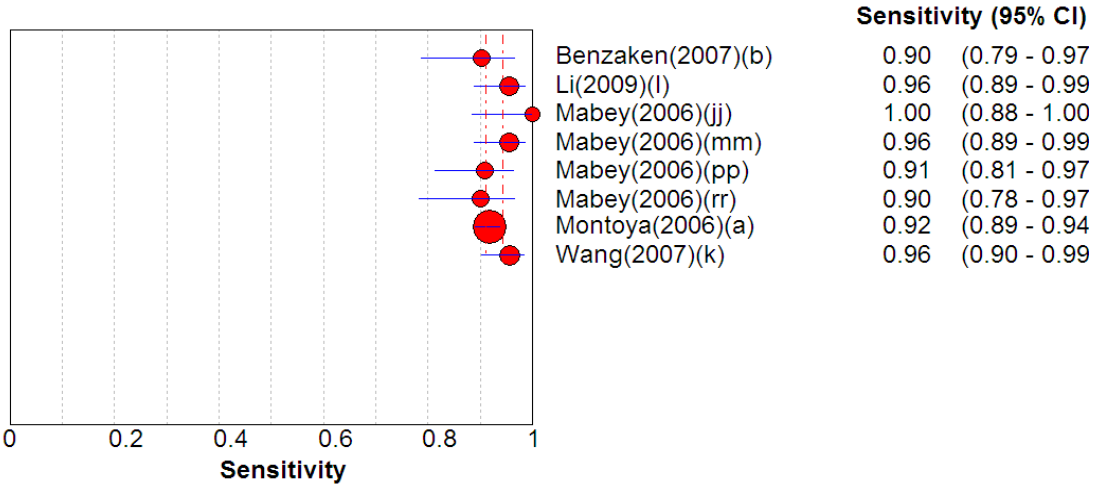

Figure 15: Forest plot of sensitivity for subgroup: Bioline, Serum, TP specific reference standard (n=8)

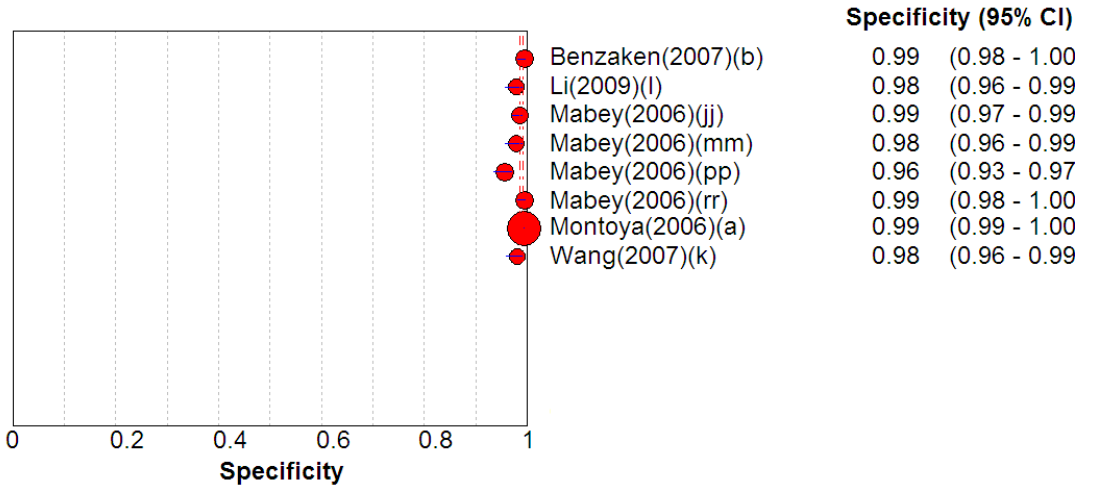

Figure 16: Forest plot of specificity for subgroup: Bioline, Serum, TP specific reference standard (n=8)

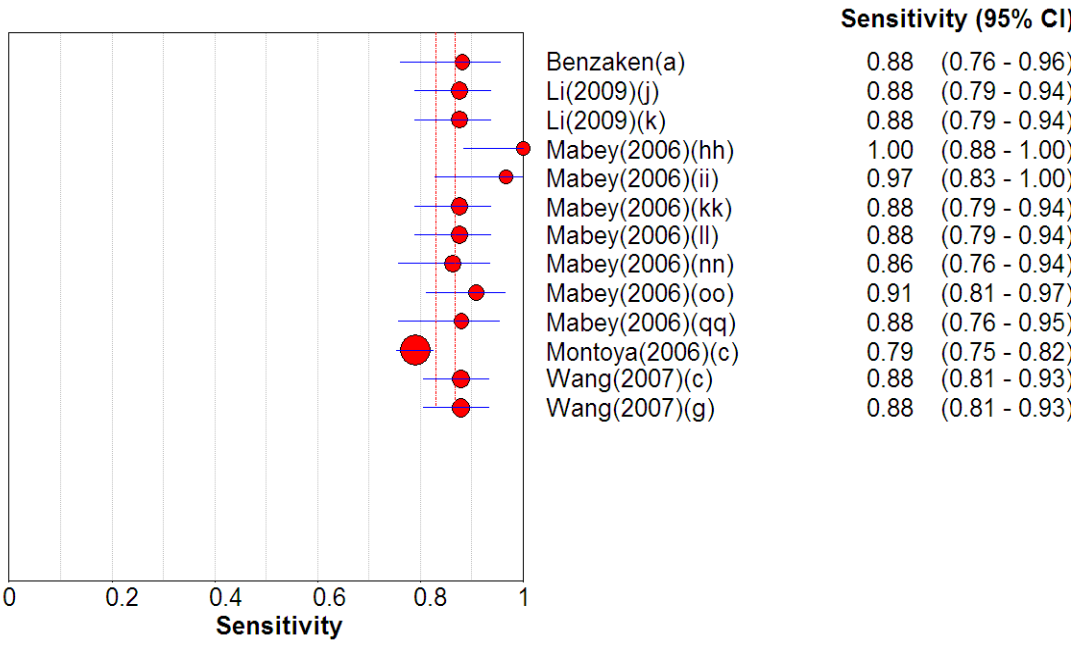

Figure 17: Forest plot of sensitivity for subgroup: Bioline, Whole Blood, TP specific reference standard (n=13)

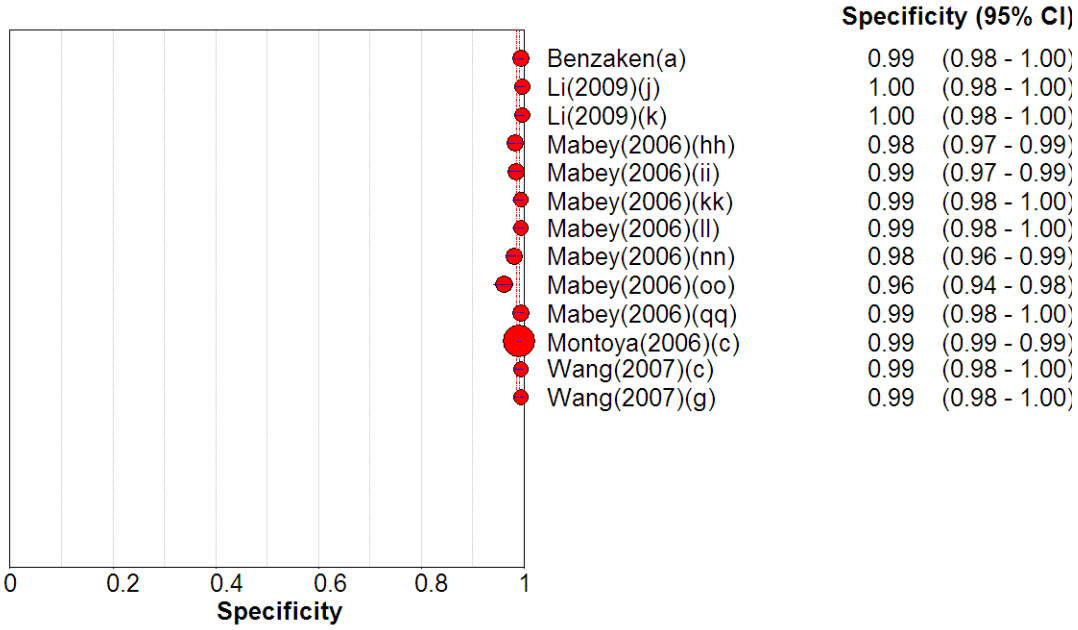

Figure 18: Forest plot of specificity for subgroup: Bioline, Whole Blood, TP specific reference standard (n=13)

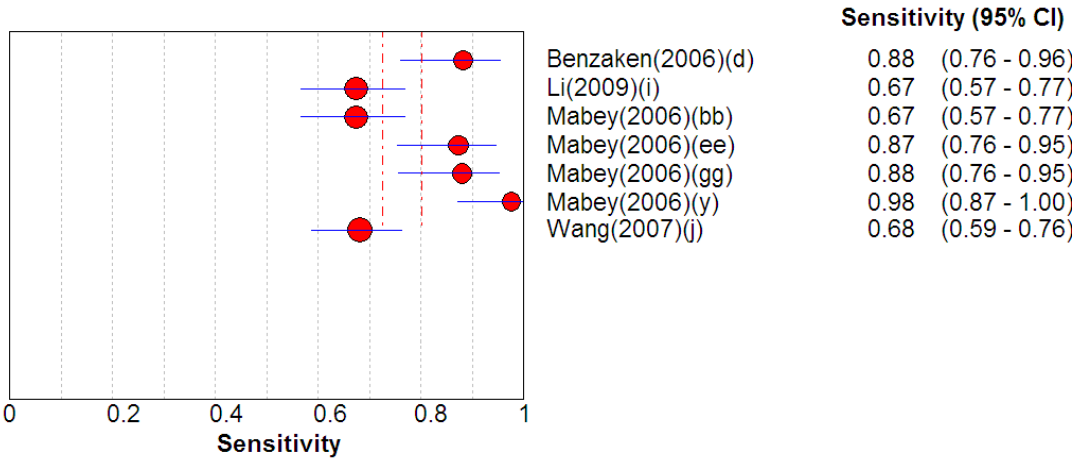

Figure 19: Forest plot of sensitivity for subgroup: Syphicheck, Serum, TP specific reference standard (n=7)

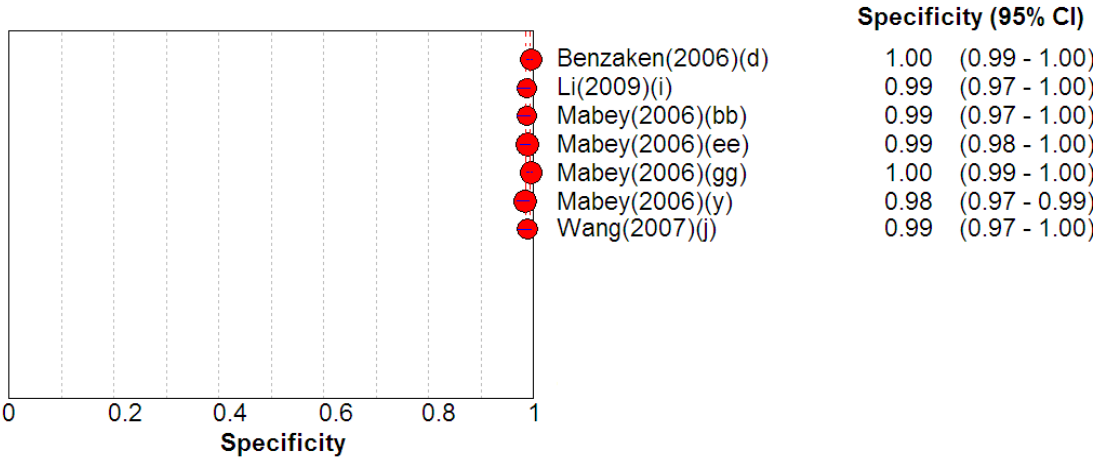

Figure 20: Forest plot of specificity for subgroup: Syphicheck, Serum, TP specific reference standard (n=7)

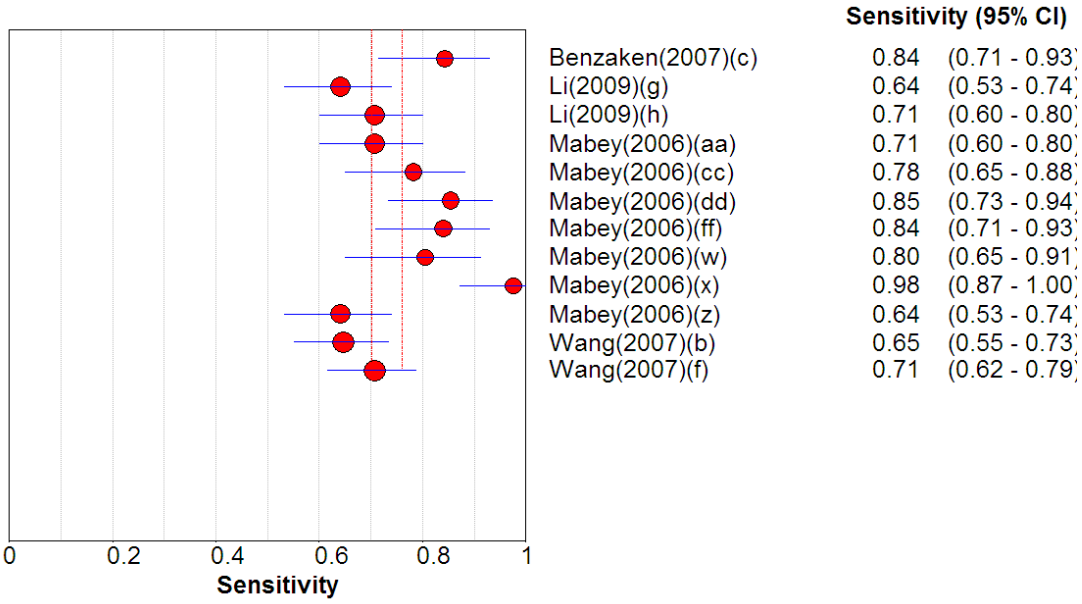

Figure 21: Forest plot of sensitivity for subgroup: Syphicheck, Whole Blood, TP specific reference standard (n=12)

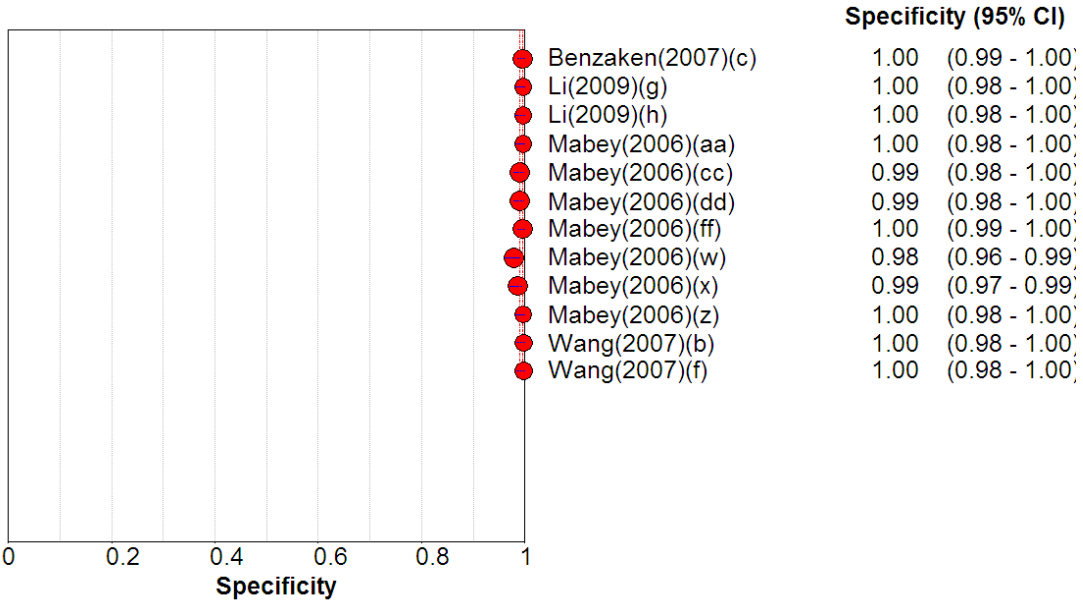

Figure 22: Forest plot of specificity for subgroup: Syphicheck, Whole Blood, TP specific reference standard (n=12)

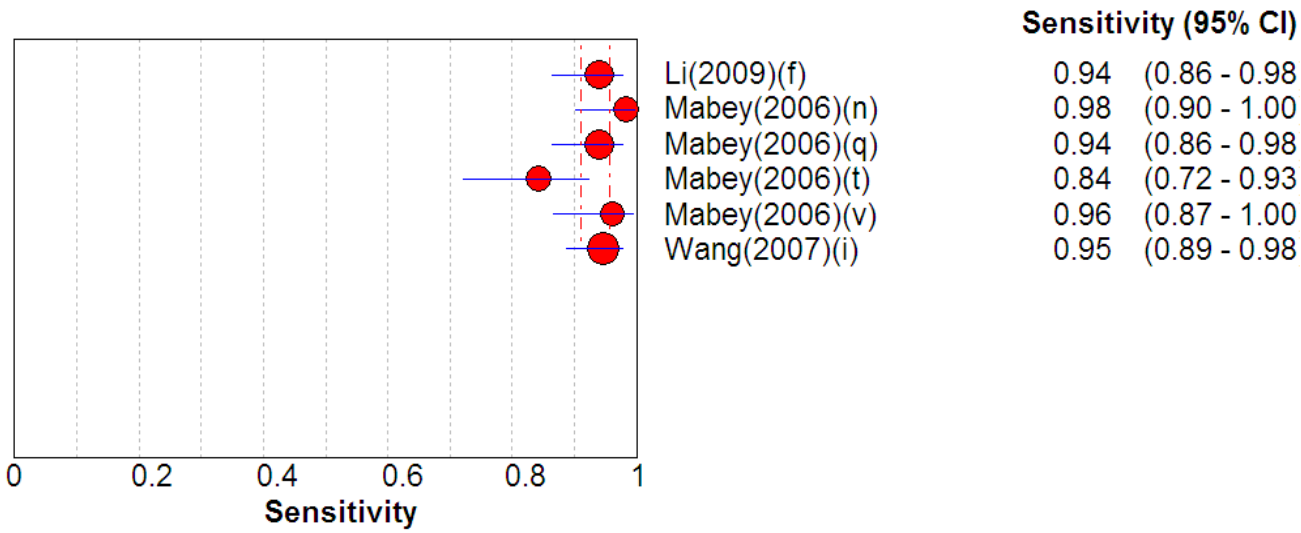

Figure 23: Forest plot of sensitivity for subgroup: Visitect, Serum, TP specific reference standard (n=6)

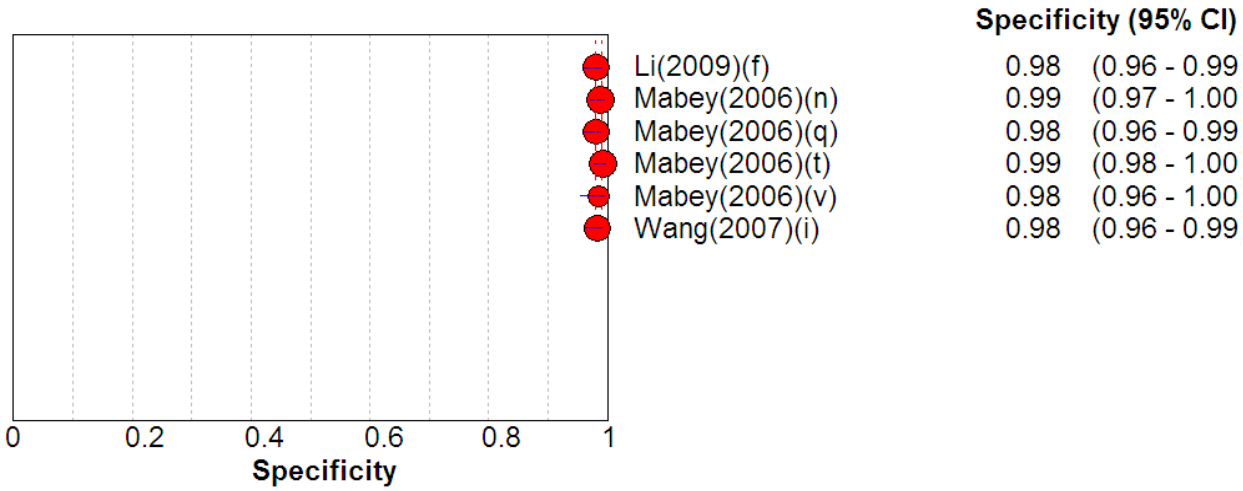

Figure 24: Forest plot of specificity for subgroup: Visitect, Serum, TP specific reference standard (n=6)

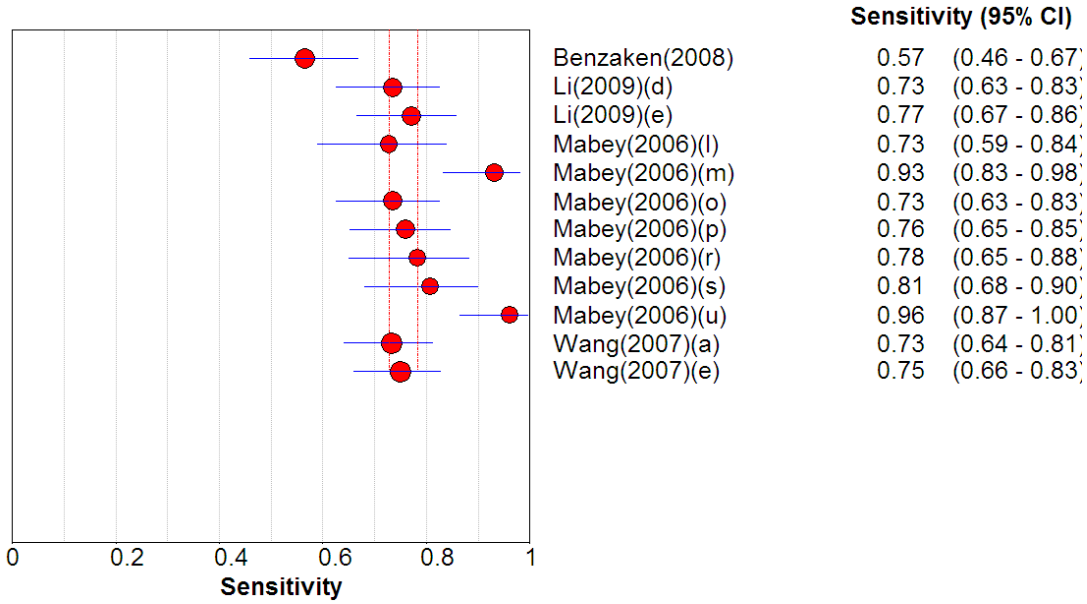

Figure 25: Forest plot of sensitivity for subgroup: Visitect, Whole Blood, TP specific reference standard (n=12)

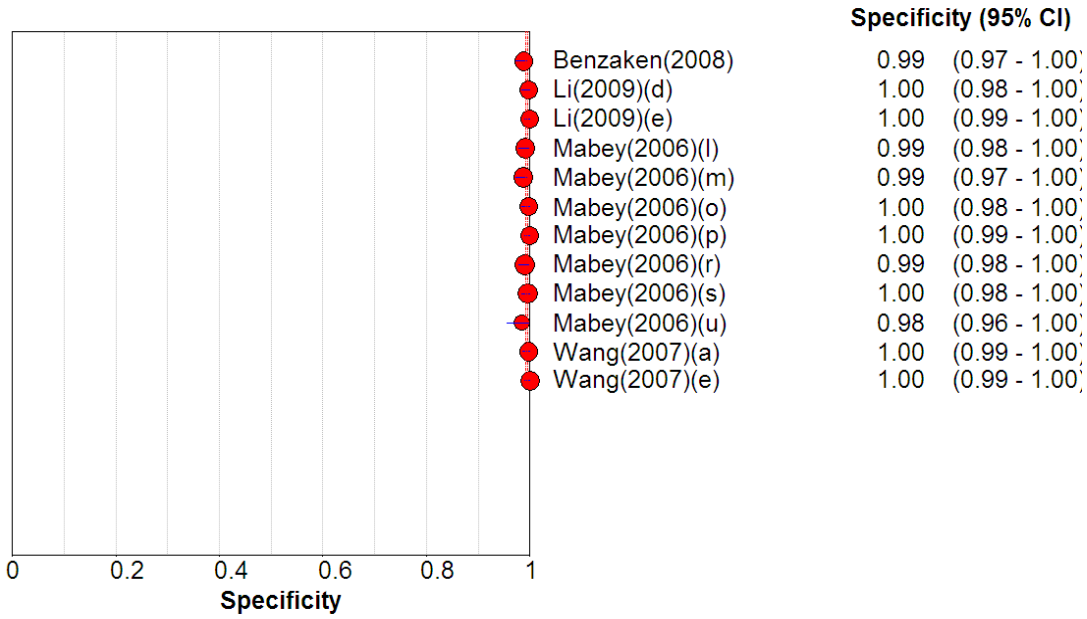

Figure 26: Forest plot of specificity for subgroup: Visitect, Whole Blood, TP specific reference standard (n=12)