PARTIAL VERIFICATION BIAS CORRECTION IN DIAGNOSTIC ACCURACY STUDIES USING PROPENSITY SCORE-BASED METHODS

WAN NOR ARIFIN BIN WAN MANSOR

UNIVERSITI SAINS MALAYSIA

2023

PARTIAL VERIFICATION BIAS CORRECTION IN DIAGNOSTIC ACCURACY STUDIES USING PROPENSITY SCORE-BASED METHODS

by

WAN NOR ARIFIN BIN WAN MANSOR

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

June 2023

ACKNOWLEDGEMENT

"If all the trees on earth were pens and the ocean (were ink), refilled by seven other oceans, the Words of Allah would not be exhausted. Surely Allah is Almighty, All-Wise." (Al-Qur'an Al-Kareem, Surah Luqman: 27)

Praise be to Allah, The All Knowing, The Omniscient.

I would like to thank my supervisor, Associate Professor Dr. Umi Kalsom Yusof, a supportive and dedicated supervisor and mentor of mine from the School of Computer Sciences, Universiti Sains Malaysia, for her guidance throughout the PhD journey.

I also thank my colleagues for sharing their knowledge, and for their feedbacks and support during the academic journey that we share together.

I am also thankful to the School of Computer Sciences, Universiti Sains Malaysia for the great learning and research environment; to Universiti Sains Malaysia for the paid study leave; and to the Ministry of Higher Education for the scholarship provided.

I would like to extend countless gratitude to my wife Wan Suryana for her patience, support and prayers, and for being a very good cook, thank you for the food; to my four children, Wan Nur Bahiyyah, Wan Muhammad Arsyad, Wan Muhammad Aqil, and Wan Muhammad Afiq for the time we spent together at home during the challenging COVID-19 pandemic; to my father for buying me books when I was young, thereby planting my love for knowledge; and to my mother for her continued support and prayers. Thank you for supporting my academic journey all this time.

Their contributions to this thesis are invaluable, may Allah Subhanahu Wa Ta'ala bless them all.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ALGORITHMS

LIST OF SYMBOLS

b	Number of bootstrap samples
B	Number of simulation runs
D	Disease status
m	Number of imputations
n	Sample size
p	Prevalence/proportion of disease
ϕ	Phi correlation coefficient
T	Test result
V	Verified/Verification status
$\mathbf{X}$	Covariate(s)
X	An auxiliary variable X with strong correlation to disease status
Y	An auxiliary variable Y with perfect correlation to disease status
Z	An auxiliary variable Z with weak correlation to disease status

LIST OF ABBREVIATIONS

AUC	Area under receiver operating characteristic curve
BG	Begg and Greenes' method
CAD	Coronary artery disease
CCA	Complete case analysis
DVT	Deep vein thrombosis
EBG	Extended Begg and Greenes' method
ECG	Electrocardiogram
EIB	Exercise-induced bronchoconstriction
EM	Expectation-maximization
FDA	Full data analysis
FeNO	fractional exhaled nitric oxide
FN	False negative
FP	False positive
IPB	Inverse probability bootstrap
IPTW	Inverse probability treatment weight
IPW	Inverse probability weight
IPW-FactLogReg-MNAR	Inverse Probability Weighted Logistic Regression by Factorization for MNAR
IPW-LogReg	Inverse Probability Weighted Logistic Regression

IPW-LogReg-MNAR	Inverse Probability Weighted Logistic Regression for MNAR
IPWE	Inverse probability weighting estimator
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MI	Multiple imputation
ML	Machine learning
MLE	Maximum likelihood estimate
MNAR	Missing not at random
MSE	Mean squared error
MSI	Mean score imputation
NI	Non-ignorable
NN	Neural networks
NPV	Negative predictive value
PPV	Positive predictive value
PS	Propensity score
PSS	Propensity score stratification
PVB	Partial verification bias
ROC	Receiver operating characteristic
RT-PCR	Reverse transcription polymerase chain reaction

RTK-Ag	Rapid test kit by antigen
SE	Standard error
SIPW	Scaled inverse probability weight
SIPW-BalResamp	Scaled Inverse Probability Weighted Balanced Resampling
SIPW-BalResamp-MNAR	Scaled Inverse Probability Weighted Balanced Resampling for MNAR
SIPW-FactBalResamp-MNAR	Scaled Inverse Probability Weighted Balanced Resampling by Factorization for MNAR
SIPW-FactResamp-MNAR	Scaled Inverse Probability Weighted Resampling by Factorization for MNAR
SIPW-Resamp	Scaled Inverse Probability Weighted Resampling
SIPW-Resamp-MNAR	Scaled Inverse Probability Weighted Resampling for MNAR
Sn	Sensitivity
Sp	Specificity
SPE	Semi-parametric efficient
SPECT	Single-photon-emission computed tomography
TN	True negative
TP	True positive

LIST OF APPENDICES

PEMBETULAN BIAS PENGESAHAN SEPARA DALAM KAJIAN KETEPATAN DIAGNOSTIK MENGGUNAKAN KAEDAH BERASASKAN SKOR KECENDERUNGAN

ABSTRAK

Ujian diagnostik baharu dinilai berbanding ujian piawaian emas dalam kajian ketepatan diagnostik. Untuk ujian diagnostik binari, prestasi dikira dengan ukuran ketepatan, yang terpenting ialah kepekaan (Sn) dan kekhususan (Sp). Ukuran ini selalunya berat sebelah disebabkan oleh pengesahan terpilih pesakit, yang dikenali sebagai bias pengesahan separa (PVB). Kaedah pembetulan PVB sedia ada berbeza dalam keupayaan dan pendekatan untuk mengendalikan andaian hilang secara rawak (MAR) atau hilang bukan secara rawak (MNAR). Daripada kaedah tersebut, kaedah sedia ada yang menggunakan skor kecenderungan (PS) menunjukkan penggunaan pemberat yang terhad dan hanya mengendalikan andaian MAR. Kaedah pembetulan berasaskan PS mempunyai potensi untuk penambahbaikan dan lanjutan di bawah andaian MAR dan MNAR. Objektif kajian ini adalah untuk mereka bentuk kaedah berasaskan PS dengan pendekatan regresi berwajaran dan pensampelan semula untuk menambah baik dan meluaskan metodologi pembetulan PVB di bawah andaian MAR dan MNAR. Tiga kaedah pembetulan PVB berasaskan PS MAR telah dicadangkan: 1) Regresi logistik berwajaran kebarangkalian songsang (IPW-LogReg), 2) Persampelan semula berwajaran kebarangkalian songsang berskala (SIPW-Resamp) dan 3) Persampelan semula seimbang berwajaran kebarangkalian songsang berskala (SIPW-BalResamp). Kaedah ini telah diperluaskan kepada MNAR dengan mencadangkan dua kaedah anggaran PS: 1) PS dengan pembolehubah bantu dan 2) pemfaktoran PS. Ini menghasilkan enam kaedah susulan MNAR berdasarkan kaedah tersebut: 1)

Regresi logistik berwajaran kebarangkalian songsang untuk MNAR (IPW-LogReg-MNAR) 2) Kebarangkalian songsang berwajaran oleh regresi logistik pemfaktoran untuk MNAR (IPW-FactLogReg-MNAR), 3) Pensampelan semula wajaran kebarangkalian songsang berskala untuk MNAR (SIPW-Resamp-MNAR), 4) Pensampelan semula wajaran kebarangkalian songsang berskala mengikut pemfaktoran untuk MNAR (SIPW-FactResamp-MNAR), 5) Pensampelan semula seimbang berwajaran songsang berskala untuk MNAR (SIPW-BalResamp-MNAR) dan 6) Kebarangkalian songsang berskala berwajaran persampelan semula seimbang dengan pemfaktoran untuk MNAR (SIPW-FactBalResamp-MNAR). Set data simulasi dijana untuk kombinasi prevalens penyakit, Sn, Sp, kebarangkalian pengesahan dan saiz sampel untuk MAR dan MNAR. Penilaian prestasi dilakukan pada set data simulasi menggunakan bias, ralat piawai dan ralat kuasa dua min. Prestasi dinilai selanjutnya pada tiga set data klinikal dengan membandingkan anggaran setelah pembetulan PVB. Untuk MAR, kaedah yang dicadangkan menunjukkan prestasi yang baik berbanding kaedah penanda aras. IPW-LogReg sepadan dengan prestasi kaedah penanda aras. SIPW-Resamp dan SIPW-BalResamp berprestasi lebih baik daripada pelbagai imputasi pada prevalens penyakit yang rendah. Untuk MNAR, IPW-Fact-LogReg-MNAR dan IPW-LogReg-MNAR berprestasi lebih baik daripada kaedah penanda aras dalam kebanyakan kes. Tambahan pula, kaedah berasaskan PS yang dicadangkan membolehkan fleksibiliti dalam menganggar PS (berasaskan model dan manual) dan dalam menganggar Sn dan Sp (berasaskan kiraan dan berasaskan model). Kaedah berasaskan PS yang dicadangkan menunjukkan prestasi yang baik dengan kebolehgunaan umum di bawah andaian MAR dan MNAR.

PARTIAL VERIFICATION BIAS CORRECTION IN DIAGNOSTIC ACCURACY STUDIES USING PROPENSITY SCORE-BASED METHODS

ABSTRACT

New diagnostic tests are evaluated in comparison to the gold standard tests in diagnostic accuracy studies. For binary diagnostic tests, the performance is quantified by accuracy measures, most important are sensitivity (Sn) and specificity (Sp). These measures are often biased owing to selective verification of the patients, known as partial verification bias (PVB). Existing PVB correction methods vary in their ability and approaches to handle missing at random (MAR) or missing not at random (MNAR) assumptions. Of the methods, the existing methods utilizing the propensity score (PS) showed limited use of weighting and only handled the MAR assumption. PS-based correction methods had the potential for improvement and extension under MAR and MNAR assumptions. This research objective was to design PS-based methods with weighted regression and resampling approaches to improve and extend PVB correction under MAR and MNAR assumptions. Three MAR PS-based methods of PVB correction were proposed: 1) Inverse probability weighted logistic regression (IPW-LogReg), 2) Scaled inverse probability weighted resampling (SIPW-Resamp) and 3) Scaled inverse probability weighted balanced resampling (SIPW-BalResamp). These methods were extended to MNAR by proposing two PS estimation methods: 1) PS with an auxiliary variable and 2) PS factorization. This resulted in six MNAR methods: 1) Inverse probability weighted logistic regression for MNAR (IPW-LogReg-MNAR) 2) Inverse probability weighted logistic regression by factorization for MNAR (IPW-FactLogReg-MNAR), 3) Scaled inverse probability weighted resampling for MNAR (SIPW-Resamp-MNAR), 4) Scaled inverse probability weighted resampling by factorization for MNAR (SIPW-FactResamp-MNAR), 5) Scaled inverse probability weighted balanced resampling for MNAR (SIPW-BalResamp-MNAR) and 6) Scaled inverse probability weighted balanced resampling by factorization for MNAR (SIPW-FactBal-Resamp-MNAR). Simulated data sets were generated for combinations of disease prevalence, Sn, Sp, verification probabilities and sample sizes for MAR and MNAR. Performance evaluation was performed on simulated data sets using bias, standard error and mean squared error. The performance was further evaluated on three clinical data sets by comparing the PVB corrected estimates. For MAR, the proposed methods perform well as compared to the benchmark methods. IPW-LogReg matches the performance of the benchmark methods. SIPW-Resamp and SIPW-BalResamp show better performance than multiple imputation at low disease prevalence. For MNAR, IPW-Fact-LogReg-MNAR and IPW-LogReg-MNAR show better performance than the benchmark method in most conditions. Furthermore, the proposed PS-based methods allow flexibility in estimating PS (model-based and manual) and in estimating Sn and Sp (count-based and model-based). The proposed PS-based methods demonstrate good performance with general applicability under MAR and MNAR assumptions.

CHAPTER 1

INTRODUCTION

1.1 Background

Diagnostic tests play a very important role in medical care, which will determine what additional tests, treatments and interventions are needed for the patients (Kosinski & Barnhart, 2003a; Zhou, Obuchowski, & McClish, 2011). Diagnostic tests help clinicians in making diagnosis by giving objective measurements of medical conditions, for example in form of blood tests and imaging procedures. Examples of common diagnostic tests are mammography to detect breast cancer, pap smear to detect cervical cancer, antigen test for detection of COVID-19, and the commonly used pregnancy test. Diagnostic tests serve several roles, which are to provide reliable information about the patient's condition, to influence clinician's plan of management and to understand disease mechanism and natural history (Zhou et al., 2011). Given these important roles, any new diagnostic tests must undergo a thorough evaluation before being used in clinical settings in form of diagnostic accuracy studies (Kosinski & Barnhart, 2003a; Linnet, Bossuyt, Moons, & Reitsma, 2012; Umemneku Chikere, Wilson, Graziadio, Vale, & Allen, 2019).

Diagnostic accuracy studies are research studies which evaluate the ability of diagnostic tests to discriminate between patients with and without the disease or medical condition (O'Sullivan, Banerjee, Heneghan, & Pluddemann, 2018; Zhou et al., 2011). The evaluation of a new diagnostic test or index test involves comparing the result of the index test with the result of a definitive gold standard (de Groot, Bossuyt, et al., 2011; Hall, Kea, & Wang, 2019; O'Sullivan et al., 2018). For example, in diagnosing COVID-19 the rapid test kit-antigen test is compared against the gold standard test by reverse transcription polymerase chain reaction, and in diagnosing coronary artery disease the non-invasive stress echocardiography is compared against the gold standard test coronary angiogram. The performance of the test is quantified by a number of accuracy measures, depending on the scale of the index test results. For index tests with binary results, sensitivity or true positive rate (the proportion of diseased patients with a positive test result) and specificity or true negative rate (the proportion of non-diseased patients with a negative test result) are commonly used indicate the accuracy (Alonzo, 2014; He & McDermott, 2012; Martinez, Achcar, & Louzada-Neto, 2006). For index tests with continuous scales, receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) are used instead to indicate the accuracy (Alonzo, 2014; Umemneku Chikere et al., 2019).

It is important to obtain valid and unbiased estimates of accuracy measures to ensure clinical validity of the tests (Umemneku Chikere et al., 2019). In ideal research situations, optimal patient selection is achieved by random selection of patients and inclusion of an appropriate spectrum of patients that can be generalized to clinical practice (Hall et al., 2019). In addition, all patients that are tested by the index test must also be verified by the gold standard test (Hall et al., 2019; Leeflang & Allerberger, 2019; Linnet et al., 2012). This ensures that the accuracy measures are estimated with minimal bias (Hall et al., 2019).

However, the evaluation of the index tests in any diagnostic accuracy studies often suffer from some forms of bias from the suboptimal selection and verification procedures (Hall et al., 2019). The introduction of bias in the studies, from the analysis point of view, leads to overestimation or underestimation of the true accuracy measures (de Groot, Janssen, et al., 2011; Hall et al., 2019). From the clinical point of view, the biased estimates of the accuracy measures are misleading, which may result in premature implementation of the tests and wrong decision making by the clinicians (de Groot, Bossuyt, et al., 2011; Hall et al., 2019; Rutjes et al., 2007).

Despite these critical impacts of bias, in real studies researchers are not always able to satisfy the perfect methodological standards in conducting diagnostic accuracy studies due to several real-world limitations such as budget constrains, rare diseases and inability to apply index tests or gold standard tests equally to all patients in a study sample (Hall et al., 2019). With regard to the latter limitation, the ascertainment of the target outcomes by the gold standard tests can be clinically infeasible due to cost, ethical and clinical considerations, and time-consuming and invasive procedures (Alonzo, 2014; de Groot, Bossuyt, et al., 2011; Hall et al., 2019; O'Sullivan et al., 2018; Pepe, 2011; Pluddemann, McCall, O'Sullivan, & Banerjee, 2019; Schmidt, Walker, & Cohen, 2015). For these reasons, patients and clinicians are less likely to proceed with verifying the presence of the disease by the gold standard test if the index test result is negative (Pluddemann et al., 2019).

Given these limitations, only a subsample of the patients is selected for verification by the gold standard test based on the result of the index test (O'Sullivan et al., 2018). In other words, patients with a positive index test result are more likely to be verified by the gold standard test because they are presumably more likely to have the disease or medical condition, while those with a negative index test result are less likely to be selected in the verification sample because they are presumably less likely to have the disease. This selective sampling may also depend on other factors, for example sex, age and clinical symptoms (de Groot, Janssen, et al., 2011; Kosinski & Barnhart, 2003a). This sampling issue creates a form of bias, specifically known as partial verification bias (PVB) (de Groot, Bossuyt, et al., 2011; Hall et al., 2019; O'Sullivan et al., 2018). PVB leads to biased estimates of the accuracy measures (de Groot, Bossuyt, et al., 2011; O'Sullivan et al., 2018; Rutjes et al., 2007), most often are biased sensitivity and specificity estimates (Alonzo, 2014; de Groot, Bossuyt, et al., 2011).

The problem with PVB may go unnoticed in published diagnostic accuracy studies, and often other experts pointed out the presence of PVB and issues stemming from PVB in these studies. For example, Schmidt and Factor (2015), Schmidt (2017), Anzola et al. (2019) and E. S. L. Pedersen and de Jong (2019) highlighted the presence of PVB in studies by Díaz et al. (2014), Feinstein, Alonso, Yang, and John (2016), Nobashi et al. (2016) and Kim et al. (2018) respectively. For diagnostic accuracy studies that suffer from PVB, correcting the bias by utilizing the available methods will improve the validity of the accuracy estimates (Alonzo, 2014; O'Sullivan et al., 2018). Methods for PVB correction are available, depending on the scale of the index test, target outcome and missing data mechanism (Alonzo, 2014; Umemneku Chikere et al., 2019).

PVB can be viewed as a missing data problem, where the disease status is missing for unverified patients (Alonzo, 2014; de Groot, Bossuyt, et al., 2011). There are two possible missing data mechanisms in PVB, which are missing at random (MAR) and missing not at random (MNAR). MAR happens when the disease status is missing (unverified) depending on the index test result and other observed variables. MNAR happens when the disease status is missing depending on the disease status itself and other unobserved variables.

With regard to the scale of the index test and target outcome, the scope of this research was on the binary index test (positive or negative) with a binary disease status (disease or no disease). Within this scope, since Begg and Greenes (1983) proposed the landmark paper on a PVB correction method based on Bayes' rule, widely known as Begg and Greenes' (BG) method, many correction methods have been proposed to improve and extend the baseline method to handle MAR or MNAR assumptions (Umemneku Chikere et al., 2019).

1.2 Problem Statements

Most existing methods that handle MAR assumption rely on unbiased distribution of disease status given index test result, specifically $P(Disease|Test)$, to correct the accuracy estimates, which are BG-based methods (Alonzo & Pepe, 2005; Day, Eldred-Evans, Prevost, Ahmed, & Fiorentino, 2022; de Groot, Janssen, et al., 2011) and multiple imputation (MI) method (Day et al., 2022; Harel & Zhou, 2006). Two MAR methods rely on propensity score (PS) for PVB correction, namely inverse probability weighting estimation (IPWE) method (Alonzo & Pepe, 2005) and propensity score stratification (PSS) method (He & McDermott, 2012). In contrast to the BG-based and MI methods, PS instead specifies a direct relationship to the verification problem, because it defines the probability of verification status given index test result or specifically $P(Verified|Test)$.

PS is widely used in sampling bias correction and different PS-based approaches have been developed in specific domains (Austin, 2011; Krautenbacher, Theis, & Fuchs, 2017; Leyrat et al., 2019; Yasunaga, 2020). However, in the context of PVB correction, only two correction methods utilized PS for correction under MAR assumption (Alonzo & Pepe, 2005; He & McDermott, 2012). These methods are limited in scope; IPWE method is limited to algebraic approach (Alonzo & Pepe, 2005) and PSS was developed for dimension reduction (He & McDermott, 2012). Of the different PS-based approaches, inverse probability weighting (IPW) approach is the most flexible; it allows generation of synthetic data and estimation by regression method (Austin, 2011; Yasunaga, 2020) with generally good performance (Austin, 2011). The use of PS-based IPW approach by utilizing synthetic data and regression has not been considered before for PVB correction. Therefore, there are potential uses of weighted regression (King & Zeng, 2001) and weighted resampling (Nahorniak, Larsen, Volk, & Jordan, 2015) through synthetic data generation to develop and extend PVB correction based on PS.

By far, the existing PVB correction methods based on PS are limited to handle MAR assumption, whereby MNAR assumption is more realistic in clinical situation (Kosinski & Barnhart, 2003a, 2003b; Pennello, 2011). Therefore, PS-based PVB correction could be extended to also handle MNAR assumption to better reflect the clinical situation.

In order to evaluate the developed methods, simulated data and real data sets were used in this research, following the practice of previous PVB correction studies (Harel & Zhou, 2006; He & McDermott, 2012; Rochani, Samawi, Vogel, & Yin, 2015; Ünal

& Burgut, 2014). Simulated data allow comparison between known parameter (i.e. Sn and Sp) values and the estimated values from analytical methods (Nahorniak et al., 2015), where in the context this research the PVB correction methods. However, methods that work on simulated settings might not work in real clinical data sets. Therefore, in addition to the evaluation done on simulated data sets, this research also included commonly used clinical data sets for evaluating PVB correction methods. Utilizing clinical data sets allows evaluation and comparison between PVB correction methods using reference data sets, following the practice of previous research on PVB correction (Harel & Zhou, 2006; Kosinski & Barnhart, 2003a).

1.3 Research Questions

Based on the highlighted problem statements, the present research posed the following questions:

i. How to develop PS-based methods with weighted regression and resampling approaches for PVB correction in diagnostic accuracy studies under MAR assumption?

ii. How to extend the developed PS-based methods of PVB correction for MAR assumption in diagnostic accuracy studies under MNAR assumption?

iii. How the developed PVB correction methods perform in clinical diagnostic accuracy study data sets.

1.4 Objectives

Based on the research questions, the general objective of this research was to develop PS-based methods with weighted regression and resampling approaches for PVB correction in diagnostic accuracy studies under MAR and MNAR assumptions. The specific objectives were as follows:

i. To develop PS-based methods with weighted regression and resampling approaches for PVB correction in diagnostic accuracy studies under MAR assumption.

ii. To extend the developed PS-based methods of PVB correction for MAR assumption in diagnostic accuracy studies under MNAR assumption.

iii. To evaluate the developed PVB correction methods in clinical diagnostic accuracy study data sets.

The developed methods were expected to improve the existing methods in the following aspects: 1) Better accuracy as reflected by smaller bias, 2) Better precision as reflected by smaller standard error, 3) More flexibility, as reflected by whether the method can accommodate other methods in the algorithm for future expansion, and whether the method can be generalized or reused for other analyses.

1.5 Research Scope

To ensure the scale of this research was manageable, it was confined within the following scopes:

i. The index test was limited to a single binary diagnostic test, focusing the research on developing correction methods within this scope.

ii. The gold-standard test outcome was binary, reflecting the binary nature of disease status.

iii. This study focused on point estimates of sensitivity and specificity. Confidence interval was not studied.

iv. The performance of the methods under model misspecification was not studied.

v. The comparison to Bayesian methods was not be considered because they are dependent on pre-determined priors.

vi. Verification dependency on other variables (covariates) was not considered to focus the development of PS-based PVB correction methods on the verification dependency on the main variables; the index test and disease status.

1.6 Thesis Outline

This thesis is organized into seven chapters. Brief descriptions of the subsequent chapters are as follows:

i. Chapter 2: *Literature review*. This chapter presents a review of the literature pertaining to the existing PVB correction methods. It also review the methods and approaches related to the proposed PS-based methods in this research.

ii. Chapter 3: *Research methodology*. This chapter describes the methodology employed in this research. Research framework, problem formulation, data sets, performance metrics and experimental setup are presented in the chapter.

iii. Chapter 4: *PVB correction under MAR assumption using PS-based methods*. This chapter describes the relevant existing methods, presents the proposed PS-based PVB correction methods for MAR assumption and the experimental results.

iv. Chapter 5: *PVB correction under MNAR assumption using PS-based methods*. This chapter describes the relevant existing method, presents the proposed PS-based PVB correction methods for MNAR assumption and the experimental results.

v. Chapter 6: *Evaluation of the proposed methods in clinical diagnostic accuracy studies data sets*. This chapter presents the results as implemented in the clinical diagnostic accuracy studies under MAR and MNAR assumptions.

vi. Chapter 7: *Conclusion*. This chapter summarizes this thesis, revisits the research objectives, highlights its contributions, and suggests potential future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, it starts with general reviews on diagnostic accuracy studies and the issue of partial verification bias. Next, the existing methods to obtain accuracy of a binary diagnostic test with a binary outcome in the presence of partial verification bias are reviewed and evaluated. Research gaps pertaining to the existing methods are discussed and potential future research directions are laid out. Lastly, the concept of propensity score and its implementation is reviewed.

2.2 Diagnostic Accuracy Studies

Diagnostic accuracy studies are studies that aim to determine the ability of a new diagnostic test to confirm (rule in) and exclude (rule out) the presence of a disease (O'Sullivan et al., 2018). Diagnostic accuracy studies are also described as research studies that evaluate the ability of a diagnostic test to discriminate between patients with and without a disease or medical condition (Zhou et al., 2011). Diagnostic accuracy studies form a part of a larger framework of diagnostic medicine, in which diagnostic medicine is the process of identifying the disease or medical condition that a patient has, and ruling out diseases or medical conditions that the patient does not have through clinical assessment of the patient's signs, symptoms and results of various diagnostic tests (Zhou et al., 2011).

Zhou et al. (2011) summarizes the main purposes of a diagnostic test. First, it provides reliable information about a patient's medical condition or disease. Second, it influences the plan of a health care provider in managing a patient. Third, it allows understanding of disease mechanism and natural history. To serve these purposes, diagnostic accuracy studies are conducted to evaluate how the tests perform and how they should be interpreted (Zhou et al., 2011). Diagnostic accuracy studies allow thorough evaluation of new diagnostic tests before the tests are deployed in clinical settings (Kosinski & Barnhart, 2003a; Linnet et al., 2012; Umemneku Chikere et al., 2019). It is generally accepted that unless a diagnostic test is relatively accurate, as determined by proper evaluation process, it will not be useful in clinical practice (Kennedy, 2016).

2.2.1 Diagnostic Accuracy Study Design

In a diagnostic accuracy study, the new diagnostic test under evaluation is referred to as the index test (Cohen et al., 2016). The index test is evaluated by comparing with the result of the index test with the result of a gold standard test (de Groot, Bossuyt, et al., 2011; Hall et al., 2019; O'Sullivan et al., 2018; Zhou et al., 2011). A gold standard test gives a definitive diagnosis that determines the true disease status of a patient, and the test may have perfect or near perfect accuracy (Cohen et al., 2016; Umemneku Chikere et al., 2019; Zhou et al., 2011). A gold standard test gives information completely different from the index test under evaluation and serves as the point of reference in determining the true disease status (Zhou et al., 2011).

In the literature, four study designs are discussed for planning diagnostic accuracy studies. Most diagnostic accuracy studies are cross-sectional studies (Hall et al., 2019;

Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996), where the index and the gold standard test results are determined simultaneously, and this study design allows determination of the prevalence of disease (Gordis, 2009). Whenever the disease is rare, case-control study is an option for the diagnostic accuracy study (Hall et al., 2019; Linnet et al., 2012), where the patients are selected based on the known true disease status (Pepe, 2011). Randomized-controlled trial is a also an emerging study design for designing a diagnostic accuracy study (Kennedy, 2016; Lijmer & Bossuyt, 2009), although it is typically associated with studies that determine treatment modalities in clinical trials (Kennedy, 2016).

Lastly, cohort studies are prototypical designs that are recommended in literature on designing diagnostic accuracy studies (Linnet et al., 2012; Pepe, 2011; Zhou et al., 2011). Cohort studies are either prospective or retropective cohort studies (Cohen et al., 2016; Zhou et al., 2011). In the retrospective design, patients who underwent both the index and gold standard tests are identified from medical records, after which the test results are determined (Cohen et al., 2016; Zhou et al., 2011). The prospective cohort study is the ideal design for a diagnostic accuracy study, in which the index test is applied to consecutive patients suspected of having the target disease, and the gold standard test is later performed for each patient irrespective of the index test result (Linnet et al., 2012; Pepe, 2011; Zhou et al., 2011). A variation of the prospective cohort study design is a two-stage design, where the verification of the disease status by the gold standard test depends on the index test result (Pepe, 2011). This two-stage design of cohort study leads to partial verification bias, which will be discussed further in this chapter. In this research, the design of the diagnostic accuracy study generally refers to the prospective cohort study and its variation, unless stated otherwise.

An ideal diagnostic accuracy study should include a representative sample of patients who will be tested in clinical practice with the test of interest, or those more or less suspected of having the target condition (Hall et al., 2019; Leeflang & Allerberger, 2019). Consecutive or random sampling of patients at risk of the target condition of interest allows inclusion of an appropriate spectrum of patients that is generalisable to clinical practices (Hall et al., 2019), without making any judgement about how likely the patients are to be tested positive or negative (Leeflang & Allerberger, 2019).

All patients that are tested by the diagnostic test under evaluation must also be tested by the gold standard test (Hall et al., 2019; Leeflang & Allerberger, 2019; Linnet et al., 2012; Schmidt & Factor, 2013). The evaluation must be done is blinded manner, where the assessment of the gold standard test must be done without knowing the index test result (Leeflang & Allerberger, 2019). A typical flow for the evaluation of a binary diagnostic test (Leeflang & Allerberger, 2019; Umemneku Chikere et al., 2019) is given in Figure 2.1.

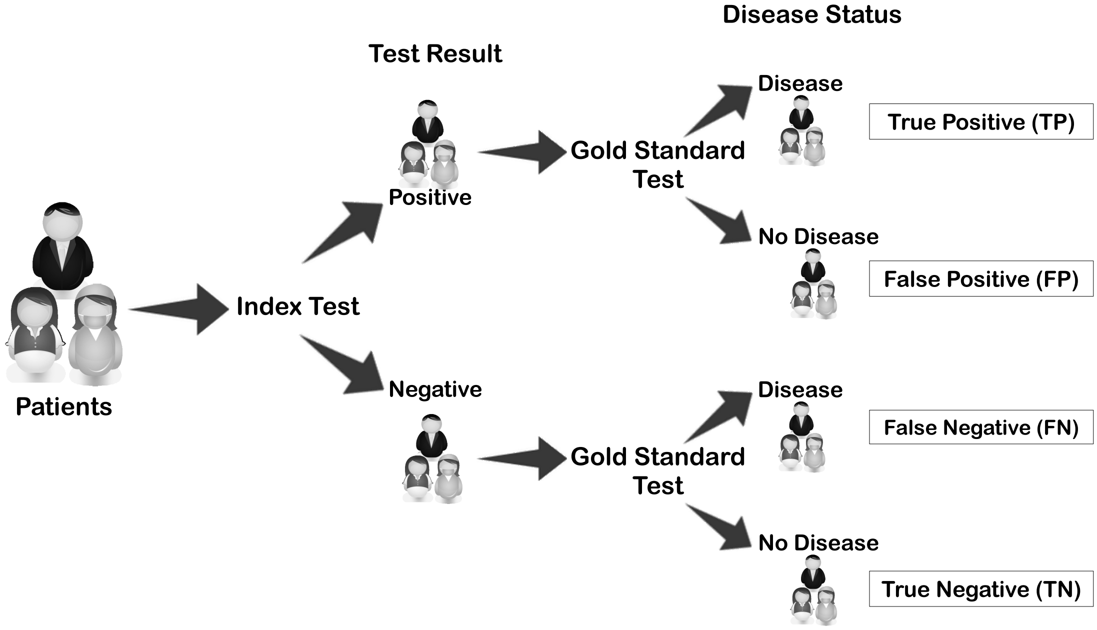

Figure 2.1: Flowchart of the evaluation of a binary diagnostic test

Having tested all patients for both tests, this will result in four combinations of results which are true positive (TP; index test positive, gold standard positive), false positive (FP; index test positive, gold standard negative), true negative (TN; index test negative, gold standard negative) and false negative (FN; index test negative, gold standard positive) (Leeflang & Allerberger, 2019; Umemneku Chikere et al., 2019). The number of patients that fall under each of these combinations allow the calculation of two basic measures of diagnostic accuracy, which are sensitivity (Sn) and specificity (Sp). Sn of a tests is its ability to detect the disease when it is present, while Sp is its ability to exclude the disease when it is absent (Zhou et al., 2011). In the ideal situation as illustrated in Figure 2.1, this provides unbiased estimates of the diagnostic accuracy measures (Umemneku Chikere et al., 2019). Sn and Sp are calculated as follows (Leeflang & Allerberger, 2019; Umemneku Chikere et al., 2019):

$$Sn = \frac{TP}{TP+FN} \tag{2.1}$$

$$Sp = \frac{TN}{TN+FP} \tag{2.2}$$

Diagnostic accuracy measures are not limited to only Sn and Sp. Other accuracy measures are discussed in the following section.

2.2.2 Diagnostic Accuracy Measures

The scale of the index test can be binary (positive or negative), ordinal (mild, moderate or severe infection) or continuous (fasting glucose level, blood pressure) (Linnet et al., 2012; Pepe, 2011; Zhou et al., 2011). Likewise, although a disease or medical condition is commonly binary (diseased or no disease), there are medical conditions of interest that are not binary, for example severity of asthma is on an ordinal scale (mild, moderate and severe asthma), and renal function as measured by glomerular filtration rate is on a continuous scale. The choice of diagnostic accuracy measures is mainly determined by the scale of the index test (Pepe, 2011; Zhou et al., 2011). For medical conditions with scales other than the binary scale, the scales can be dichotomized so that accuracy measures for binary disease status can be used (Zhou et al., 2011). Although this research focused on binary index test with binary disease status, this section gives an overview of other important measures for binary index test as well as non-binary index test.

There are a number of intrinsic accuracy measures, where the term intrinsic here denotes the test's inherent ability to correctly detect a condition when it is truly present and correctly rule out a condition when it is truly absent (Zhou et al., 2011). The most important quality of these measures is that the measures are not affected by the prevalence of the disease (Pepe, 2011; Zhou et al., 2011). An important implication of this property is that measures estimated from samples from a population will be applicable to other populations with different prevalence rates (Zhou et al., 2011). Another important implication of this property is that whenever the study designs are not suitable for estimating the prevalence of disease, which is typically estimated in a cross-sectional study, the choice of the study design will not affect the intrinsic accuracy measures.

For the binary index test, the available intrinsic accuracy measures are Sn, Sp, odds ratio, Youden's index and likelihood ratio, where Sn and Sp were already defined before. Odds ratio is the ratio between the odds of a positive test result relative to a negative test result among patients with the disease and the odds of a positive test result relative to a negative test result among patients without the disease (Zhou et al., 2011). Odds ratio is calculated based on Sn and Sp values as follows:

$$Odds\ ratio = \frac{Sn/(1-Sn)}{(1-Sp)/Sp} \tag{2.3}$$

Youden's index is calculated as follows,

$$Youden's\ index = Sn + Sp - 1 \tag{2.4}$$

Odds ratio and Youden's index are often used in meta-analyses of diagnostic accuracy tests (Zhou et al., 2011). Likelihood ratio is the ratio of the probability of getting a specific index test result (either positive or negative) in patients with the disease to the probability in patients without the disease (Deeks & Altman, 2004; Zhou et al., 2011). Likelihood ratio is defined separately for positive and negative likelihood ratios, calculated as

$$Positive\ likelihood\ ratio = \frac{Sn}{1-Sp} \tag{2.5}$$

and

$$Negative\ likelihood\ ratio = \frac{1-Sn}{Sp} \tag{2.6}$$

Likelihood ratio is not specific to the binary index test; it is also applicable to ordinal and continuous index tests, where it can be defined for an interval of test values and for the results of one side of a decision threshold (Zhou et al., 2011).

Other commonly used accuracy measures for the binary index test are overall accuracy (also referred to simply as *accuracy* or precisely as *probability of a correct test result*), and positive and negative predictive values (Linnet et al., 2012; Pepe, 2011; Umemneku Chikere et al., 2019; Zhou et al., 2011). Overall accuracy summarizes Sn and Sp in a single number, which is the weighted average of these measures (Zhou et al., 2011). Overall accuracy is calculated as

$$Overall\ accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.7}$$

This is a common measure of accuracy in other field of study such as machine learning and statistics to quantify the classification accuracy of a model or classifier (i.e. the inverse of the error rate) (James, Witten, Hastie, & Tibshirani, 2013). Positive predictive value (PPV) is the probability that the disease is present when the index test is positive, while negative predictive value (NPV) is the probability that the disease is absent when the index test result is negative (Linnet et al., 2012; Zhou et al., 2011). PPV and NPV are calculated as

$$PPV = \frac{TP}{TP+FP} \tag{2.8}$$

and

$$NPV = \frac{TN}{TN+FN} \tag{2.9}$$

Whenever PPV or NPV is equal to one, this indicates that the index test predicts the disease perfectly (Pepe, 2011). However, these measures are dependent on the disease prevalence as estimated in the diagnostic accuracy study, thus they are not intrinsic accuracy measures (Zhou et al., 2011). Nevertheless, these measures have their own advantages; overall accuracy is easy to calculate and provides combined measures of Sn and Sp in a single measure (Zhou et al., 2011), PPV and NPV are clinically important in interpreting a test result (probability that the patient has or does not have the disease given a positive or negative test result) (Linnet et al., 2012; Pepe, 2011; Zhou et al., 2011).

For the ordinal index test, the same intrinsic accuracy measures used for the binary index test are used, where a cutoff point for dichotomization of the ordinal scale is set (Zhou et al., 2011). For the continuous index test, the available intrinsic accuracy measures are receiver operating characteristic (ROC) curve and the area under the ROC curve (AUROC) (Linnet et al., 2012; Pepe, 2011; Umemneku Chikere et al., 2019; Zhou et al., 2011). ROC curve is a plot of Sn on the y-axis versus (1 - Sp) on the x-axis at several different cutoff points of the index test result (Hosmer, Lemeshow, & Sturdivant, 2013; Linnet et al., 2012). In other words, it is a plot of positive likelihood ratio at different cut off points. From the curve, a suitable combination of Sn and Sp that maximizes the sum of the two can selected, corresponding to the cut off point of the index test value (Linnet et al., 2012). ROC curve allows visual inspection of how the test performs (Zhou et al., 2011). AUROC is the area under the ROC curve, which is used to indicate the ability of a test to discriminate between patients with the disease and patients without the disease (Hosmer et al., 2013). AUROC provides an overall measure of diagnostic ability and represents the proportion by which patients with the disease have a higher index test result than patients without the disease for all possible pairs of patients with and without the disease in the diagnostic accuracy study (Linnet et al., 2012). Because this research focused on the binary index test, the plotting of ROC curve and the calculation of the AUROC are not described.

Having described all the accuracy measures, accurate estimation of these measures assumes an ideal research setup as shown in Figure 2.1. However, diagnostic accuracy studies are subject to a number of biases, broadly classified into the bias due to suboptimal patient selection and the bias due to interpretation and verification of the index test (Hall et al., 2019; Kea, Hall, & Wang, 2019; Whiting et al., 2004). Relevant to this research, verification bias falls under both of these classification as detailed in the next section.

2.2.3 Diagnostic Accuracy Studies and Verification Bias

Accurate and consistent confirmation of disease status or medical condition is crucial in diagnostic accuracy studies (O'Sullivan et al., 2018). Verification bias, also known as work-up bias and referral bias, occurs during evaluation of diagnostic test accuracy when there is a difference in testing strategy between groups of patients, leading to different ways of verifying the disease of interest (Pluddemann et al., 2019). This bias occurs because the result for the gold standard test is often missing in some of the patients (Naaktgeboren et al., 2016). The verification of the disease by the gold standard can be incomplete for several reasons, which are:

i. **Study design**. The verification by the gold standard test is planned or undesirable for cost efficiency, technical reason and ethical reason (de Groot, Bossuyt, et al., 2011; O'Sullivan et al., 2018; Pepe, 2011; Pluddemann et al., 2019; Schmidt & Factor, 2013; Schmidt et al., 2015)

ii. **Clinical practice**. The verification is not done when the clinical plausibility of having the disease of interest is low, for example patients with less severe symptoms and negative index test result (Naaktgeboren et al., 2016). When the gold standard test involves invasive, time-consuming and clinically risky procedures, the verification will be limited to only selected groups of patients at high risk of having the disease (Alonzo, 2014; Naaktgeboren et al., 2016; O'Sullivan et al., 2018; Pepe, 2011; Pluddemann et al., 2019; Schmidt & Factor, 2013).

iii. **Infeasibility**. The verification is not done when the procedure involved in the gold standard test is technically impossible to perform, for example it is impossible to obtain a biopsy of cancerous tissue sample when there is no cancerous lesion observed in the patient (Alonzo, 2014; de Groot, Bossuyt, et al., 2011; Naaktgeboren et al., 2016).

Verification bias is categorized into two types, which are partial verification bias and differential verification bias (de Groot, Bossuyt, et al., 2011; Hall et al., 2019; Pluddemann et al., 2019). In partial verification bias, patients with a positive index test are preferentially selected to be tested with the gold standard test (Hall et al., 2019). On the other hand, in differential verification bias, patients with a positive index test will be verified with the gold standard test, while patients with a negative index test will be verified with an alternative reference test (de Groot, Bossuyt, et al., 2011; Hall et al., 2019). For example, in a study evaluating the accuracy of the D-dimer test for diagnosing pulmonary embolism, those tested positive with the D-dimer test are verified by the ventilation-perfusion scan (gold standard test), while those tested negative with the D-dimer test are only verified by clinical routine follow-up (alternative reference test) (Pluddemann et al., 2019). Although the alternative reference test is not as good as the gold standard test, it might be chosen based on a number of reasons, such as being less invasive or cheaper (Hall et al., 2019). As the issue of partial verification bias is central to this thesis, it is further elaborated in the next section.

2.3 Partial Verification Bias

Partial verification bias (PVB) occurs in a diagnostic accuracy study when only a subsample of patients is selected for verification by the gold standard test based, while the rest of the study sample are unverified by the gold standard test (O'Sullivan et al., 2018; Umemneku Chikere et al., 2019), where the verification rate depends on the index test result (Schmidt & Factor, 2013; Schmidt et al., 2015) and other variables (de Groot, Bossuyt, et al., 2011). The verification rate is selective (non-random), where patients with a positive index test result are more likely to be selected for verification by the gold standard test (de Groot, Bossuyt, et al., 2011; Schmidt & Factor, 2013). This selective sampling may also depend on high clinical suspicion based on other variables, for example, gender, age and clinical symptoms (de Groot, Bossuyt, et al., 2011; de Groot, Janssen, et al., 2011; Kosinski & Barnhart, 2003a). In simpler words, PVB occurs in a diagnostic accuracy study when the verification by the gold standard test depends on the index test result; patients with a positive index test result are more likely to be verified by the gold standard test because they are presumably more likely to have the disease, while patients with a negative index test result are less likely to be selected for verification because they are presumably less likely to have the disease. The flow of evaluation of a binary diagnostic test under PVB is given in Figure 2.2.

As an example of diagnostic accuracy studies with PVB issue, a diagnostic accuracy study by (Kim et al., 2018) investigated the diagnostic accuracy of fractional exhaled nitric oxide (FeNO) in diagnosing exercise-induced bronchoconstriction (EIB).

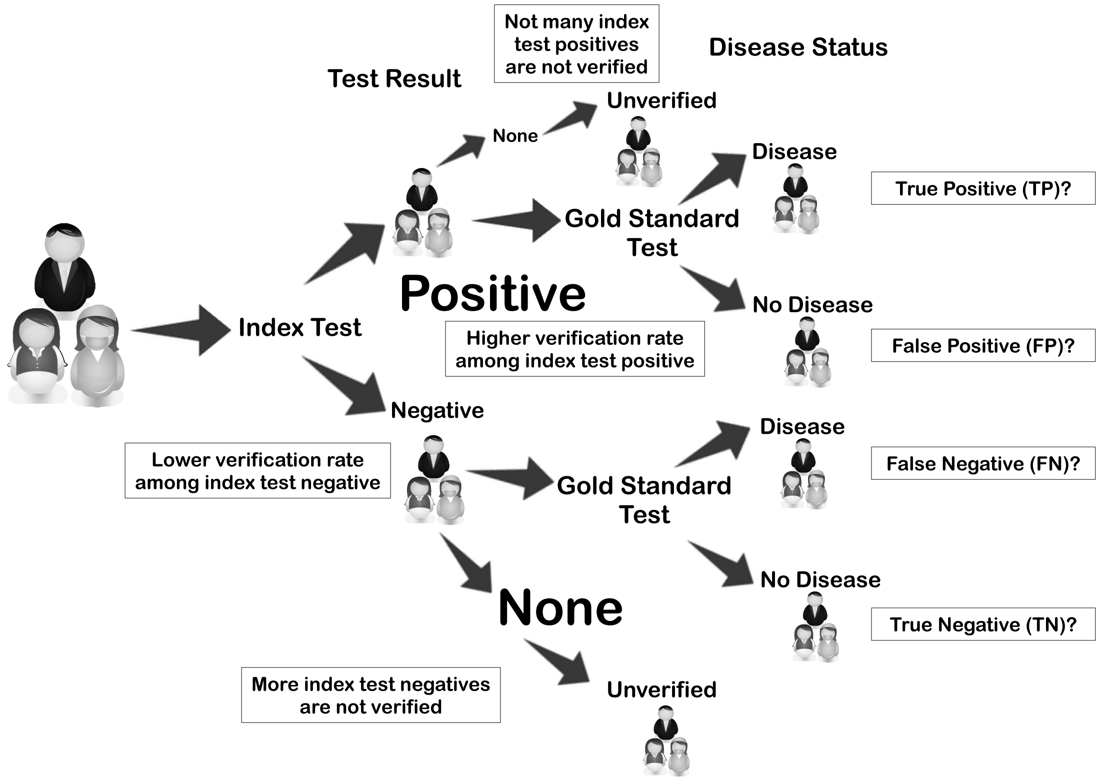

Figure 2.2: Flowchart of the evaluation of a binary diagnostic test under partial verification bias

The study considered only a subsample of asthmatic children who had both FeNO (the index test) and EIB (the gold standard test) results to obtain the sensitivity and specificity estimates, while the rest of the sample without the EIB result was not considered. E. S. L. Pedersen and de Jong (2019) pointed out that the study suffers from PVB, resulting in an overestimation of the sensitivity estimate. In another example, Porte et al. (2020) evaluated the performance of a new rapid test kit for diagnosing COVID-19 against the gold standard test by reverse transcription polymerase chain reaction. There was a possibility of PVB in their study, as convenient sampling was performed owing to the shortage of available test kits with a ratio of 2:1 for positive and negative samples respectively. Several other diagnostic studies with PVB issue (Ahn, Kim, Sohn, Choi, & Na, 2015; Díaz et al., 2014; Feinstein et al., 2016; Nobashi et al., 2016) were highlighted by other researchers (Anzola et al., 2019; Schmid & Cohen, 2017;

Schmidt, 2017; Schmidt & Factor, 2015).

2.3.1 Impacts of Partial Verification Bias

The impacts of PVB in diagnostic accuracy studies can be divided into the impact on statistical estimates and the impact on clinical practice. With regard to its impact on statistical estimates, PVB generally results in biased estimates of diagnostic accuracy measures. (de Groot, Bossuyt, et al., 2011; Hall et al., 2019; Naaktgeboren et al., 2016; O'Sullivan et al., 2018; Rutjes et al., 2007). Specifically, it often affects the estimates of Sn and Sp (Alonzo, 2014; de Groot, Bossuyt, et al., 2011; de Groot, Janssen, et al., 2011), where it will frequently cause an overestimation of Sn (Hall et al., 2019; Kea et al., 2019; O'Sullivan et al., 2018) and an underestimation of Sp (Alonzo, 2014; de Groot, Janssen, et al., 2011). Because the verification rate is low among those with negative index test result, this results in a lower number of FN, therefore increases the estimate of Sn (Hall et al., 2019; Schmidt et al., 2015). On the contrary, because the verification rate is high among those with positive index test result, this results in a higher number of FP, therefore reduces the estimate of Sp (Schmidt et al., 2015). Predictive values are generally unaffected by PVB (Schmidt et al., 2015; Zhou, 1994).

Regarding the impact of PVB in clinical practice, there are many implications of relying on biased estimates. PVB leads to clinically invalid tests (Umemneku Chikere et al., 2019), premature implementation of new tests (Rutjes et al., 2007), missed diagnosis (Hall et al., 2019), unnecessary clinical procedure (Hall et al., 2019), inefficient diagnostic testing, unnecessary costs and wrong clinical decision (de Groot, Bossuyt, et al., 2011).