PhD THESIS DECLARATION

The undersigned

SURNAME Pramov

FIRST NAME Aleksandar

PhD Registration Number 1626993

Thesis title: Noncompliance in Screening Trials

PhD in Statistics

Cycle 27^{th}

Candidate's tutor Professor Marco Bonetti

Year of thesis defence 2016

DECLARES

Under his responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28^{th} December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove

true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio istituzionale ad accesso aperto" (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

   - thesis Noncompliance in Screening Trials;
   - by Pramov, Aleksandar;
   - defended at Università Commerciale "Luigi Bocconi" – Milano in 2016;
   - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is ab-

solved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results, and is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date 15.11.2015

SURNAME Pramov

FIRST NAME Aleksandar

Contents

List of Figures

List of Tables

Acknowledgments

The road to the PhD title is long and difficult and one needs both academic and moral support during some hard periods. I was fortunate enough to have a supervisor who was able and willing to give me both. Therefore, I'd like to express my sincere gratitude to Prof. Marco Bonetti for his unwavering support throughout the period of my research. I feel that our countless discussions in the past two years have made me a better researcher and a better statistician. He managed to create a working atmosphere of palpable optimism, one that was always there for me to rely on when I ran into difficulties.

I was further fortunate enough to participate in the Bocconi PhD Program in Statistics. It allowed me to have close contact with top researchers that gave me much of what I know today about the field. Hence, I'd like to thank all the Bocconi academic staff involved in my PhD program, in particular its director Prof. Sonia Petrone for her continuous support during both my study and research phase.

During my B.Sc. and M.Sc. degrees I had the great pleasure to study under Prof. Ulrich Küsters and Dr. Holger Kömm who lit my spark of interest for research and for the science of Statistics. Without them, I surely would not have taken the road that has led me to the current thesis.

I thank Doctor Luigi Benecchi for the motivating example of the screening trial that is discussed in this manuscript. I sincerely thank Dr. Stuart Baker for his many comments and suggestions on how to explain the model for noncompliance in a better way, based on which I was able to substantially improve the manuscript.

I would also like to thank my internal PhD thesis reviewers - Prof. Rebecca Graziani and Prof. Fabrizia Mealli for their very helpful comments and suggestions.

Finally, my wife and my parents have certainly given me be biggest possible support that I could ever wish for. This manuscript is as much the result of their love and support, as it is of my efforts. Thank you.

Abstract

I consider a two-arm, unpaired, screen-positive randomized screening trial for a binary disease status. The aim is to evaluate the diagnostic performance of two competing diagnostic tests, which may recommend disease verification. The main contribution of this thesis is twofold: First, I discuss the specific case in which the recommendation from one of the tests also provides the would-be diagnostic recommendation of the other. I call the latter test "nested" within the former. I show that the nestedness property offers additional information and discuss how to capitalize on it. Second, I allow for all-or-none noncompliance with respect to the diagnostic test recommendation for disease status verification and I call this type of noncompliance "post-screening" noncompliance. Unlike earlier work, I do not assume that compliance is independent from disease status, but rather allow for the disease status to be missing ignorably, given the latent compliance class and the assigned test outcome. I define relevant measures of relative accuracy in the case of post-screening noncompliance and develop the associated inferential procedures for them. In particular, I discuss the modeling and inference for partially identified stratum-specific relative measures of accuracy. I explore the finite sample performance of these inferential procedures and discuss issues related to optimal sample size and minimal total cost for a proposed prostate cancer screening trial.
Keywords: Cost-effectiveness, disease verification, nested diagnostic test, partial identification, principal stratification, screen-positive trial, unpaired design

Chapter 1

Introduction

Disease screening is an integral part of any public health system. The main idea is that by regularly screening the population for particular diseases, necessary action can be undertaken to increase the lifetime of the individual and potentially reduce the public health costs. There are many aspects to consider when defining what a good screening policy should have: e.g. the time interval between two screening events, the type of disease to screen for, the age and gender group which should undergo the screening for that particular disease, the type of treatment to be administered to the patient after a positive disease status verification etc.

Another main aspect which lies in the focus of this thesis, is the choice of accurate diagnostic tests which should be administered to the patient. Ideally, every patient who is diseased must be flagged as such by the diagnostic test and patients who are not diseased should not get a positive diagnostic test result, i.e. overdiagnosis should be avoided. In other words, a medical diagnostic test can also be interpreted as a disease revealing procedure: patients who are diseased should get revealed as such, patients who are not diseased should not be revealed as such.

One tool to make such a choice between competing medical diagnostic tests is to conduct a randomized screening trial, which has the aim of comparing the accuracy between two (or more) of them. Typically, one does a two-arm randomized screening trial, where one of

the procedures is new, improved and hopefully more accurate (in some relative or absolute sense), while the other procedure is a routine procedure, already approved by the medical authorities. Another important component when making this choice is the economic one - even if there is statistically significant evidence that the new (and more expensive) procedure is more accurate than the old and established one, the additional cost must be set against the marginal improvement in accuracy.

One important property of a screening trial is whether the patients who are screened negative get their disease status verified. If this is not the case, i.e. if disease status verification is restricted to patients who have had a positive diagnostic test result, then such a trial is called a screen-positive trial. Cancer screening trials are a prime example of this, as it is not ethical to administer invasive disease verification procedures (e.g. biopsy) unless there is some statistically significant signal (i.e positive diagnostic test outcome) to do so.

Another classification criterion for two-arm screening trials is the one of paired vs. unpaired trials. In the former, both diagnostic tests are administered to each patient. If any of them (or both) have a positive diagnostic test result, the patient's disease status is verified clinically - typically by using the gold standard procedure for this. On the hand, unpaired trials administer, per protocol, one test to each arm. If the assigned test is positive then, per protocol, the patients's disease status gets verified. Unpaired tests are necessary, e.g. when it is not technically possible to administer the two diagnostic tests on the same person (see e.g. Alonzo and Kittelson (2006)). Moreover, as it will be shown in Section 3.3, an unpaired screening trial design might be preferable from an economic point of view as well. Figure 1.1 illustrates the points made above.

Using the screening trial classification terminology that is reviewed in the previous section, I now discuss the main research question and its motivating example.

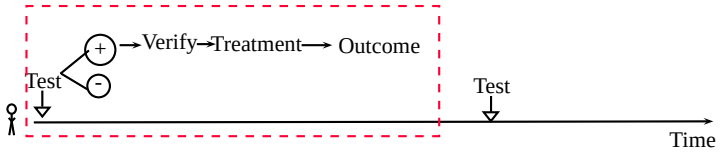


Figure 1.1: A general screen-positive screening procedure.

1.1 Motivating example

The aim of this manuscript is to construct and discuss a general inferential procedure to evaluate relative diagnostic test accuracy in an unpaired, two-arm randomized screening trial with two binary medical diagnostic tests. The trial is screening for a disease, the verification of which is of invasive nature (e.g. through biopsy). The thesis focuses on defining and analyzing a specific case of diagnostic tests, in which the results from one of the tests give the would-be test result of the other. I shall call the latter test "nested" within the former.

In addition, patients may exhibit non-compliance with the screening protocol. In particular, per protocol, disease status verification follows only after a positive screening test result, which makes the trial screen-positive (Pepe, 2003). However, deviation from the protocol in the form of an all-or-none noncompliance (Baker, 1997) with the recommendation for disease status verification is allowed for. Thus, one can have a situation where the disease status is possibly nonignorably missing Rubin (1974, 1976, 1978), despite a positive test screening result. Likewise, it is possible for a patient to obtain a disease status verification in spite of a negative test. Such type of noncompliance is observed after the patient has obtained the diagnostic result and I shall call this type of noncompliance "post-screening noncompliance."

The endpoint of the screening trial is the disease status verification. In particular, medical treatment procedure outcomes such as, e.g. survival time after surgery are not considered. Here, the screening trial is understood to be a comparison between two competing disease revealing mechanisms. This helps to formulate the problem in a way which allows for causal interpretation of the relevant metrics for diagnostic accuracy. Note that screening trials

usually repeatedly collect measurements of patients' characteristics over time. Here, the focus is on a static analysis (one-period screening), i.e. I look at a snapshot of the population at a given time point.

The initial motivation and the notions of nested screening tests and post-screening non-compliance behind this work came from a planned, unpaired, screen-positive, randomized screening trial for prostate cancer in the city of Cremona, Italy. While the methodology outlined in this thesis holds in general for any type of screening trial with the outlined design, for the purposes of clarity it is instructive to consider this planned trial as a practical, problem-motivating example.

The Cremona trial will aim to compare the accuracy of two prostate cancer (PC) diagnostic tests. The endpoint of the study is the disease status verification of the results from the diagnostic tests - a patient might have "high-grade" prostate cancer ($PC = 1$) or have "low-grade or no prostate cancer" ($PC = 0$). The two diagnostic tests - the Prostate Serum Antigen (PSA) test and the Prostate Health Index (PHI) - are binary-valued functions of selected biomarkers and a positive test suggests that a biopsy be performed to identify the true disease status of the patient. It is assumed that the biopsy is the gold standard for identification of PC. An interesting feature is that the PSA test is not proprietary, and indeed it is a known function (the PSA test is positive, if the value of the PSA biomarker is bigger than a chosen threshold, it is thus a step-function of said biomarker) of a subset of the selected biomarkers of the PHI test. Thus, although the trial has an unpaired design, additional information of the would-be PSA test result is available to the data analyst, even for patients who were randomized to be screened by the other test. In such a situation, I call the PSA test "nested" within the PHI test. The aforementioned additional biomarkers used for the latter test may be very expensive compared to those (e.g. routine, clinical measurements) used in the former test.

To the best of my knowledge, the way to exploit such type of additional information has not been addressed explicitly before. Moreover, existing methodology which considers

nonnested unpaired tests, does not exploit the additional information that the "nestedness" offers. The motivation for looking at such type of nested tests is twofold: First, from a practical perspective, more and more tests are based on a functional combination of biomarkers (say, gene expression levels) and are often compared with existing tests based on a subset of these biomarkers. Indeed, I believe that due this reason, the manuscript addresses a germane problem for cancer screening Moreover, the post-screening noncompliance is also relevant to cancer screening, as often disease verification procedures are painful and patients may choose to avoid them.

It will be shown that in such a design involving a nested test, some underlying population parameters which would be nonidentifiable in a two-arm design without a nested test, become identifiable from the observed data distribution. Second, it will be shown that a design involving a nested diagnostic test would allow us to construct additional test statistics for relevant hypotheses of relative test accuracy. This consideration is also related to design issues, which will be discussed in Chapter 3.

1.2 The screening trial setup

To formalize the argument and introduce some basic notation, consider a randomized screening trial with two arms - arm 0 and arm 1. In each arm, a patient is assigned to undergo one of the two diagnostic tests to screen for the presence of a disease D. Let D be binary (or be meaningfully dichotomized): $D \in \{0, 1\}$ where 0 stands for "no disease" and 1 for "disease". As an example, in a prostate cancer screening trial, $D = 1$ might be "high-grade" prostate cancer, while $D = 0$ may be "none or low-grade" prostate cancer. Let $Z \in \{0, 1\}$ indicate the assignment to a test (0 for arm 0 and 1 for arm 1). Without loss of generality, assume $P(Z = 0) = 0.5$.

Per protocol, each patient is randomly assigned to conduct only one of the diagnostic tests (i.e. an unpaired design, see Pepe (2003)). An example of such a study could be

Alonzo and Kittelson (2006) where the two diagnostic tests for cervical cancer are mutually exclusive. The outcome of each diagnostic test is a binary recommendation for further disease status verification - 1 advising the patient to verify the disease status and 0 advising for no verification. Throughout, it assumed that there are no compliance issues on the path between randomization and obtaining the test recommendation, i.e. one always observes the recommendations from the assigned tests of all patients. Let $T_0 \in \{0, 1\}$ and $T_1 \in \{0, 1\}$ denote the two diagnostic test recommendations.

Let us now formalize the notion of a "nested" screening test: Let T_1 be a binary-valued function (the exact functional form may or may not be explicitly known to the data analyst, as is the case when the diagnostic test is proprietary) of a certain set of patient characteristics $\mathcal{B}$ which are measured and known (e.g. the PSA biomarker, patient's age etc.) Let T_0 be another (the functional form is here known to the data analyst, e.g. the standard PSA diagnostic test is a step-function of the PSA biomarker) binary-valued function of $\mathcal{A}$, where $\mathcal{A}$ is another set of patient's characteristics, such that $\mathcal{A} \subseteq \mathcal{B}$. It is then clear that if a patient obtains the diagnostic test T_1, the data analyst will be automatically in the position to know the would-be value of T_0, even if the patient was not scheduled to undergo T_0.

Definition 1.2.1 *Let $\mathcal{B}$ be a set of measured patients' characteristics. Let $\mathcal{A}$ be another set of patients' characteristics, such that $\mathcal{A} \subseteq \mathcal{B}$. Let the binary-valued diagnostic test T_1 be a function of $\mathcal{B}$ and binary-valued T_0 be a function of $\mathcal{A}$. T_0 is then called a "nested" diagnostic test within the diagnostic test T_1.*

Note, that even when having "nested" tests, the screening trial design considered here is still unpaired - per protocol, patients receive only one of the tests and disease status verification is obtained only if the assigned test recommends it. In what follows, it is assumed that T_0 is nested within T_1. I will discuss the non-nested case later.

To distinguish between the factual and counterfactual outcomes of the test results, let us denote the (observed, i.e. factual) test recommendation of the assigned test as $T(Z)$,

with $T(Z) = Z \cdot T_1 + (1 - Z) \cdot T_0$. Similarly, $T(1 - Z)$ is the test recommendation of the unassigned test. Note that since T_0 is nested within T_1, whenever $Z = 1$, both $T(Z)$ and $T(1 - Z)$ are observed. To simplify notation, a patient-specific subscript m is omitted: e.g. in $T_{m,1}(Z)$ which indicates that the result refers to a given patient m, it would hold that $T_1(Z) = T_{m,1}(Z)$.

Let $R^D(T(Z), Z) \in \{0, 1\}$ indicate the revealing of the true D, as a function of the arm assignment Z and the result of the assigned test. Indeed, $R^D(T(Z), Z)$ is the outcome of interest in the study, as by measuring diagnostic accuracy one is comparing two disease status revealing procedures. Note that, per design, Z influences R^D only through $T(Z)$, thus, for each patient it holds $R^D(T(Z), Z) = R^D(T(Z))$. To simplify notation, I will write R^D instead, when the dependence on $T(Z)$ is implied by the context. Similarly, $R^{T_0}(Z) \in \{0, 1\}$ and $R^{T_1}(Z) \in \{0, 1\}$ indicate whether the T_0 and T_1 recommendations are revealed, as functions of Z. Throughout, an indicator variable $R = 0$ stands for "value missing." I assume that the test result or the disease status revealing of a given patient (as well as any other variable) does not depend on the assignment status and the diagnostic test results of other patients, i.e. $T(Z) = T(\boldsymbol{Z})$ and $R^D(Z, T(Z)) = R^D(\boldsymbol{Z}, \boldsymbol{T}(\boldsymbol{Z})))$, where the vector $\boldsymbol{Z}$ contains the treatment assignments for all patients in the study. Another assumption is that that there are no "hidden" tests, i.e. all patients get the same version of their assigned diagnostic test. These two assumptions amount to the well known Stable Unit Treatment Value Assumption (SUTVA) postulated by Rubin (1974, 1978).

1.3 Noncompliance in screening trials

I now introduce the issue of post-screening noncompliance. This is a relevant practical phenomenon, which is encountered often both in clinical and screening trials, e.g. the patient scheduled for a clinical disease verification might fail to show up for it, in particular when the verification procedure is painful. For cancer screening studies, Gareen (2007) gives an

excellent overview of the types of noncompliance that can be experienced in practice. The author outlines three types of noncompliance: (I) noncompliance with respect to screening test assignment; (II) post-screening noncompliance (which she calls is "noncompliance with follow-up"); and (III) noncompliance with the assigned treatment for the disease (e.g. surgery). Figure 1.2 illustrates those three types.

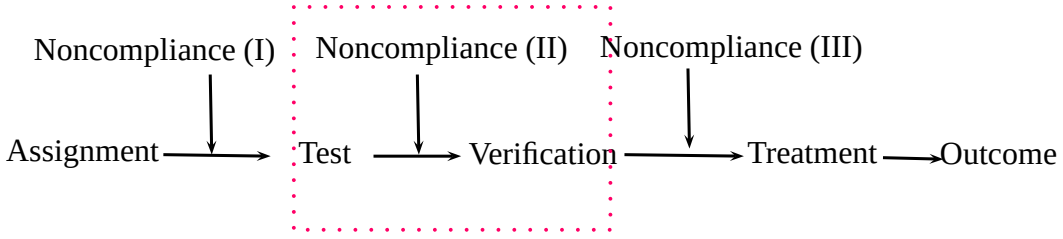


Figure 1.2: Types of noncompliance in a screening trial.

In the screening trial of interest here, there is no noncompliance of type (I). Every patient obtains the result of the originally assigned test. As the revealing of the disease diagnosis is the endpoint of the study, I neither consider prostate cancer treatment, nor noncompliance with respect to it, that is, of type (III). Instead, the focus of this thesis is on noncompliance of type (II), that is, noncompliance with respect to the recommendation for disease status verification (i.e. post-screening noncompliance). As an example for this, consider a patient who, upon obtaining a positive test recommendation, decides to not show up for the disease status verification and thus never has her true D status revealed. Throughout, the assumption is made that such a behavior would be recorded, i.e. known to the data analyst (see Gareen (2007) for more on collecting data on noncompliance).

The variable $Q(T(Z = z)) \in \{0, 1\}$ is introduced to indicate the observed decision of the patient with respect to proceeding with disease verification. As a consequence, it describes the compliance behavior with the test recommendation. I consider this decision-indicating variable to be a function only of the test result, not of the assignment. This assumption would be realistic in settings such as (double)-blind randomization. As an example, $Q(T(Z = z)) = T(Z = z)\ , \forall z \in \{0, 1\}$, would mean that the patient complies with the test recommendation

of the assigned test. $Q(0) = 1$ means that the patient does not comply with (the negative) the test recommendation of the assigned test, and decides instead to also undergo the test in the unassigned arm where she might obtain a positive recommendation for disease status verification. $Q(1) = 0$ means that the patient has received a positive test recommendation for disease status verification by the test in the assigned arm, but never undertook it and thus never revealed the true D. When one allows for post-screening noncompliance, the definition of R^D must be extended to include $Q(T(Z))$, and again to simplify notation I write R^D to indicate $R^D(T(Z), Q(T(Z)))$ when the functional dependence is clear from the context. Lastly, let U collect all unobserved variables which are associated by the disease and (potentially) cause noncompliance, e.g. family history of the disease or other personal background characteristics. Figure 1.3 gives a schematic description of the possible test and compliance outcomes for the screening trial setup that is discussed in this thesis, in terms of the developed notation in this and the previous sections.

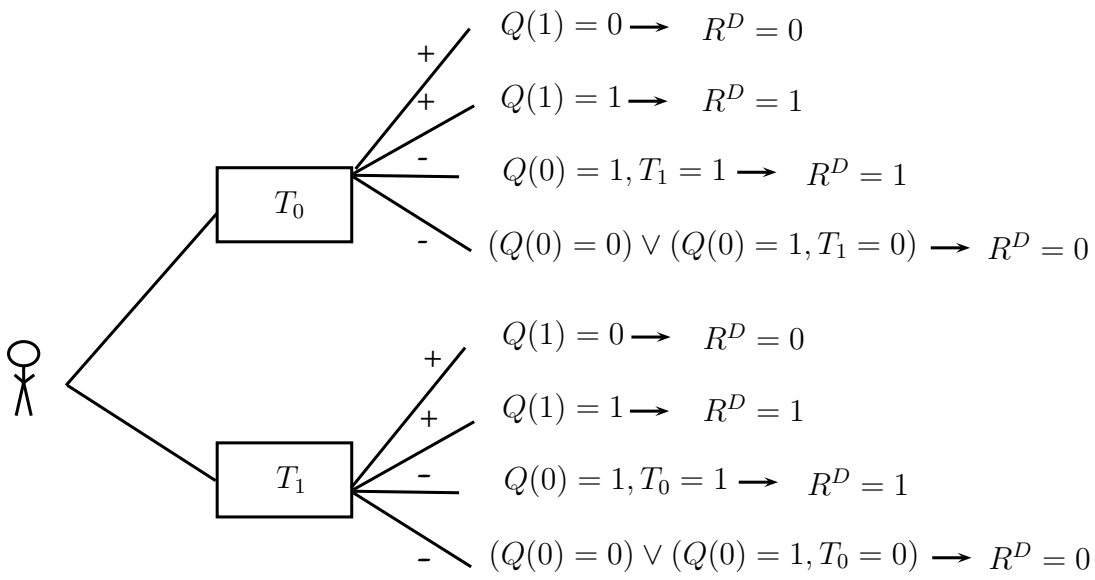


Figure 1.3: Schematic representation of a screen-positive, unpaired, two-arm screening trial with post-screening noncompiance.

1.4 Measures of accuracy

Without loss of generality, let the working hypothesis of the data analyst be that T_1 is a more accurate diagnostic test than T_0. The question is which measures of diagnostic test accuracy might be of interest here. Since in the discussion in this manuscript is for a screen-positive study, $P(D)$ is not identifiable from the observed data distribution.

Thus, I will work with two well-known measures of relative accuracy, namely the relative True Positive Rate (rTPR) and the relative False Positive Rate (rFPR) (Schatzkin et al. (1987), Cheng and Macaluso (1997), Pepe and Alonzo (2001), Pepe (2003)):

$$rTPR = \frac{P(T_1 = 1|D = 1)}{P(T_0 = 1|D = 1)} = \frac{P(T_1 = 1, D = 1)}{P(T_0 = 1, D = 1)}$$
$$rFPR = \frac{P(T_1 = 1|D = 0)}{P(T_0 = 1|D = 0)} = \frac{P(T_1 = 1, D = 0)}{P(T_0 = 1, D = 0)}$$

Under the working hypothesis of T_1 being more accurate than T_0 in relative terms, ideally $rTPR$ would be high and $rFPR$ would be low. Later on, these measures will be re-defined to account for the post-screening noncompliance. Throughout this thesis, the focus is on modeling and developing inferential procedures for $rTPR$, as both modeling and inference for $rFPR$ follows in a similar way. For the sake of brevity and to avoid tedious repetition, analysis of $rFPR$-based measures will be done only when it is meaningful to emphasize on any possible differences to the procedures for $rFPR$.

1.5 Related work

Summarizing the screening trial concept described above, the goal is to develop a general inferential procedure which estimates $rTPR$ and $rFPR$ in a screening trial with the features described above - a two-arm, screen-positive, unpaired screening trial for a binary disease via two nested diagnostic tests in the presence of post-screening noncompliance.

With regards to related work, note that in a screening trial with these properties, one is confronted with three major problems: First, the screening trial is screen-positive with an unpaired design. This means that the disease prevalence cannot identified from the observed data distribution. Problems of this type have been extensively discussed in the literature - among which Cheng and Macaluso (1997), Pepe and Alonzo (2001), Pepe (2003), Alonzo et al. (2002, 2004), Alonzo and Pepe (2005) and Broemeling (2007).

Next, as an additional problem, there may be post-screening noncompliance. Noncompliance with respect to assignment has been extensively studied in the case of clinical trials (see, e.g. Angrist et al. (1996), Frangakis and Rubin (2002), O'Malley and Normand (2005), Chen et al. (2009), Lui (2011) for a frequentist and Imbens and Rubin (1997), Schwartz et al. (2011), Mealli and Pacini (2013) for Bayesian approaches to its modeling and inference and Lui (2011) for a general discussion for noncompliance in clinical trials where the outcome is binary). In such types of noncompliance problems in clinical trials, one typically has noncompliance with respect to the assignment. Moreover, usually there is an active treatment (e.g. taking aspirin vs. taking placebo), a well defined outcome and the aim of analysis is to estimate and infer on the causal effect of that treatment on the outcome. Typical measures for this goal are the average treatment effect (ATE) and subpopulation-specific measures, such as the compliers average causal effect (CACE).

In addition to noncompliance with the assignment in clinical trials, some papers discuss the additional challenge added by a potentially missing outcome. Some earlier work on this has been done in Frangakis and Rubin (1999) for a rather specific trial design with some restrictions on the possibility of patients to randomized to placebo to cross-over to the treatment group. In the case of a binary outcome and binary compliance behavior (all-or-none) Mealli and Rubin (2002) discuss how to take proper account for the adjustment on the post-treatment variables that are the compliance behavior and the missing outcome and propose different set of assumptions which reduce the number of potential outcomes. Within the same setup of noncompliance and missing data, more recently, Chen et al. (2009) consider

cases in which the latent ignorability assumption (i.e. the model on which the outcome is not missing ignorably, if conditioned on the latent compliance strata and other relevant(and observable) quantities) does not hold.

Some authors model compliance behavior explicitly, on the basis of an assumed functional dependence (e.g. a regression model) between the compliance behavior and some pre-treatment covariates (e.g Barnard et al. (2003) and Bartolucci and Farcomeni (2013)). This can be particularly useful when one is confronted with non-ignorably missing outcomes. In a randomized trial comparing the effectiveness of two different teaching procedures for breast self examination, Mealli et al. (2004) proposes a nonignorable model for the missing data process and models compliance behavior on basis of relevant pre-treatment covariates.

Others, e.g. Cuzick et al. (1997), Baker (2000) also discuss noncompliance with respect to the assignment to an active treatment in the case of screening trials. A major difference in comparison to clinical trials, is that in a screening trial one typically has a diagnostic test occuring between assignment and outcome. And while the majority of the literature on noncompliance for screening trials assumes noncompliance of type (I), post-screening noncompliance and noncompliance with clinical treatment assignment have received much less attention (Gareen (2007)).

In this manuscript, the latter (the clinical treatment assignment, e.g. surgery) is not discussed, but it is worth noting that post-screening noncompliance is closely linked to the concept of verification bias (see, e.g. Alonzo (2005)), which arises in screen-positive studies. Noncompliance leads (additionally) to missingness of the disease status. If there were full compliance, then the disease status would still be missing for some patients (due to their assigned test giving a negative result) - but it would be missing at random (MAR), conditionally on the test results. Alonzo (2005) assumes that the disease status is MAR, conditional on recorded variables relevant for compliance (e.g. age, family history etc.).

Instead, in this thesis, this is not assumed - compliance may depend on the disease status and thus it is possible that the disease status is missing nonignorably. In particular, I do not

assume any specific model for compliance behavior - e.g. a regression model which predicts compliance based on some covariates.

Note that here, it is natural to consider endpoint of the study to be the indicator for disease status verification - $R^D(Z)$. This approach is beneficial for several reasons: First, it allows to disentangle the accuracy of the tests (in the form of the metrics presented above) from the (correct or false) disease status verification as a function of exposure to either one of the two tests. Furthermore, it allows to interpret the measures of relative accuracy for the complying patients in a causal way. Lastly, by employing this way of analysis, one can use the already present extensive literature on the principal stratification (PS) to model type noncompliance with the principal stratification method (Permutt and Hebel (1989), Baker and Lindeman (1994), Angrist et al. (1996), Frangakis and Rubin (2002), Baker et al. (2015)). In a nutshell, the PS method manages to account for the exhibited post-randomization noncompliance and to divide the population into subgroups (principal strata), such that the noncompliance behavior can be treated as a pre-treatment (i.e pre-trial) covariate, just as e.g. age would. Hence, conditioning on this newly defined subgroup would not introduce a potential bias and the result would be interpretable in a causal way. The downside is that often the belonging to this subgroup is not known for all patients. A way out of this is to assume a specific model for the noncompliance (which is not done here), or to conduct modeling and inference for a partially identified quantity in the presence of the ambiguous subgroup belonging. Problems of this type are often the subject of interest not only in the field of clinical and screening trials, but also of policy analysis. Thus, there are similarities between methods used in field of Econometrics and in Biostatistics which address similar problems. Indeed, the seminal paper of Angrist et al. (1996) establishes conditions under which instrument variables estimators can be interpreted as CACE under the Rubin Causal Model framework (Rubin (1974),Rubin (1978)) of potential outcomes. For the widely used PS method, Mealli and Pacini (2008) compare differences and similiraties to selection models Heckman (1974) in the case of nonignorable missingness and non-compliance.

In the particular case of post-screening noncompliance discussed here, the PS method allows us to gain more insight into what separates noncompliance issues in a clinical trial from noncompliance issues in a screening trial - in particular post-screening noncompliance in screening trials where disease status verification is not administered for every patient.

The work outlined in this thesis has some similarities with the work of Baker (2000). Baker considers a prostate screening trial, in which subjects are given active treatment (finasteride) vs. placebo. The considered outcome is the prevalence of prostate cancer, i.e. the question is whether or not finasteride lowers the prevalence of prostate cancer. Again, a biopsy-like verification for the true disease status of a patient is available only upon a positive diagnostic test. In the study outline in Baker (2000), that is the PSA test, which he calls an "auxiliary" variable. That is, a variable which occurs post-randomization and before the outcome of interest. Additionally, in the study considered by the author, there might be pre-screening noncompliance: some patients may decide to obtain finasteride outside the protocol or they might refuse to take the assigned finasteride.

Here, one can also view the two diagnostic tests as"auxiliary" variables which appear after randomization and before the outcome — disease status verification. However, unlike the trial design described by Baker (2000), there might be post-screening noncompliance and there is no active treatment that affects the prevalence of prostate cancer. Rather, the problem here is formulated as a comparison of two competing disease revealing mechanisms. That is to say, the endpoint of the study is the revealing of the disease status, and not the diagnosis itself as in Baker (2000).

Lastly, by virtue of the nonignorable missingness both of compliance type and disease status, it is not possible to point-identify the relevant measures of accuracy. However, one can provide stratum-specific nonparametric bounds for some compliance types. Thus, this work employs notions from the literature on partial identification to conduct inference for these bounds - see, e.g. Manski (1990), Horowitz and Manski (2000), Manski (2003), Imbens and Manski (2004), Stoye (2009), Woutersen and Ham (2013). In the case of noncompliance

in a screening trial, Cheng and Small (2006) construct bounds for effects in 3-arm trials and conduct a Bonferroni type, as well as bootstrap inference for them. Note that by deciding not to model compliance explicitly, e.g. by a regression model one loses on precision but gains on inferential credibility (Manski (2003)). Indeed, the bounds which are developed always hold under the assumptions that come as natural by the trial design and can be seen as doing modeling and inference "worst case" scenario in which the data analyst does not wish to or indeed has no adequate covariates or additional information to model noncompliance explicitly.

1.6 Thesis outline

The outline of this thesis is as follows: In Chapter 2, Section 2.1, I will present and discuss a model for the screening trial under the assumption of full compliance, with a nested diagnostic test. This is a convenient way to analyze the additional information which is coming from the nestedness property and its added value for the modeling without having to deal with the added complication of post-screening noncompliance. In following Section 2.2, I use the principal stratification method to define and model the post-screening noncompliance within the screening trial setup with a nested diagnostic test. Furthermore, I will address the issue of lack of identifiability of the relevant measures of relative accuracy and provide new versions of them, such that they become point-identifiable or at least partially identifiable. Chapter 3 addresses the inferential procedures for the models described in Chapter 2. Moreover, in Section 3.3 I will discuss the problem of a choice of design (paired vs. unpaired) when confronted with a nested diagnostic test. Chapter 4 shows some simulations which help evaluate the finite sample properties of the inferential procedures described above. I conclude my work in the last Chapter 5

Chapter 2

Model based on principal strata

2.1 Model with full compliance

In the context of a prostate cancer screening trial under full compliance with nested diagnostic tests - T_1 and T_0 as in Definition 1.2.1, consider the following Definition 2.1.1:

Definition 2.1.1 *A screening trial is called a two-arm, unpaired, screen-positive nested screening trial with full compliance if the following are true:*

1. $0 < P(Z = 1) < 1$ *(randomization)*

2. $(R^{T_1} = 1) \Rightarrow (R^{T_0} = 1)$ *(nestedness)*

3. $(T(Z) = 0) \Leftrightarrow (R^D = 0)$ *(screen-positive study with full compliance)*

Table 2.1 (top panel) illustrates the observed data configurations for $(Z, T(Z), R^D)$ under full compliance. The bottom panel will be addressed in the next section. Note that when all patients comply, it holds that $Q(T(Z)) = T(Z)$ for a given $Z = z$. That is to say, patients follow protocol and undergo disease status verification for $T(Z) = 1$ and do not verify their disease status when $T(Z) = 0$.

Assignment (Z)	Result of assigned test ($T(Z)$)	D status verified ($R^D = 1$)
Top panel: Model with full compliance		
T_1	1	yes
T_1	0	no
T_0	1	yes
T_0	0	no
Bottom panel: Model with post-screening noncompliance		
T_1	1	yes
T_1	1	no (patient did not comply)
T_1	0	yes (patient did not comply)
T_1	0	no
T_0	1	yes
T_0	1	no (patient did not comply)
T_0	0	yes (patient did not comply)
T_0	0	no

Table 2.1: Observed data configurations for $(Z, T(Z), R^D)$ under the models of full compliance (Section 2.1) and of post-screening noncompliance (Section 2.2).

Let the population parameters of the true underlying model under full compliance for the outlined screening trial be:

$$\pi_{11}^{c,d} = P(T_1 = 1, T_0 = 1, D = d)$$
$$\pi_{10}^{c,d} = P(T_1 = 1, T_0 = 0, D = d)$$
$$\pi_{01}^{c,d} = P(T_1 = 0, T_0 = 1, D = d)$$
$$\pi_{00}^{c,d} = P(T_1 = 0, T_0 = 0, D = d)$$

The superscript "c" is used to emphasize that all patients comply with the diagnostic test recommendation. In the next section, this notation will be extended to include noncomplying patients, with "c" indicating the subgroup of patients who comply. Let $\boldsymbol{\pi}^F$ be a vector collecting all of the 8 probabilities for all values $d \in \{0, 1\}$: $\boldsymbol{\pi}^F = \left(\pi_{11}^{c,1}, \pi_{10}^{c,1}, \ldots, \pi_{01}^{c,0}, \pi_{00}^{c,0}\right)$. It is assumed that $\boldsymbol{\pi}^F > \mathbf{0}$ and $\sum_d \pi^{c,d} = 1$. The latter assumption is natural and the former essentially ensures that there are no empty sets in the populations partition according to the levels of the vector (T_1, T_0, D). This assumption is natural when dealing with contingency

tables and will later be useful for estimation (see Agresti (2002)).

For general quantities such as vector of underlying parameters, vector of parameters of the observed data distribution, expressions for the likelihood and measures of relative accuracy, the superscript "F" (which stands for "full compliance") will be used to indicate that they are under the model with full compliance (e.g. $rTPR^F$ and $rFPR^F$ for the relative true positive rate and relative false positive rate respectively). In the next section, "F" will be replaced by "NC" for noncompliance. Note that in the model under full compliance, $rTPR^F$ and $rFPR^F$ can be expressed in terms of observable quantities. Under the assumptions of randomization and SUTVA which were made in Section 1.2:

$$rTPR^F = \frac{P(R^D = 1|D = 1, Z = 1)}{P(R^D = 1|D = 1, Z = 0)} = \frac{P(T(Z) = 1|D = 1, Z = 1)}{P(T(Z) = 1|D = 1, Z = 0)} = \frac{P(T_1 = 1|D = 1)}{P(T_0 = 1|D = 1)}$$
$$rFPR^F = \frac{P(R^D = 1|D = 0, Z = 1)}{P(R^D = 1|D = 0, Z = 0)} = \frac{P(T(Z) = 1|D = 0, Z = 1)}{P(T(Z) = 1|D = 0, Z = 0)} = \frac{P(T_1 = 1|D = 0)}{P(T_0 = 1|D = 0)}$$

Before deriving the observed data distribution, the parameters of which would allow for the estimation of $rTPR$ and $rFPR$, the model of true underlying distribution must be defined. The type of data considered here is categorical and the levels of the vector (T_1, T_0, D) define elementary cells of a contingency table. It is then natural to model the vector of random counts $\left(N_{11}^{c,1}, \ldots N_{00}^{c,0}\right)$ of patients via a multinomial sampling scheme. Such underlying "true" multinomial model based on the population parameters in $\boldsymbol{\pi}^F$ gives rise to the following likelihood

$$L_{true}^F \propto \prod_{d\in\{0,1\}} \left(\pi_{11}^{c,d}\right)^{n_{11}^{c,d}} \cdot \left(\pi_{10}^{c,d}\right)^{n_{10}^{c,d}} \cdot \left(\pi_{01}^{c,d}\right)^{n_{01}^{c,d}} \cdot \left(\pi_{00}^{c,d}\right)^{n_{01}^{c,d}}$$

A crucial point in the analysis is that the model of the observed data distribution is different from the true underlying model defined above. To derive it, first let $\boldsymbol{Y}$ collect all recorded patient data: $\boldsymbol{Y} = (T(Z), T(1-Z), D, Z)$ and assume that for a sample n patients one has $\boldsymbol{Y}_1 \ldots \boldsymbol{Y}_n \overset{i.i.d}{\sim} f(\boldsymbol{Y})$. Following standard approach of integrating over the missing data (here,

D status and unobserved test outcomes) in the full model (see e.g. Little and Rubin (2002)),

$$f(\boldsymbol{Y^{obs}}) = \int f(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}|Z) \cdot P(Z) d\boldsymbol{Y}^{mis}$$

Integrating over the missing values results into collapsing of cells of the contingency table defined by the levels of $\boldsymbol{Y}$ (Bishop et al. (2007)). Note that in the model with full compliance, D is MAR - its missingness is solely dependent on the assigned test, which in turn is observed for all patients. The resulting expression is proportional to a product multinomial sampling scheme for two independent screening trial arms. Thus, the observed data likelihood (conditionally, within each arm) and for the whole trial is given by:

$$L_1^F \propto \prod_d \left[\left(\pi_{11}^{c,d}\right)^{n_{11}^{c,d}} \cdot \left(\pi_{10}^{c,d}\right)^{n_{10}^{c,d}} \right] \cdot \left(\sum_d \pi_{01}^{c,d} \right)^{\sum_d n_{01}^{c,d}} \cdot \left(\sum_d \pi_{00}^{c,d} \right)^{\sum_d n_{00}^{c,d}}$$

$$L_0^F \propto \left(\pi_{11}^{c,1} + \pi_{01}^{c,1}\right)^{\left(n_{11}^{c,1} + n_{01}^{c,1}\right)} \cdot \left(\pi_{11}^{c,0} + \pi_{01}^{c,0}\right)^{\left(n_{11}^{c,0} + n_{01}^{c,0}\right)} \cdot \left(\sum_d \pi_{10}^{c,d} + \pi_{00}^{c,d} \right)^{\sum_d \left(n_{10}^{c,d} + n_{00}^{c,d}\right)}$$

It is convenient to re-parametrize it by using conditional probabilities, defined by the observable features of the patient:

$$p_{T(Z),T(1-Z)|Z}^{c,d} = P\left(T(Z), T(1-Z), D|Z = z\right)$$

in the full compliance model. The re-parametrized likelihood (conditionally within each arm) is then given by

$$L_1^F \propto \left(p_{11|1}^{c,1}\right)^{n_{11,1}^{c,1}} \cdot \left(p_{11|1}^{c,0}\right)^{n_{11,1}^{c,0}} \cdot \left(p_{10|1}^{c,1}\right)^{n_{10,1}^{c,1}} \cdot \left(p_{10|1}^{c,0}\right)^{n_{10,1}^{c,0}} \cdot \left(p_{01|1}^{c,*}\right)^{n_{01,1}^{c,*}} \cdot \left(p_{00|1}^{c,*}\right)^{n_{00,1}^{c,*}}$$

$$L_0^F \propto \left(p_{1*|0}^{c,1}\right)^{n_{1*,0}^{c,1}} \cdot \left(p_{1*|0}^{c,0}\right)^{n_{1*,0}^{c,0}} \cdot \left(p_{0*|0}^{c,*}\right)^{n_{0*,0}^{c,*}}$$

where $*$ is used to indicate that a variable is not unambiguously identifiable from the observed data, as e.g. the case of a patient's disease status D assigned to the arm 1 with both

the $T_1 = T(Z)$ and $T_0 = T(1-Z)$ diagnostic tests negative. Let all nine parameters from L_1^F and L_0^F be collected in the vector of parameters of the observed data distribution: $\boldsymbol{p}^F = \prod_{d\in\{0,1\}} (p^{c,d}_{11|1}, p^{c,d}_{10|1}, p^{c,*}_{01|1}, p^{c,*}_{00|1}, p^{c,d}_{1*|0}, p^{c,*}_{0*|0})$.

From the expressions of the likelihood, one can see that more cells are identifiable from arm 1, than from arm 0. This is due to the nestedness property, as one has more information about the would-be outcome of the unassigned test. Indeed, from the assumption of randomization, a link can be established between $\boldsymbol{p}^F$ and $\boldsymbol{\pi}^F$. For the analysis in this section, consider columns 1-7 ("Assignment", "Test outcomes" and "Full compliance") from Table 2.2 (the remaining columns will be discussed in Section 2.2).

Note that $\boldsymbol{\pi}^F$ is a noninjective function of the vector of parameters of the true underlying model. Thus, there will be some cells in the underlying data model, which will not be identifiable from the observed model. Under the full compliance model, there are nine parameters of the observed data distribution. To see the added value of nestedness, consider that for rows $j = 1, \ldots, 6$, the result of the T_0 test can be deduced from the arm 1. This allows for (some of) the corresponding in $\boldsymbol{\pi}^F$ to be factored out of the likelihood, thus making them point-identifiable. In contrast, in rows $j = 7, 8, 9$, the outcome of T_1 is not known, thus less parameters of $\boldsymbol{\pi}^F$ are point-identifiable. Note that in a model with full compliance, both $rTPR^F$ and $rFPR^F$ are point-identifiable, since

$$rTPR^F = \frac{\left(\pi^{c,1}_{11} + \pi^{c,1}_{10}\right)}{\left(\pi^{c,1}_{11} + \pi^{c,1}_{01}\right)}$$
$$rFPR^F = \frac{\left(\pi^{c,0}_{11} + \pi^{c,0}_{10}\right)}{\left(\pi^{c,0}_{11} + \pi^{c,0}_{01}\right)}$$

The next section introduces a model for post-screening noncompliance based on the principal stratification method.

Arm	True test outcomes			Full compliance model (F)			Noncompliance model (NC)		
Z	T_1	T_0	D	j	$\boldsymbol{p^F}$	$\boldsymbol{\pi}^F$	q	$\boldsymbol{p^{NC}}$	$\boldsymbol{\pi}^{NC}$
1	1	1	yes	1	$p^{c,1}_{11\|1}$	$\pi^{c,1}_{11}$	1	$p^{*,1}_{11\|1}$	$\pi^{c,1}_{11}+\pi^{i,1}_{11}$
1	1	0	yes	2	$p^{c,1}_{10\|1}$	$\pi^{c,1}_{10}$	2	$p^{*,1}_{10\|1}$	$\pi^{c,1}_{10}+\pi^{i,1}_{10}$
1	0	1	yes				3	$p^{i,1}_{01\|1}$	$\pi^{i,1}_{01}$
1	1	1	no	3	$p^{c,0}_{11\|1}$	$\pi^{c,0}_{11}$	4	$p^{*,0}_{11\|1}$	$\pi^{c,0}_{11}+\pi^{i,0}_{11}$
1	1	0	no	4	$p^{c,0}_{10\|1}$	$\pi^{c,0}_{10}$	5	$p^{*,0}_{10\|1}$	$\pi^{c,0}_{10}+\pi^{i,0}_{10}$
1	0	1	no				6	$p^{i,0}_{01\|1}$	$\pi^{i,0}_{01}$
1	1	1	?				7	$p^{r,*}_{11\|1}$	$\pi^{r,1}_{11}+\pi^{r,0}_{11}$
1	1	0	?				8	$p^{r,*}_{10\|1}$	$\pi^{r,1}_{10}+\pi^{r,0}_{10}$
1	0	1	?	5	$p^{c,*}_{01\|1}$	$\pi^{c,1}_{01}+\pi^{c,0}_{01}$	9	$p^{*,*}_{01\|1}$	$\begin{pmatrix}\pi^{c,1}_{01}+\pi^{c,0}_{01}\\ +\pi^{r,1}_{01}+\pi^{r,0}_{01}\end{pmatrix}$
1	0	0	?	6	$p^{c,*}_{00\|1}$	$\pi^{c,1}_{00}+\pi^{c,0}_{00}$	10	$p^{*,*}_{00\|1}$	$\sum_s\sum_d \pi^{s,d}_{00}$
0	?	1	yes	7	$p^{c,1}_{1*\|0}$	$\pi^{c,1}_{11}+\pi^{c,1}_{01}$	11	$p^{*,1}_{1*\|0}$	$\begin{pmatrix}\pi^{c,1}_{11}+\pi^{i,1}_{11}\\ +\pi^{c,1}_{01}+\pi^{i,1}_{01}\end{pmatrix}$
0	1	0	yes				12	$p^{i,1}_{01\|0}$	$\pi^{i,1}_{10}$
0	?	0	?	8	$p^{c,*}_{0*\|0}$	$\begin{pmatrix}\pi^{c,1}_{10}+\pi^{c,0}_{10}\\ +\pi^{c,1}_{00}+\pi^{c,0}_{00}\end{pmatrix}$	13	$p^{*,*}_{00\|0}$	$\begin{pmatrix}\sum_d\left(\pi^{c,d}_{10}+\pi^{r,d}_{10}\right)\\ +\sum_s\sum_d \pi^{s,d}_{00}\end{pmatrix}$
0	?	1	?				14	$p^{r,1}_{1*\|0}$	$\begin{pmatrix}\pi^{r,1}_{11}+\pi^{r,0}_{11}\\ +\pi^{r,1}_{01}+\pi^{r,0}_{01}\end{pmatrix}$
0	1	0	no				15	$p^{i,0}_{01\|0}$	$\pi^{i,0}_{10}$
0	?	1	no	9	$p^{c,0}_{1*\|0}$	$\pi^{c,0}_{11}+\pi^{c,0}_{01}$	16	$p^{*,0}_{1*\|0}$	$\begin{pmatrix}\pi^{c,0}_{11}+\pi^{i,0}_{11}\\ +\pi^{c,0}_{01}+\pi^{i,0}_{01}\end{pmatrix}$

Table 2.2: **Link between the vectors of parameters of the observed data distribution $\boldsymbol{p}^F, \boldsymbol{p}^{NC}$ and the parameters of the true underlying distribution $\boldsymbol{\pi}^F, \boldsymbol{\pi}^{NC}$ under the model of full compliance (F) in columns 5-7, and noncompliance (NC) in columns 8-10.**
For FC rows $j = 1-6$: By nestedness property, one can deduce result of T_0 diagnostic test in arm 1. For FC rows $j = 5, 6$: if T_1 is negative in arm 1, the disease status D is not observed. For FC rows $j = 7, 8$: if T_0 is positive, there is no nestedness so T_1 is not known. For FC row $j = 9$, if T_0 is negative and since there is no nestedness T_1 is not known; also D is not observed. For NC rows $q = 1, 4$ (arm 1): by nestedness, the probability of having $D = 1, T_1 = 1, T_0 = 1$ can be identified, in difference to row $q = 11$ (arm 1), where there is no nestedness. For NC rows $q = 3, 6, 12, 15$: from noncompliance in a two-arm study, one can fully identify these cells, together with the stratum (insisters). These can be used to identify (some) compliers from cells $q = 2, 5, 11, 16$. For NC rows $q = 1$ to $q = 10$: from nestedness, when T_1 is known, T_0 is also known. For NC rows $q = 7, 8, 14$: from noncompliance in a two-arm study, the stratum (refusers) is known, but the disease status D is not. See Section 2.1 for details on (F) and Section 2.2 for details on (NC).

2.2 Model with post-screening noncompliance

Allowing for post-screening noncompliance has implications for $rTPR$, $rFPR$ and their causal interpretation, for the true underlying model, and for the inferential procedures. This section extends the discussion of the previous section. First, a new definition which is similar, but not identical, to Definition 2.1.1 is provided in order to account for post-screening noncompliance:

Definition 2.2.1 *A screening trial is called a two-arm, unpaired, screen-positive nested screening trial with post-screening noncompliance if the following are true:*

1. $0 < P(Z = 1) < 1$ *(randomization)*
2. $(R^{T_1} = 1) \Rightarrow (R^{T_0} = 1)$ *(nestedness)*
3. $(T_0 = 0$ *and* $T_1 = 0) \Rightarrow (R^D = 0)$ *(screen-positive study with possible post-screening noncompliance)*

As evident from both the last point of Definition 2.2.1 and the bottom panel of Table 2.1, the failure of a patient to verify the disease status does not imply that the assigned test was negative. This is different than in Definition 2.1.1, where failure to verify implied that the assigned test is negative, since now the patient may decide not to comply with a positive recommendation for disease status verification. Relatedly, note that $(R^D = 1) \Rightarrow (Q(0) = 1 \text{ or } Q(1) = 1)$. In other words, for patients with a negative test recommendation from their assigned test, it is still possible to reveal their true status, if they decide not to comply with the protocol.

Note that now, without making any further assumptions at this stage, R^D is indeed a function of three variables - the assignment Z, the test result $T(Z)$ and the decision whether or not to follow the test recommendation for disease status verification — $Q(T(Z))$. As in the previous section, by the assumption of randomization, Z does not influence R^D directly, but only through $T(Z)$.

To account for the complication which arises from post-screening noncompliance, namely that it is possible to have $Q(T(Z)) \neq T(Z)$, the post-screening noncompliance is modeled explicitly. Generally, conditioning on a post-treatment variable - e.g. the observed decision $Q(T(Z))$, may lead to inferential bias, as noted by Rosenbaum (1984).

A possible way to deal with all-or-none compliance, is to use the principal stratification method. The main idea is the following: for any individual patient "m", use the potential outcomes of the joint vector $(T_m(Z), Q_m(T(Z)))$ for $Z \in \{0, 1\}$ to define principal strata. The advantage of using these is that they can be then considered as being pre-trial covariates characterizing the post-screening compliance behavior of the patients (Frangakis and Rubin, 2002). Typically, in clinical trials, the principal strata are built around the decision of the patient (factual and counterfactual) to comply with the treatment arm assignment (e.g. to take aspirin vs. to take placebo). This is different from how the strata are built here, since here there is no pre-screening noncompliance. Here, they are instead built on the potential outcomes for the compliance with the disease status verification.

In order to make the principal stratification framework operational, we first consider the aforementioned joint potential values of the diagnostic test outcome $T(Z)$ and the decision whether to follow the recommendation $Q(T(Z))$ for $Z \in \{0, 1\}$. It is instructive to first consider the set of all possible potential outcomes and then use reasonable assumptions (some of them holding by design) to simplify that set step-by-step. These assumptions will be summarized later in this section. Initially, the levels of the elements of the vector

$$(T(Z), T(1-Z), Q(T(Z)=1), Q(T(1-Z)=1), Q(T(Z)=0), Q(T(1-Z)=0))$$

taken for $Z = 0, 1$ define $2^7 = 128$ potential outcomes. Fortunately, one can use some assumptions that hold by design to reduce that big number. We consider them step-by-step:

The first assumption which holds by design is *randomization*.

Assumption 2.2.1 *Randomized screening trial*

The screening trial is randomized, i.e. the probability of assignment $P(Z)$ is an a-priori chosen, fixed number.

Since the screening trial is randomized, one has that $(T(Z)|Z = 1) = T_1$ and $(T(Z)|Z = 0) = T_0$, are both independent of Z and hence one can simplify the number down to $2^5 = 32$ potential outcomes. For the remaining text, $T(Z)$ and $T(1 - Z)$ can be thus understood as "assigned" and "unassigned" diagnostic tests respectively.

Assumption 2.2.2 *Blinded screening trial*

The screening trial is blinded. Thus, the patients do now know which test they are undergoing. From this, a realistic restriction to be put is $Z \perp\!\!\!\perp Q(T), Q(T - Z)|T(Z)$.

By design the randomized screening trial is *blinded*, and thus it holds that $Z \perp\!\!\!\perp Q(T), Q(T - Z)|T(Z)$ for all values of $Z \in \{0, 1\}$. In other words, the assignment has no direct effect on the decision whether to follow the recommendation or not. Intuitively, that would make sense in the case of a blinded trial - a patient's decision is solely based on her background characteristics and her test result. Formally, this means $Q(T(Z), Z) \equiv Q(T(Z))$ Since, per design, the patient doesn't know which test she is actually receiving, then the assignment plays no role in her decision. This reduces the number of potential outcomes from 32 to 16. These 16 potential outcomes (for patient "m") can be summarized in the following Table 2.3

Note that some of the quantities in several columns are a-priori counterfactuals, i.e. these outcomes are not just potentially unobserved, but actually non-existent. As an example, consider the first line where we have a patient assigned to arm $Z = z$, who is positive on both diagnostic tests, i.e. has the pair $(T_1 = 1, T_0 = 1)$. For that patient, the outcome $Q(T(Z) = 0)$ involves a quantity which is a-priori counterfactual - it does not exist, as both of his diagnostic tests are positive. Following Frangakis and Rubin (2002), I do not consider estimands which involve outcome values that are a-priori counterfactual. In other words, since per design we cannot control for the post-treatment variable $(Q(T(Z)))$, i.e.

$T_m(Z)$	$T_m(1-Z)$	$Q_m(T_m(Z)=1)$	$Q_m(T_m(1-Z)=1)$	$Q_m(T_m(Z)=0)$	$Q_m(T_m(1-Z)=0)$
1	1	1	1	[0]	[0]
1	0	1	[1]	[0]	0
0	1	[1]	0	0	[0]
0	0	[1]	[1]	0	0
1	1	1	1	[1]	[1]
1	0	1	[1]	[1]	1
0	1	[1]	1	1	[1]
0	0	[1]	[1]	1	1
1	1	0	0	[0]	[0]
1	0	0	[0]	[0]	0
0	1	[0]	0	0	[0]
0	0	[0]	[0]	0	0
1	1	0	0	[1]	[1]
1	0	0	[0]	[1]	1
0	1	[0]	1	1	[1]
0	0	[0]	[0]	1	1

Table 2.3: Table of potential outcomes of $(T(Z), T(1-Z), Q(Z), Q(1-Z))$ after making the assumptions of randomization and blinded design of the screening trial. Values denoted by "[·]" define a-priori counterfactuals, i.e. quantities which are not only potentially unobserved, but also non-existent.

the decision of a patient whether to follow the recommendation or not, I thus do not define principal strata on quantities which are a-priori counterfactual. Due to the assumption of randomization, one can now interpret $T(Z)$ as the "assigned" diagnostic test and $T(1-Z)$ as the "unassigned" test, regardless of whether $Z = 0$ or $Z = 1$. Consider the following example: e.g. for a patient having being assigned to $Z = 0$, the potential outcome of, say $(Q(T(Z=0)=1)=0), Q(T(1-0)=1)=0))$ describes the behavior of refusing to do the follow-up diagnostic verification, regardless of the diagnostic test outcome. If that patient would have been assigned to Z=1, then the pair $(Q(T(1)=1)=0), Q(T(0)=1)=0))$ would describe the same behavior.

I now further simplify the above table so that it includes only existent quantities. These are summarized in Table 2.4. Indeed, those will be the quantities around which the basic principal strata are built and will be the only quantities involved in any of the estimands of relative diagnostic test accuracy.

	$(T(Z), T(1-Z))$			
	(1,1)	(1,0)	(0,1)	(0,0)
$(Q(T(Z)), Q(T(1-Z)))$	(1,1)	(1,0)	(0,0)	(0,0)
	(1,1)	(1,1)	(1,1)	(1,1)
	(0,0)	(0,0)	(0,0)	(0,0)
	(0,0)	(0,1)	(1,0)	(1,1)

Table 2.4: Table of potential outcomes of $(Q(T(Z), Q(T(1-Z))$ after making the assumptions of randomization and blinded design of the screening trial without considering the a-priori counterfactuals. $T(Z)$ and $T(1-Z)$ represent the outcomes of the assigned and unassigned test respectively, in the randomized screening trial.

Lastly, I now marginalize over the outcome of the unassigned test $T(1-Z)$, and can thus define the principal strata by considering the possible outcomes of the pair $(Q(T(Z) = 1), Q(T(Z) = 0))$, thus generating a partition of the population in four groups:

1. A person who would follow the recommendation and undergo disease status verification if told to do so, and would not undergo it if told not to, would have the pair $(Q(1) = 1, Q(0) = 0)$. Such patients are called "Compliers". In Section 2.1, all patients fell in this stratum.

2. In contrast, a person who would never undergo disease status verification regardless of the test outcome would have $(Q(1) = 0, Q(0) = 0)$. Such patients would be "Refusers."

3. Another case is that of a patient who insists on disease status verification, regardless what the test outcome is. The patient's principal stratum would be characterized by the pair $(Q(1) = 1, Q(0) = 1)$. Such individuals are called "Insisters."

4. Finally, patients who do the opposite of what told to, are called a "Defiers" and have the pair $(Q(1) = 0, Q(0) = 1)$.

Note that in the literature of noncompliance for clinical trials, patients characterized by $(Q(1) = 0, Q(0) = 0)$ and $(Q(1) = 1, Q(0) = 1)$ are usually called "never-takers" and "always-takers" respectively. However, since here there is no active treatment to be "taken" (as it would be for e.g. aspirin pill vs. placebo), the terms refusers and insisters are more appropriate. These terms are introduced in Cuzick et al. (2007) and have

Test result	Observed decision	Compliance (latent) strata
$T(Z) = 1$	$Q(T(Z)) = 1$	$S = (1, 1)$ or $(1, 0)$
$T(Z) = 1$	$Q(T(Z)) = 0$	$S = (0, 0)$ or $(0, 1)$
$T(Z) = 0$	$Q(T(Z)) = 1$	$S = (1, 1)$ or $(0, 1)$
$T(Z) = 0$	$Q(T(Z)) = 0$	$S = (0, 0)$ or $(1, 0)$

Table 2.5: Table of observed decisions in a two-arm, unpaired, screen-positive screening trial for a binary disease with post-screening noncompliance and the corresponding latent stratum membership.

an intuitive interpretation here. Letting S denote the compliance stratum, one thus has: $S = (Q(1), Q(0)) = (1, 0)$ for Compliers (c); $S = (Q(1), Q(0)) = (1, 1)$ for Insisters (i); $S = (Q(1), Q(0)) = (0, 0)$ for Refusers (r) and $S = (Q(1), Q(0)) = (0, 1)$ for Defiers (d). Summarizing the notation for the strata:

$$S = (Q(1), Q(0)) = \begin{cases} (1,0) & \text{complier(c)} \\ (1,1) & \text{insister(i)} \\ (0,0) & \text{refuser(r)} \\ (0,1) & \text{defier(d)} \end{cases}$$

An indicator function, R^S, is introduced to indicate whether the specific belonging to a compliance stratum is point-identified (i.e. unambiguously revealed) ($R^S = 1$) or not ($R^S = 0$). A key point here is that without additional assumptions, the group of subjects with observed value $Q(T(Z) = 1) = 1$ is a mixture of compliers and insisters and thus the individual membership of each patient to each one of this strata is not identifiable. Similarly, for $Q(T(Z) = 0) = 0$, one observes decisions that may equally arise from a mixture of compliers and refusers without being able to identify the individual stratum membership of each patient. Lastly, the decisions from a mixture of defiers and insisters, as well as defiers and refusers, will be observed together. Table 2.5 summarizes these facts.

Note that it is allowed for the compliance status S to be associated with D and U in a (possibly) unknown way. This is a natural consequence of the fact that compliance behavior is likely to be related to common personal background and/or clinical characteristics of the patient.

In addition to the aformentioned SUTVA, and Assumptions 2.2.1 and 2.2.2, several further (realistic) assumptions will help further with the identifiability problems evident from Table 2.5.

Assumption 2.2.3 *"Monotonicity" (Permutt and Hebel 1989; Baker and Lindeman 1994; Imbens and Angrist 1994), i.e. there are no defiers:* $P(S = (0, 1)) = 0$.

This assumption is often made when dealing with noncompliance in clinical trials and is indeed reasonable here. With the monotonicity assumption the second and third rows in Table 2.5 can be simplified and one is able to identify

1. insisters who have switched to the unassigned group upon obtaining a negative test recommendation

2. refusers who have refused to follow a positive recommendation and never showed up to obtain their disease status verification.

Indeed, by making the monotonicity assumption, one can disregard the last row in Table 2.4 and subsequently the lines involving the defiers for the definition of the stratum membership variable S.

In addition, the following assumptions of conditional independence are made.

Assumption 2.2.4 *1.* $R^D \perp\!\!\!\perp D|S, T(Z)$ *- which amounts to the latent ignorability assumption (Baker 1998; Frangakis and Rubin 1999) describing the notion that the revealing of the D status is solely a function of the principal stratum and the test outcome. The interpretation is the following: if the data analyst would know the principal stratum of a patient as well as her (assigned) diagnostic test outcome of a patient, D would not be nonignorably missing.*

2. $R^D \perp\!\!\!\perp Z|S, T(Z)$ - *i.e. there is no direct effect of* Z *on* R^D. *Imbens and Angrist (1994) introduced the term "exclusion restriction" for this type of assumption. This assumption is realistic in the case of a blinded trial.*

Finally, the following assumption concerning the behavior of a patient with a positive diagnostic test is made: A patient who has already received a positive recommendation from her assigned test, would not choose to also undergo the unassigned test. This seems to be a reasonable assumption for an economical behavior of the patient, as in most cases she would have to pay for the unassigned test by himself, outside the protocol. Let k be the number of test recommendations that a patient obtains during the trial (both within and outside the study protocol)

Assumption 2.2.5 *No unnecessary tests are performed:* $P(k > 1|T(Z) = 1) = 0$.

It is important to consider the following remark purveys the main message of the above considered principal stratification modeling and assumptions:

Remark 2.2.1 *In a screening trial as described in Definition 2.2.1], the value of the missingness indicator* R^D *of the disease status* D*, is fully determined by both* T *and* S.

With the help of these assumptions, the likelihood of the true underlying model can be now extended to account for the post-screening noncompliance. To build it, I extend the notation from the previous Section 2.1. Using $s \in \{c, i, r\}$ as a shorthand notation for the three compliance strata of compliers, insisters and refusers, let:

$$\pi_{11}^{s,d} = P(T_1 = 1, T_0 = 1, D, S)$$
$$\pi_{10}^{s,d} = P(T_1 = 1, T_0 = 0, D, S)$$
$$\pi_{01}^{s,d} = P(T_1 = 0, T_0 = 1, D, S)$$
$$\pi_{00}^{s,d} = P(T_1 = 0, T_0 = 0, D, S)$$

Instead of "F", now use the superscript "NC" is used to indicate that the relevant likelihoods, measures of relative accuracy and parameter vectors are under the model with post-screening noncompliance. As in Section 2.1, the first step is to give the likelihood of the true underlying model and consequently to derive the observed data distribution.

Let $\boldsymbol{\pi^{NC}}$ be a vector collecting all of the 24 probabilities for all values of $s \in \{c, i, r\}$ and $d \in \{0, 1\}$: $\boldsymbol{\pi}^{NC} = \left(\pi_{11}^{c,1}, \pi_{10}^{c,1}, \ldots, \pi_{01}^{r,0}, \pi_{00}^{r,0}\right)$. Assume that $\boldsymbol{\pi}^{NC} > \mathbf{0}$, and that all probabilities sum up to one. As argued in the previous section, the vector (T_1, T_0, D, S) defines elementary cells of a contingency table. Even though the stratum membership is not known for all patients and neither are their true D statuses, the aim is still to make inference for $\boldsymbol{\pi^{NC}}$. For a given sample of patients, I again model the random counts $\left(N_{11}^{s,1}, \ldots, N_{00}^{s,0}\right)$ of patients falling within these cells by a multinomial model. Such underlying "true" multinomial model gives rise to the likelihood

$$L_{true}^{NC} \propto \prod_{d \in \{0,1\}} \prod_{s \in \{c,i,r\}} \left[\left(\pi_{11}^{s,d}\right)^{n_{11}^{s,d}} \cdot \left(\pi_{10}^{s,d}\right)^{n_{10}^{s,d}} \cdot \left(\pi_{01}^{s,d}\right)^{n_{01}^{s,d}} \cdot \left(\pi_{00}^{s,d}\right)^{n_{01}^{s,d}} \right]$$

Naturally, the model with full compliance described in Section 2.1 is a special case, as when $P(S = c) = 1$, then $L_{true}^{NC} = L_{true}^{F}$.

To derive the observed data distribution first note that D is missing nonignorably, as R^D is a function of S, which in turn is also missing nonignorably. To see this, consider that per Assumption 2.2.5, a complier with a positive test would not go and obtain the other test. Then, $P(R^S = 1 | S = r, T(Z) = 1, T(1-Z) = 1, Z = 1) = 0$, but $P(R^S = 1 | S = c, T(Z) = 1, T(1-Z) = 1, Z = 1) = 1$.

The observed data distribution is derived from the full model (Little and Rubin (2002)) of the joint distribution of the data and the missingness indicators. Here, there is a general pattern of nonignorable missingness with a known mechanism. The mechanism is indeed known, since, e.g. $P(R^D = 1 | S, D, Z, T(Z), T(1-Z))$ is either 0 or 1 for all patients, i.e. the probability of the missingness indicator is degenerate and does not depend on any further

parameters.

Let $\boldsymbol{Y} = (Y^{obs}, Y^{mis}) = (T(Z), T(1-Z), D, S, Z)$ collect the relevant features of the patients in the trial, where $T(Z)$ is the result on the assigned test and $T(1-Z)$ is the result of the unassigned test. Define a vector of missingness indicator functions $\boldsymbol{R} = (R^{T(Z)}, R^{T(1-Z)}, R^D, R^S, R^Z)$ of the same length as $\boldsymbol{Y}$. Each element of $\boldsymbol{R}$ has the value of 1 if the value of the corresponding variable in $\boldsymbol{Y}$ is observed.

The following Table 2.6 contains all the possible observed data configurations that can arise from the trial (see also Figure 1.3). These are the data which are available to the analyst there are 16 distinct cases:

case	Z	$T(Z)$	$T(1-Z)$	D	S	$R^{T(Z)}$	$R^{T(1-Z)}$	R^D	R^S	R^Z
1	1	1	1	1	? (c or i)	1	1	1	0	1
2	1	1	0	1	? (c or i)	1	1	1	0	1
3	1	0	1	1	i	1	1	1	1	1
4	1	1	1	0	? (c or i)	1	1	1	0	1
5	1	1	0	0	? (c or i)	1	1	1	0	1
6	1	0	1	0	i	1	1	1	1	1
7	1	1	1	?	r	1	1	0	1	1
8	1	1	0	?	r	1	1	0	1	1
9	1	0	1	?	? (c or r)	1	1	0	0	1
10	1	0	0	?	? (c or r)	1	1	0	0	1
11	0	1	?	1	? (c or i)	1	0	1	0	1
12	0	0	1	1	i	1	1	1	1	1
13	0	0	?	?	? (c or r)	1	0	0	0	1
14	0	1	?	?	r	1	0	0	1	1
15	0	0	1	0	i	1	1	1	1	1
16	0	1	?	0	? (c or i)	1	0	1	0	1

Table 2.6: All possible observed data configurations from $(\boldsymbol{Y}, \boldsymbol{R})$. The "?" indicate structurally unknown data.

In more detail, each row entry describing possible data configurations in the screening trial as per Definition 2.2.1 can be explained in the following way:

- Case 1: Patient is randomized to arm 1, has both tests positive, is diseased and has the disease revealed. The patient can be either an "insister" or a "complier" (it is not

known which one).

- Case 2: Patient is randomized to arm 1, has a positive T_1 test and a negative T_0 test. D is revealed. The patient can be either an "insister" or a "complier".

- Case 3: Patient is randomized to arm 1, has a negative T_1 test. The patient revealed her D status (positive) because the unassigned T_0 test was positive. The patient is a known "insister".

- Cases 4, 5 and 6: Patients have the same features as patients 1, 2 and 3 respectively, but for $D = 0$.

- Case 7: Patient is randomized to the arm 1 and has both tests positive. The patient refused to verify D status and is thus a "refuser".

- Case 8: Patient is randomized to arm 1, has T_1 test (the assigned test) positive and T_0 (the unassigned) negative. The patient refused to verify D status and is thus a "refuser".

- Case 9: Patient has the assigned test T_1 negative and the unassigned test T_0 positive. The recommendation was not to do verify D, it is observed that the patient followed that recommendation. Neither the D status, nor S is known, since the patient can be either a "refuser" or a "complier."

- Case 10: Patient was assigned to the arm 1 and has both tests negative. Neither the compliance status, nor the D status is known.

- Case 11: Patient was assigned to arm 0 and has the assigned T_0 positive. Neither the would-be result on the T_1 test, nor the stratum belonging is known. The patient has $D = 1$.

- Case 12: Patient was assigned to arm 0 and has the assigned T_0 negative. The patient did not comply with the protocol and performed the T_1 test, which was positive and thus T_1 test result, the D status (positive) and the stratum - "insister" are known.

- Case 13: Patient has the assigned test T_0 negative. The recommendation was not to do verify D, it is observed that the patient followed that recommendation. Hence, D and S are not known.

- Case 14: Patient has the assigned test (T_0) positive but refused to verify D. It is thus known that the patient is a "refuser" but her D is not known.

- Case 15: Patient is like patient 12 in everything but the D status, ($D = 0$).

- Case 16: Patient is like patient 11 in everything but the D status, ($D = 0$).

Note that some elements of $\boldsymbol{R}$ are redundant, e.g. regardless of the value of Z, T_0 is always observed. R^{T_1} can be re-expressed a function of Z, R^D and R^S, hence can also be dropped from $\boldsymbol{R}$. Obviously, R^Z is always equal to 1. Thus, in the subsequent analysis, let $\boldsymbol{R} = (R^D, R^S)$.

Consider the joint p.m.f. of the patient features data $\boldsymbol{Y}$ and the vector with the missingness indicators $\boldsymbol{R}$, i.e. $f(\boldsymbol{R}, \boldsymbol{Y})$, which following Little and Rubin (2002) I call the "full model." Note that the levels of $(\boldsymbol{R}, \boldsymbol{Y})$ define elementary cells (see, e.g. (Bishop et al., 2007)) of a contingency table, whose counts can be modeled with the multinomial distribution. Some of those are structural zeros, as e.g. the cell defined by $(R^D = 1, S = \text{"}r\text{"}, D, T(Z), T(1-Z), Z)$ is empty, as no "refuser" will ever have the D status revealed. All the cells which contain structural zeros are ignored, as they do not influence the maximum likelihood estimation here. Also, note that Z is included as a level in the cell.

Now suppose an i.i.d sample of n patients is given to the data analyst, such that $(\boldsymbol{R}, \boldsymbol{Y})_1, \ldots (\boldsymbol{R}, \boldsymbol{Y})_n \overset{i.i.d}{\sim} f_{\boldsymbol{R},\boldsymbol{Y}}$. Then, by integrating over the missing data (i.e. collapsing cells) one effectively ends up with cells corresponding to the cases in Table 2.6. Formally,

$$f_{(\boldsymbol{R},\boldsymbol{Y}^{obs})_1,\ldots,(\boldsymbol{R},\boldsymbol{Y}^{obs})_n}(\cdot) = \int\limits_{\boldsymbol{Y}^{mis}} \prod_{j=1}^{n} f(\boldsymbol{r}_j, \boldsymbol{y}_j) d\boldsymbol{Y}^{mis} = \int \cdots \int \prod_{j=1}^{n} f(\mathbf{Y}, \mathbf{R})_j d\mathbf{Y_1^{mis}} \ldots d\mathbf{Y_n^{mis}}$$

Integrating the full model over the missing values (D, S, and the unobserved test outcomes), similarly to the approach in the previous section, the observed data likelihood (conditionally within each arm) is given by

$$L_1^{NC} \propto \left[\prod_d \left(\sum_{s\in\{c,i\}} \pi_{11}^{s,d}\right)^{\sum\limits_{s\in\{c,i\}} n_{11}^{s,d}}\right] \cdot \left[\prod_d \left(\sum_{s\in\{c,i\}} \pi_{10}^{s,d}\right)^{\sum\limits_{s\in\{c,i\}} n_{10}^{s,d}}\right] \cdot \left[\sum_d \pi_{11}^{r,d}\right]^{\sum\limits_d n_{11}^{r,d}} \cdot$$
$$\cdot \left[\sum_d \pi_{10}^{r,d}\right]^{\sum\limits_d n_{10}^{r,d}} \cdot \left[\sum_{s\in\{c,r\}} \sum_d \pi_{01}^{s,d}\right]^{\sum\limits_{s\in\{c,r\}} \sum\limits_d n_{01}^{s,d}} \cdot \left[\sum_s \sum_d \pi_{00}^{s,d}\right]^{\sum\limits_s \sum\limits_d n_{00}^{s,d}} \cdot \left[\prod_d \left(\pi_{01}^{i,d}\right)^{n_{01}^{i,d}}\right]$$

$$L_0^{NC} \propto \left[\prod_d \left(\sum_{s\in\{c,i\}} (\pi_{11}^{s,d} + \pi_{01}^{s,d})\right)^{\sum\limits_{s\in\{c,i\}} (n_{11}^{s,d} + n_{01}^{s,d})}\right] \cdot \left[\prod_d \left(\pi_{10}^{i,d}\right)^{n_{10}^{i,d}}\right] \cdot$$
$$\cdot \left[\left(\sum_d \left(\pi_{10}^{c,d} + \pi_{10}^{r,d}\right) + \sum_s \sum_d \pi_{00}^{s,d}\right)^{\sum_d \left(n_{10}^{c,d} + n_{10}^{r,d}\right) + \sum_s \sum_d n_{00}^{s,d}}\right] \cdot$$
$$\left[\sum_d \left(\pi_{11}^{r,d} + \pi_{01}^{r,d}\right)^{\sum_d \left(n_{11}^{r,d} + n_{01}^{r,d}\right)}\right]$$

It is convenient to extend the notation for the parameters from Section 2.1 to include the stratum and thus now have

$$p_{T(Z),T(1-Z)|Z}^{s,d} = P\left(T(Z), T(1-Z), D, S | Z = z\right)$$

For convenience, the likelihood (conditionally within each arm) can be then re-expressed so that

$$\begin{aligned} L_1^{NC} \propto & \left(p_{11|1}^{*,1}\right)^{n_{11|1}^{*,1}} \cdot \left(p_{10|1}^{*,1}\right)^{n_{10|1}^{*,1}} \cdot \left(p_{01|1}^{i,1}\right)^{n_{01|1}^{i,1}} \cdot \left(p_{11|1}^{*,0}\right)^{n_{11|1}^{*,0}} \cdot \left(p_{10|1}^{*,0}\right)^{n_{10|1}^{*,0}} \cdot \\ & \left(p_{01|1}^{i,0}\right)^{n_{01|1}^{i,0}} \cdot \left(p_{11|1}^{r,*}\right)^{n_{11|1}^{r,*}} \cdot \left(p_{10|1}^{r,*}\right)^{n_{10|1}^{r,*}} \cdot \left(p_{01|1}^{*,*}\right)^{n_{01|1}^{*,*}} \cdot \left(p_{00|1}^{*,*}\right)^{n_{00|1^{*,*}}} \\ L_0^{NC} \propto & \left(p_{1*|0}^{*,1}\right)^{n_{1*|0}^{*,1}} \cdot \left(p_{01|0}^{i,1}\right)^{n_{01|0}^{i,1}} \cdot \left(p_{00|0}^{*,*}\right)^{n_{00|0}^{*,*}} \cdot \left(p_{1*|0}^{r,1}\right)^{n_{1*|0}^{r,1}} \cdot \left(p_{01|0}^{i,0}\right)^{n_{01|0}^{i,0}} \cdot \\ & \left(p_{1*|0}^{*,0}\right)^{n_{1*|0}^{*,0}} \end{aligned}$$

Again, "$*$" is used to denote that either a test result, the D status or the compliance stratum does not factor out of the likelihood, i.e. is not point-identifiable. All parameters from the above equations are collected in a vector $\boldsymbol{p}^{NC}$.

From Table 2.2, one can easily see the complication which arises from the post-screening noncompliance. As an example, if one takes the first row in both models, it is clear that compliers who have both diagnostic tests positive are never point-identified. Such group of patients will always be observed together with a group of insisters who also have both tests positive. However, as before, the nestedness does give us deeper information on the underlying vector of parameters $\boldsymbol{\pi}^{NC}$. Considering the last three columns of the table: From rows $q = 3, 6, 12, 15$ due to the noncompliance in the study, one can fully identify these cells, together with the stratum (insisters). This is quite important, as these parameters can then be used to identify (some) compliers from cells $q = 2, 5, 11, 16$. Similarly, from rows $q = 7, 8, 14$ the stratum (refusers) is known, but the D statuses are not.

Partitioning the population in the three principal strata has implications for the measures of accuracy which are of interest - $rFPR$ and $rTPR$. While both were identifiable in the case of the model with full compliance (Section 2.1), by including S in the analysis, the

following holds for $rTPR$ and $rFPR$

$$rTPR = \frac{P(T_1 = 1|D = 1)}{P(T_0 = 1|D = 1)} = \frac{P(T_1 = 1, D = 1)}{P(T_0 = 1, D = 1)} = \frac{\pi_{11}^{c,1} + \pi_{10}^{c,1} + \pi_{11}^{i,1} + \pi_{10}^{i,1} + \pi_{11}^{r,1} + \pi_{10}^{r,1}}{\pi_{11}^{c,1} + \pi_{01}^{c,1} + \pi_{11}^{i,1} + \pi_{01}^{i,1} + \pi_{11}^{r,1} + \pi_{01}^{r,1}}$$
$$rTPR = \frac{P(T_1 = 1|D = 1)}{P(T_0 = 1|D = 1)} = \frac{P(T_1 = 1, D = 1)}{P(T_0 = 1, D = 1)} = \frac{\pi_{11}^{c,0} + \pi_{10}^{c,0} + \pi_{11}^{i,0} + \pi_{10}^{i,0} + \pi_{11}^{r,0} + \pi_{10}^{r,0}}{\pi_{11}^{c,0} + \pi_{01}^{c,0} + \pi_{11}^{i,0} + \pi_{01}^{i,0} + \pi_{11}^{r,0} + \pi_{01}^{r,0}}$$

Since one never observes the true disease status D of refusers, the above ratio is not identifiable from the observed data distribution. Analogously, the same holds for $rFPR$. Note that if there were only compliers, then the ratio would be point-identifiable from the observed data, as shown in the previous section, since $P(S = \text{“}c\text{”}) = 1$.

All is not lost however, as there are versions of the $rTPR$ and $rFPR$ which are indeed of interest to the data analyst and *are* identifiable. One such version is the relative true positive and relative false positive rates based on the joint stratum of insisters and compliers:

Definition 2.2.2 *Relative ratios based on the joint stratum of compliers and insisters:*

$$rTPR_{i+c} = \frac{P(R^D(Z) = 1, T(Z) = 1, D = 1|Z = 1)}{P(R^D(Z) = 1, T(Z) = 1, D = 1|Z = 0)} = \frac{\pi_{11}^{c,1} + \pi_{10}^{c,1} + \pi_{11}^{i,1} + \pi_{10}^{i,1}}{\pi_{11}^{c,1} + \pi_{01}^{c,1} + \pi_{11}^{i,1} + \pi_{01}^{i,1}} \quad (2.1)$$
$$rFPR_{i+c} = \frac{P(R^D(Z) = 1, T(Z) = 1, D = 0|Z = 1)}{P(R^D(Z) = 1, T(Z) = 1, D = 0|Z = 0)} = \frac{\pi_{11}^{c,0} + \pi_{10}^{c,0} + \pi_{11}^{i,0} + \pi_{10}^{i,0}}{\pi_{11}^{c,0} + \pi_{01}^{c,0} + \pi_{11}^{i,0} + \pi_{01}^{i,0}} \quad (2.2)$$

These ratios are indeed point-identifiable from the data as the following proposition shows (with obvious adjustments, it holds for $rFPR_{i+c}$):

Proposition 2.2.3 *$rTPR_{i+c}$ can be re-expressed from the observable data and it is thus point-identifiable*

$$rTPR_{i+c} = \frac{P(R^D(Z) = 1, T(Z) = 1, D = 1|Z = 1)}{P(R^D(Z) = 1, T(Z) = 1, D = 1|Z = 0)} \quad (2.3)$$

Of even greater interest might be stratum-specific ratios for each stratum separately — the insisters, the compliers and the refusers. However, as mentioned before, the data analyst

would never obtain data for the true disease status of any refusers. Moreover, it can be argued that since they refuse to verify their diagnosis, any screening policy implementation would be without an effect on them. Hence, the focus for the remaining part of the thesis would be on insisters, compliers and their joint stratum. For the compliers-specific rTPR one has:

Definition 2.2.4 *Compliers-specific $rTPR$ and $rFPR$:*

$$rTPR_c = \frac{P(T_1 = 1|D = 1, S = c)}{P(T_0 = 1|D = 1, S = c)} = \frac{P(T_1 = 1, D = 1, S = c)}{P(T_0 = 1, D = 1, S = c)} = \frac{\pi_{11}^{c,1} + \pi_{10}^{c,1}}{\pi_{11}^{c,1} + \pi_{01}^{c,1}}$$
$$rFPR_c = \frac{P(T_1 = 1|D = 0, S = c)}{P(T_0 = 1|D = 0, S = c)} = \frac{P(T_1 = 1, D = 0, S = c)}{P(T_0 = 1, D = 0, S = c)} = \frac{\pi_{11}^{c,0} + \pi_{10}^{c,0}}{\pi_{11}^{c,0} + \pi_{01}^{c,0}}$$

In a similar way, one can define $rTPR_i$ and $rFPR_i$ for insisters. Frangakis and Rubin (2002) show that any comparison between two potential outcomes within same stratum does represent a causal effect and it is worth noting the following, in difference to the compliers-only and insisters-only strata:

$$\frac{P(R^D = 1|S = i, D = 1, Z = 1)}{P(R^D = 1|S = i, D = 1, Z = 0)} = \frac{\pi_{11}^{i,1} + \pi_{10}^{i,1} + \pi_{01}^{i,1}}{\pi_{11}^{i,1} + \pi_{01}^{i,1} + \pi_{10}^{i,1}} = 1 \tag{2.4}$$

The reasoning behind this is simple: an insister will always obtain the disease status verification from the assigned test, or try to obtain via the unassigned test. Hence, exposing an insister to either group has a relative causal effect of 1 on the disease status revealing outcome.

It is very important to note that this does *not* mean that both diagnostic tests are equally accurate for the stratum of insisters. Indeed, the ratio in expression (2.4) need not be equal to $rTPR_i$. Indeed, although the contrast here has a causal interpretation, it bears no useful information regarding the relative sensitivity of the diagnostic tests within the set of insisters. By the same token, it is easy to show that $\frac{P(R^D=1|S=i+c,D=1,Z=1)}{P(R^D=1|S=i+c,D=1,Z=0)} = 1$, which also not necessarily equal to $rTPR_{i+c}$.

This reasoning shows the advantage of using principal strata to model the compliance behavior here (as opposed to doing analysis solely based on the patients whose disease status was recorded): it allows us to disentangle the causal effect of the exposure for a specific screening policy to the relative sensitivity of the two diagnostic tests.

By formulating the whole screening trial as a comparison between two competing disease-revealing mechanisms, one is thus able to show the limitations of using the causal effect of Z on R^D as means to make overall statements about the relative test accuracy. Thus, in the case of post-screening noncompliance, any policy implementation in regards to which test should be administered to the public should carefully disentangle the effect of exposure (to one of the two tests) on the disease revealing from the actual relative accuracy of the diagnostic tests. These two would coincide *only* for the compliers.

After introducing the stratum-specific ratios and showing that $rTPR_{i+c}$ and $rFPR_{i+c}$ are indeed point-identifiable from the observed data distribution, one has to consider that the stratum-specific ratios are not point-identified, but rather partially identified (cf. Manski (2003)). Recall that the stratum membership is not known a-priori, so that one can not directly observe the data coming from the individual probability mass functions in the nominator and denominator of the two individual sensitivity ratios - $rTPR_i$ and $rTPR_c$.

To point-identify $rTPR_c$ or $rTPR_i$, additional assumptions on the probability of belonging to the different strata $P(S = s)$ are needed. Indeed, compliance behavior can be modeled as a function of a vector of some baseline covariates $\mathbf{X}$.

In absence of such baseline covariates and/or a meaningful model for compliance, one can formulate nonparametric bounds for $rTPR_c$ and thus partially identify it, as opposed to point-identifying it through an assumed model for $P(S)$. This has the advantage of making inference for the measures of interest more credible (Manski (2003)), since there are no modeling assumptions about $P(S)$. Nonparametric bounds for $rFPR_c$, $rTPR_i$ and $rFPR_i$ can be constructed in an analogous way. Nonparametric bounds here are understood as bounds that do not depend on specific modeling assumptions for $P(S)$.

The next proposition gives a lower and an upper bound for $rTPR_c$. The bounds hold regardless whether the diagnostic tests are nested or not, but results in the next section will show, the nested structure allows for the bounds' estimation.

Proposition 2.2.5 $\left(\frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}} \leq 1\right) \Leftrightarrow \left(\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}} \leq rTPR_c \leq \frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}}\right)$

See the Appendix for a proof.

Recall that a sensitivity ratio bigger than 1 indicates that the T_1 test has a higher probability to correctly predict the presence of a disease than the T_0 test. It is then natural to conduct hypothesis testing for $H_0 : rTPR_c \leq 1$ and indeed the result in Proposition 2.2.5 is useful for that.

Corollary 2.2.6 *From Proposition 2.2.5 it follows*

$$min\left(\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}}, \frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}}\right) \leq \frac{\pi_{11}^{c,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{01}^{c,1}} \leq max\left(\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}}, \frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}}\right) \tag{2.5}$$

$$min\left(\frac{\pi_{10}^{i,1}}{\pi_{01}^{i,1}}, \frac{\pi_{11}^{i,1}+\pi_{11}^{c,1}+\pi_{10}^{i,1}}{\pi_{11}^{i,1}+\pi_{11}^{c,1}+\pi_{01}^{i,1}}\right) \leq \frac{\pi_{11}^{i,1}+\pi_{10}^{i,1}}{\pi_{11}^{i,1}+\pi_{01}^{i,1}} \leq max\left(\frac{\pi_{10}^{i,1}}{\pi_{01}^{i,1}}, \frac{\pi_{11}^{i,1}+\pi_{11}^{c,1}+\pi_{10}^{i,1}}{\pi_{11}^{i,1}+\pi_{11}^{c,1}+\pi_{01}^{i,1}}\right) \tag{2.6}$$

While it is the aim for $rFPR$ and $rTPR$ to have a causal interpretation, it is important to justify why exactly I employ the principal stratification method. In fact, Pearl (2011) cautions against the usage of principal strata without them having a meaningful interpretation. Indeed, an alternative here would be to adjust for noncompliance simply by considering the patients whose disease status was recorded. This can be still be done without introducing principal strata. Mealli and Mattei (2012) discuss the objections raised by Pearl (2011) and argue that the PS methodology has its justification, in particular they write that "Principal stratification does not always answer the causal question of primary interest, but it often provides useful insights..."

Indeed, for the problem discussed in this manuscript, I believe that using PS does provide useful insights on the data generating process is justified for two reasons: First, it can be argued that the primary objective for a screening trial which screens for a disease such as, e.g. prostate cancer, is to provide evidence for policy makers to help them decide which one of the two competing procedures should be implemented (and funded) in a health care system. By the same token, this is relevant for, e.g. health insurance companies, which would arguably decide to fund only one of

several competing screening procedures. These type of policy considerations are indeed important when an existing (and cheap) diagnostic test is compared to a new test (which is more expensive) in terms of accuracy.

Furthermore, one can argue that compliance behavior is likely to be related to some important common background characteristics among the patients, which were unobserved (see e.g. (Gareen, 2007, p. 347) for the special case of noncompliance in screening trials). As an example, consider a case where the old test performs better (in terms of relative accuracy) than the new one for the group of insisters, but worse for the group of compliers. Such information would have direct policy implications - arguably, the health policy maker would choose to fund the one test which affects a bigger group of the population. That could very well be the group of compliers and stratum-specific ratios would be of bigger interest in this case.

The above described line of thought of a policy maker, could be opposed to the point of a view of a clinician. A clinician does not know a-priori, whether the patient is, e.g. a complier or an insister. Hence, the clinician cannot intervene on the compliance status and give the "better" test based on a stratum-specific evaluation of their relative diagnostic accuracy.

Indeed, she might be interested in the overall effect of exposing a patient to either one of the two tests, i.e. in the causal effect for a given patient j: $\left(R_j^D(1) \quad \text{vs.} \quad R_j^D(0)\right)$ or the causal effect of assignment on the disease revealing within the subpopulation with certain disease status level $\left(R_j^D(1) \quad \text{vs.} \quad R_j^D(0)\right) \big| D$ (e.g. among already diseased patients), without the need to know exactly *how* the patient obtains the disease status verification, as long as the verification *is* indeed obtained. But in this case, the data analyst would lose the more detailed information about stratum-specific effects. The PS method gives us the tools to model these and the nested test property allows to conduct inference for these stratum-specific bounds.

The main message here is that by employing the PS method, the analyst can model and estimate both the stratum-specific and union based metrics. Indeed, by using the PS method, the data analyst does not lose information, but gains more.

For the rest of this manuscript, the focus will lie on $rTPR_{i+c}$ and subsequently on $rTPR_c$. All the arguments for the relative false positive rates and the insisters-specific ratios follow in

an analogous way. The next chapter shows how to estimate these relevant measures of relative accuracy and how to construct appropriate statistical test hypotheses for them.

Thus, as long as it can be reasonably assumed that the joint stratum of insisters and compliers represent the bigger part of the population, $rTPR_{i+c}$ and $rFPR_{i+c}$ are useful measures for the relative accuracy of the two screening tests.

Chapter 3

Inference

3.1 Inference for the model with full compliance

The vector of parameters for the full compliance Multinomial model — $\boldsymbol{p}^F$, can be estimated by using MLE. Since the target of inference is $rTPR^F$, recall that $rTPR^F = \left(\pi_{11}^{c,1} + \pi_{10}^{c,1}\right) / \left(\pi_{11}^{c,1} + \pi_{01}^{c,1}\right)$.

Thus, it is convenient to use a continuous function of $\boldsymbol{p}^F$ to re-express $rTPR^F$ and hence the estimator would be:

$$rTPR(\widehat{\boldsymbol{p}}^F) = (\widehat{p}_{11|1}^{\,c,1} + \widehat{p}_{10|1}^{\,c,1})/\widehat{p}_{1,*|0}^{\,c,1} \tag{3.1}$$

By the functional invariance property of the MLE (cf. Casella and Berger (2002)), this estimator is also MLE. To find its asymptotic distribution, first note that this estimator is a function from parameters in each arm. One can re-parametrize the original problem in terms of one big Multinomial model and by employing the multivariate central limit theorem and subsequently the delta method, one can find its asymptotic distribution (see Rao (1973), Agresti (2002) and Bishop et al. 2007 for a specific application for the Multinomial distribution and van der Vaart (2000) for a general discussion on the delta method).

Proposition 3.1.1 *Under the assumptions in Definition 2.1.1 and the assumed i.i.d sampling*

scheme, as $n \to \infty$*:*

$$\sqrt{n}\left(rTPR^F\left(\widehat{\boldsymbol{p}}^F\right) - rTPR^F\left(\boldsymbol{p}^F\right)\right) \xrightarrow{D} N\left(0, \sigma^F_{rTPR}(\boldsymbol{p}^F)^2\right)$$

As a result, for large n, it holds that $rTPR(\widehat{\boldsymbol{p}}^F) \approx N(rTPR^F, \left(\sigma^F_{rTPR}(\boldsymbol{p}^F)\right)^2/n)$. Details on the inferential procedure and the full form of the variance can be found in the appendix.

It can be shown that if one has the same structure of the screening trial (two-arm, unpaired, screen positive with two binary diagnostic tests screening for a binary disease), *except* for the nested property of T_0 within T_1, the expression for the asymptotic variance of the estimator will be the same. The result is not surprising: While one can point-identify more parameters of $\boldsymbol{\pi^F}$ when there is a nested diagnostic test, for the estimation of $rTPR$ under full compliance, the nested test property of the trial does not add any additional information when compared to a *nonnested* unpaired design of a two-arm, screen-positive screening trial (see e.g. Pepe and Alonzo (2001) for more general details on such a design).

However, apart from estimation, a relevant point of analysis of diagnostic test accuracy is the testing for relative diagnostic superiority vs. inferiority in terms of $rTPR$ and $rFPR$. Now, by using the nested property, it is possible to construct an additional test statistic which does indeed exploit the additional information from the nestedness. In particular, consider a relevant hypothesis for the researcher of the type

$$H_0 : rTPR^F \leq 1 \qquad \text{vs.} \qquad H_1 : rTPR^F > 1$$

and note that this statement is equivalent to the hypotheses statement

$$H_0 : (\pi^{c,1}_{10} - \pi^{c,1}_{01}) \leq 0 \qquad \text{vs.} \qquad H_1 : (\pi^{c,1}_{10} - \pi^{c,1}_{01}) > 0$$

With the help of Table 2.2, one can construct two different test statistics. The first test statistic, $T^F_1(\widehat{\boldsymbol{p}}^F)$, follows the construction of $rTPR(\widehat{\boldsymbol{p}}^F)$ and thus does not employ additional information coming from the nested property. On the other hand, the second test statistic, $T^F_2(\widehat{\boldsymbol{p}}^F)$, exploits

the nestedness property and that more parameters of $\boldsymbol{\pi}^F$ factor out of $L_1^F \cdot L_0^F$ and become point-identifiable, e.g. $\widehat{p}_{11|1}^{\,c,0}$. If there were no nestedness, this additional information (i.e. identifiability) on $\boldsymbol{\pi}^{\boldsymbol{F}}$ would not have been available.

$$T_1^F(\widehat{\boldsymbol{p}}^F) = \widehat{p}_{11|1}^{\,c,1} + \widehat{p}_{10|1}^{\,c,1} - \widehat{p}_{1,*|0}^{\,c,1}$$
$$T_2^F(\widehat{\boldsymbol{p}}^F) = \widehat{p}_{10|1}^{\,c,1} - \widehat{p}_{11|1}^{\,c,0} - \widehat{p}_{01|1}^{\,c,*} + \widehat{p}_{1,*|0}^{\,c,0}$$

Using similar techniques as above, one can find the asymptotic distributions of scaled versions of $T_1^F(\widehat{\boldsymbol{p}}^F)$ and $T_2^F(\widehat{\boldsymbol{p}}^F)$ under the null and the alternative hypotheses (see Appendix for details on deriving the corresponding expressions for the asymptotic variances of these test statistics). Since both statistics are complicated (multidimensional) functions of the original parameters and the assignment probability, one could compare them through their power functions for given sets of parameter values. Analytic comparison between the two power functions is difficult and neither test statistic makes the other not admissible. This issue is discussed in the last chapter.

Throughout, I have omitted describing the modeling and inference for $rFPR^F$. Indeed, the construction of the estimators and relevant hypothesis tests for $rFPR^F$ are analogous are briefly described in the appendix.

3.2 Inference for the model with post-screening noncompliance

3.2.1 Large sample hypothesis tests for the stratum-specific measures of relative accuracy

Similar to the approach in Section 2.1, I now introduce functions of $\boldsymbol{p}^{NC}$ which can be re-expressed in terms $\boldsymbol{\pi}^{NC}$, i.e. of the parameters of the true underlying model with noncompliance. As before,

the goal is to find an MLE for $rTPR_{i+c}$. Consider that

$$rTPR_{i+c}^{NC}(\widehat{\boldsymbol{p}}^{NC}) = \frac{\left(\widehat{p}_{11|1}^{*,1} + \widehat{p}_{10|1}^{*,1}\right)}{\widehat{p}_{1,*|0}^{*,1}} \tag{3.2}$$

which, by the invariance property of MLE, gives a consistent estimator for $rTPR_{i+c}$. Again, using standard results for asymptotics of MLE, the following proposition holds

Proposition 3.2.1 *Under the assumed i.i.d sampling scheme as $n \to \infty$:*

$$\sqrt{n}\left(rTPR_{i+c}^{NC}\left(\widehat{\boldsymbol{p}}^{NC}\right) - rTPR_{i+c}^{NC}\left(\boldsymbol{p}^{NC}\right)\right) \xrightarrow{D} N\left(0, \sigma_{rTPR_{i+c}}^{NC}(\boldsymbol{p}^{NC})^2\right)$$

So that for large n:

$$rTPR_{i+c}^{NC}(\widehat{\boldsymbol{p}}^{NC}) \approx N\left(rTPR_{i+c}^{NC}(\boldsymbol{p}^{NC}), \frac{\left(\sigma_{rTPR_{i+c}}^{NC}(\boldsymbol{p}^{NC}\right)^2}{n}\right)$$

Details for the exact expressions and a proof can be found in the appendix.

To test the relevant hypothesis of inferiority of the new test vs. the old one, I proceed as in Section 3.1 and note that the composite hypothesis

$$H_0 : rTPR_{i+c} \leq 1 \qquad \text{vs.} \qquad rTPR_{i+c} > 1$$

is equivalent to the hypothesis

$$H_0 : \left(\pi_{10}^{c,1} + \pi_{10}^{i,1}\right) - \left(\pi_{01}^{c,1} + \pi_{01}^{i,1}\right) \leq 0 \qquad \text{vs.} \qquad \left(\pi_{10}^{c,1} + \pi_{10}^{i,1}\right) - \left(\pi_{01}^{c,1} + \pi_{01}^{i,1}\right) > 0 \tag{3.3}$$

Here, too, using the nested diagnostic test property, one can construct two competing (in terms of statistical power) test statistics to conduct asymptotic tests. The first test statistic, $T_1^{NC}(\widehat{\boldsymbol{p}}^{NC})$, follows the construction of $rTPR_{i+c}^{NC}(\widehat{\boldsymbol{p}}^{NC})$ and thus does not employ additional information coming from the nested property. In contrast, $T_2^{NC}(\widehat{\boldsymbol{p}}^{NC})$, exploits the fact that from patients in arm 1, the

sum of $\left(\pi_{11}^{c,1}+\pi_{11}^{i,1}\right)$ becomes point-identifiable. A major difference from Section 2.1 is that sum cannot be "split" without further assumptions (e.g. a regression model) for $P(S)$. Thus, unlike in Section 2.1, here $\pi_{11}^{c,1}$ is *not* point-identifiable.

$$T_1^{NC}(\widehat{\boldsymbol{p}}^{NC}) = \widehat{p}_{11|1}^{\,*,1} + \widehat{p}_{10|1}^{\,*,1} - \widehat{p}_{1,*}^{\,*,1}$$
$$T_2^{NC}(\widehat{\boldsymbol{p}}^{NC}) = \widehat{p}_{10|1}^{\,*,1} + \widehat{p}_{1*|0}^{\,r,1} + \widehat{p}_{1*|0}^{\,*,0} - \widehat{p}_{01|1}^{\,*,*} - \widehat{p}_{11|1}^{\,*,0} - \widehat{p}_{01|1}^{\,i,0} - \widehat{p}_{11|1}^{\,r,*} - \widehat{p}_{01|1}^{\,i,1}$$

Using similar techniques as above, I can find the asymptotic distributions of $T_1^{NC}(\widehat{\boldsymbol{p}}^{NC})$ and $T_2^{NC}(\widehat{\boldsymbol{p}}^{NC})$ under the null and the alternative hypotheses (details of derivations for the estimator and the test statistics of this section, as well as their asymptotic variances, can be found in the appendix). As with the model under full compliance, analytic comparison between the two powers appears to be difficult and no test makes the other not admissible. Empirical size and power with simulated data evaluations will be discussed in the next Chapter 4.

Apart from these point-identified measures of relative accuracy, the previous chapter discussed stratum-specific relative measures of accuracy, namely $rTPR_c$, $rTPR_i$ and their stratum-specific relative false positive rate counterparts. It was shown that these stratum-specific measures are not point-identified from the observed data distribution. Rather, the analyst can point-identify some bounds for these stratum-specific measures which can still be of interest, e.g. for the policy maker, even if the focus lies on a point identified measure (e.g. $rTPR_{i+c}$). To proceed with the inference for these bounds, first consider the following two functions of the parameter vector $\boldsymbol{\pi}^{NC}$ the estimation of which would give us the two bounds on the logarithmic scale:

$$g_1\left(\boldsymbol{\pi}^{NC}\right) = \log\left(\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}}\right) \qquad g_2\left(\boldsymbol{\pi}^{NC}\right) = \log\left(\frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}}\right) \tag{3.4}$$

Using the estimated parameters of the observed data distribution $\widehat{\boldsymbol{p}}^{NC}$, the estimators of these two functions are as follows:

$$g_1\left(\widehat{\boldsymbol{p}}^{NC}\right) = \log\left(\frac{\widehat{p}^{\,*,1}_{10|1} - \widehat{p}^{\,i,1}_{01|0}}{\widehat{p}^{\,*,1}_{1*|0} - \widehat{p}^{\,i,1}_{01|1} - \widehat{p}^{\,*,1}_{11|1}}\right) \tag{3.5}$$

$$g_2\left(\widehat{\boldsymbol{p}}^{NC}\right) = \log\left(\frac{\widehat{p}^{\,*,1}_{11|1} + \widehat{p}^{\,*,1}_{10|1} - \widehat{p}^{\,i,1}_{01|0}}{\widehat{p}^{\,*,1}_{1*|0} - \widehat{p}^{\,i,1}_{01|1}}\right) \tag{3.6}$$

At this stage it is important to observe that the "nestedness" of the tests allow us to identify $g_1\left(\boldsymbol{\pi}^{NC}\right)$. This is so, because one can use information from arm 1 to (by randomization) estimate $p^{*,1}_{11|1}$. *Without* nestedness, all other assumptions and properties of the trial being equal, one can still compute (and estimate) $g_2\left(\boldsymbol{\pi}^{NC}\right)$. This is so, because one would be able to estimate $\left(p^{*,1}_{11|1} + p^{*,1}_{10|1}\right)$, even without being able to split the sum of these two elements.

As it will be shown later, this would still allow us to conduct relevant hypothesis tests even without the additional information coming from the nestedness. However, nestedness is needed when one wants to construct the bounds, as both g_1 and g_2 are needed as per the expressions in inequalities (2.5).

To find the asymptotic distribution of these two estimators, one can use the delta method as done for the previous estimators and given the smooth nature of $g_1(\cdot)$ and $g_2(\cdot)$ obtain weakly consistent estimators. Indeed, the choice of taking a logarithmic function was inspired by the fact that in simulations, for some specific population parameters of the underlying population, it showed faster convergence to the standard normal distribution. One could just as easily take the square root function or any other smooth function. Again by employing the delta method, for large n, one can find the asymptotic distributions of the two estimators. For large n, it holds that $g_1\left(\widehat{\boldsymbol{p}}^{NC}\right) \approx N\left(g_1\left(\boldsymbol{p}^{NC}\right), \frac{\sigma^2_{g_1}(\boldsymbol{p}^{NC})}{n}\right)$ and $g_2\left(\widehat{\boldsymbol{p}}^{NC}\right) \approx N\left(g_2\left(\boldsymbol{p}^{NC}\right), \frac{\sigma^2_{g_2}(\boldsymbol{p}^{NC})}{n}\right)$. The sampling variances $\sigma_{g_1^2}(\boldsymbol{p}^{NC})/n$ and $\sigma_{g_2^2}(\boldsymbol{p}^{NC})/n$ have rather complicated forms and I use `Mathematica 7.0` to derive them. They are not given here, however, the procedure to obtain them are analogous to the ones described for, e.g. the variance of $rTPR^{NC}_{i+c}\left(\widehat{\boldsymbol{p}}^{NC}\right)$ in the Appendix by using the delta method.

As already mentioned, inference for $rFPR_c$ follows in an analogous manner. One can define new functions $g_1(\cdot)^{(rFPR)}$ and $g_2(\cdot)^{(rFPR)}$, such that:

$$g_1^{rFPR}\left(\widehat{\boldsymbol{p}}^{NC}\right) = \log\left(\frac{\widehat{p}_{10|1}^{\,*,0} - \widehat{p}_{01|0}^{\,i,0}}{\widehat{p}_{1*|0}^{\,*,0} - \widehat{p}_{01|1}^{\,i,0} - \widehat{p}_{11|1}^{\,*,0}}\right) \tag{3.7}$$

$$g_2^{rFPR}\left(\widehat{\boldsymbol{p}}^{NC}\right) = \log\left(\frac{\widehat{p}_{11|1}^{\,*,0} + \widehat{p}_{10|1}^{\,*,0} - \widehat{p}_{01|0}^{\,i,0}}{\widehat{p}_{1*|0}^{\,*,0} - \widehat{p}_{01|1}^{\,i,0}}\right) \tag{3.8}$$

To ease notation, I drop the $(rFPR)$ superscript, when it is clear from the context. Note that, e.g. $g_1(\cdot)$ and $g_1^{(rFPR)}$ are similar and they just select different cells of $\boldsymbol{p}^{NC}$ to obtain the corresponding measure of relative accuracy.

Before moving on to actual inference on the bounds expressed in inequalities in expression (2.5), note that one can use, e.g. $g_2(\widehat{\boldsymbol{p}}^{NC})$ as a test statistic for an asymptotic test of the type

$$H_0 : rTPR_c \leq 1 \quad \text{vs. } H_1 : rTPR_c > 1$$

This hypothesis would be of interest for the health policy maker to judge the relative accuracy of the two tests and subsequently take a decision on which (one) test to implement based on the largest elementary subpopulation (i.e. *not* the union of compliers and insisters, but either insisters or compliers only). While $rTPR_c$ is not identifiable from the observed data distribution, from Proposition 2.2.5, one can use $g_2(\widehat{\boldsymbol{p}}^{NC})$ conduct the test. In other words, the above hypothesis test for $rTPR_c$ is equivalent to testing the following:

$$H_0 : g_2(\boldsymbol{p}^{NC}) \leq 0 \quad \text{vs. } H_1 : g_2(\boldsymbol{p}^{NC}) > 0$$

The asymptotic distribution of the test statistic under the H_0 can be easily obtained via the delta-method under the restriction $\pi_{10}^{c,1} = \pi_{01}^{c,1}$ as $g_2^{H_0}(\widehat{\boldsymbol{p}}^{NC}) \approx N\left(0, \frac{\sigma_{g_2}^{H_0}(\boldsymbol{p}^{NC})^2}{n}\right)$. As outlined above, a test for $rFPR_c$ follows analogously with minor adjustments to $g_2(\boldsymbol{p}^{NC})$ to take into account the cells with $D = 0$ instead of $D = 1$. It is indeed of great added value that even though the compliers and insisters are *not* identified from the observed data distribution (apart from insisters who have

received a negative result on the assigned diagnostic test and have subsequently underwent the unassigned diagnostic test), one is able to conduct a useful hypothesis test on the stratum-specific measures of relative accuracy for insisters and for compliers separately.

3.2.2 Large sample confidence interval for the partially identified stratum-specific measure of relative accuracy

Overview and related literature

After having established some useful estimators for the elements which are necessary to estimate the two bounds, I now discuss an analytic approach for the construction of asymptotic confidence interval (CI).

The notion of *partial identification* has "had a long but sparse history in statistical theory" as Manski (2003) notes. Consider a parametric model, in which the parameter of interest does not factor out of the likelihood (e.g. $\pi_{11}^{c,1}$ in the model with post-screening noncompliance). That population parameter is then not *point-identifiable* from the observed data distribution. However, it may be *partially identified*, via some point-identifiable bounds.

Indeed, whenever a parameter of interest is not point-identifiable from the observed data distribution, one is confronted with the choice of either making further assumptions (which are possibly unverifiable from the observed data) in order to point-identify it or to refrain from making them while accepting the fact that any inferential conclusions would be done for a set of possibly population parameters, rather than a point.

Manski (2003) discusses this trade-off in the paradigm of what he calls *The Law of Decreasing Credibility*, which he postulates as "The credibility of inference decreases with the strength of the assumptions maintained." The way this concept applies to the problem of identifying population parameters in the case of outlined screening trial in this thesis is the following: Naturally, one can always consider a compliance-specific model, which, on the basis of some external covariates $\boldsymbol{X}$ models e.g. $E(S|\boldsymbol{X})$ via a regression model. Indeed, some authors have considered such models in the case of noncompliance in a clinical trial (see e.g. Barnard et al. (2003) and Bartolucci and

Farcomeni (2013)). By continuing this logic, if we do assume a specific model for the compliance, then we could model the probability of belonging to a specific stratum. In this case, e.g. $rTPR_c$ would become point-identifiable, as we would be able to conduct inference for $\pi_{11}^{c,1}$, provided that the compliance model is correct. The latter is an assumption and it is often untestable from the observed data. Considering the Law of Decreasing Credibility, inference for $rTPR_c$ will vary in its degree of credibility, depending on how strong (or "plausible") our assumptions for the compliance model are.

Instead, here I operate under the premise that compliance cannot be realistically modeled explicitly, as is the case in the absence of adequate covariates and/or a reasonable model.

Moreover, there inferential interest here does not lie on a population parameter, but rather a function of a vector of population parameters ($\boldsymbol{\pi}^{NC}$), some of which are not point-identifiable, namely the interest lies on $rTPR_c$, $rTPR_i$ and their counterpart false positive rates. The goal is to construct a (large sample based) confidence interval for these partially identified quantities.

In particular, when dealing with an (univariate) partially identified parameter, there are lower and upper bounds for the parameter, which represent its identification region.

When conducting inference for such bounds, there are generally two approaches:

- one can either construct CI that cover the identification region for a fixed level $1-\alpha$
- alternatively, one can construct CI that cover the true (unknown) parameter with a specific $1-\alpha$ level.

To make matters more formal, consider the following general example: Let $\theta \in \mathbb{R}$ be a parameter of interest. Assume that θ cannot point-identified, but that there are lower and upper bounds, which are point-identified, such that $\theta_l \leq \theta \leq \theta_u$ where θ_l, θ_u are finite. The identification region for θ is then $\mathcal{I}(\theta) = [\theta_l, \theta_u] \subset \mathbb{R}^1$.

- Main Option 1: Construct CI such that θ is (asymptotically) covered with probability of (at least) $1-\alpha$ (where δ_l, $\delta_u \geq 0$ and may depend on the data):

$$P\left(\theta \in \left[\ \widehat{\theta}_l - \delta_l; \widehat{\theta}_u + \delta_u \right]\right) \geq 1-\alpha$$

- Main Option 2: Construct CI such that it covers (asymptotically) $\mathcal{I}(\theta)$ at a desired level (where δ_l^*, $\delta_u^* \geq 0$ and may depend on the data):

$$P\left(\mathcal{I}(\theta) \subset \left[\ \widehat{\theta}_l - \delta_l^*; \widehat{\theta}_u + \delta_u^* \right]\right) \geq 1 - \alpha$$

Clearly, when the CI covers $\mathcal{I}(\theta)$ with probability $1-\alpha$, it also covers θ with probability at least $1-\alpha$ (Imbens and Manski (2004)).

$$P\left(\theta \in \left[\ \widehat{\theta}_l - \delta_l; \widehat{\theta}_u + \delta_u \right]\right) \geq P\left(\mathcal{I}(\theta) \subset \left[\ \widehat{\theta}_l - \delta_l; \widehat{\theta}_u + \delta_u \right]\right) = 1 - \alpha$$

One criteria of classifying existing literature on confidence intervals for partially identified parameters follows exactly the options outlined above. For a case in which the object of inference is the entire region $\mathcal{I}(\theta)$, Horowitz and Manski (2000) derive bounds and provide inference following Option 2 for the case of MAR covariate data.

Generally speaking, often the inferential interest lies on the parameter itself rather than on its identification region, i.e. the CI outlined in main option 1. Work on this has been done by e.g. Imbens and Manski (2004), Stoye (2009) and more recently by Chernozhukov et al. (2007), Woutersen and Ham (2013). Comparison of the works of these authors brings in another possible classification criteria, namely, the type inferential procedure for the CI (in a frequentist framework). While Imbens and Manski (2004) and subsequently Stoye (2009) provide analytical form for their (large sample based) CI, Chernozhukov et al. (2007) provides the CI based on a developed simulation procedure which is able to deliver inference for both the identification region and/or the parameter of interest. Bootstrap-based methods could be used as well, although Andrews (2000) cautions against their usage when the parameter of interest can lie on the boundary. Woutersen and Ham (2013) propose a Bootstrap-based procedure which aims to deal with doing inference on non-differentiable functions of the parameters, which indeed, is the case in the inequalities (2.5) for $rTPR_c$ and in (2.6) for $rTPR_i$.

Cheng and Small (2006) discuss an application and similar discussion of bounds for partially identified parameters in the case of a 3-arm clinical trial with noncompliance, where (some) bounds

are also non-differentiable functions of the underlying population parameters. The authors consider Bonferroni-type bootstrap-based procedures.

Proposed procedure

As before, in this section the focus is on the compliers. The results for the insisters are analogous. I concentrate on finding an analytic, frequentist (large sample) based procedure. A discussion for a possible bootstrap approach and a general insight for a possible Bayesian procedure for the CI construction are mentioned in the concluding Chapter 5.

Ideally, a CI following Option 1 would be preferred, as one is really interested in $rTPR_c$, i.e. the function of the parameters itself, not its identification region. While Imbens and Manski (2004), Stoye (2009) deal with such type of intervals, which cover the true parameter and are shorter than intervals which cover the whole identification region, their work requires joint asymptotic normality of the lower and upper bound.

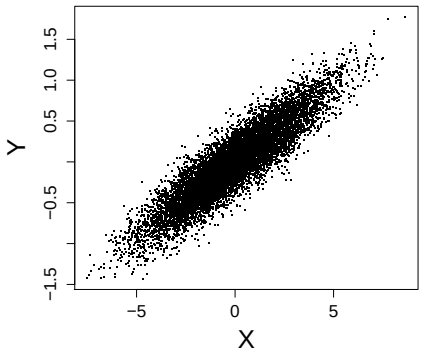
Here, this requirement is clearly not fulfilled, as the bounds are based on the minimum and the maximum of asymptotic bivariate normal estimators $g_1(\widehat{\boldsymbol{p}}^{NC})$ and $g_2(\widehat{\boldsymbol{p}}^{NC})$.

Hence, alternatively, one could look at an interval which covers the whole identification region. Horowitz and Manski (2000) provide such an interval which takes into account the joint distribution of the upper and the lower bound. However, their method also relies on normality of the joint distribution of lower and the upper bound.

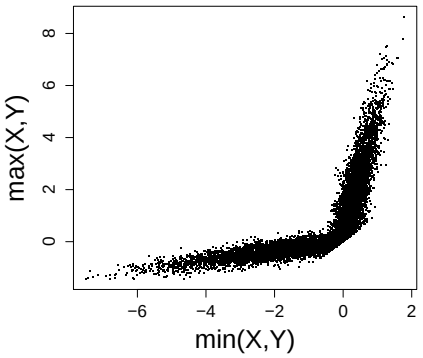
Another approach, which can be always done, is to work with Bonferroni type of CI. The drawback is that such bounds are very conservative, as they

- a) might not take into account the joint distribution of the lower and the upper bound
- b) cover the whole identification region and not just the parameter of interest.

Nevertheless, such an approach has the advantage of being easy to conduct. While the Bonferroni bounds represent the "worst case" in the sense of being the most wide, they can still be of good use in case they are "informative", in the sense of being non-trivial. Hence, I decide to construct the CI for $rTPR_c$ in this way.



(a) A scatterplot of a 10000 samples from the distribution of (X,Y).



(b) A scatterplot of (min(X,Y),max(X,Y) using the same data.

Figure 3.1: Minimum and maximum of correlated bivariate normal random variables.

Let $LB_c(\boldsymbol{\pi^{NC}}) = min\left(\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}}, \frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}}\right)$ and $UB_c(\boldsymbol{\pi^{NC}}) = max\left(\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}}, \frac{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{11}^{i,1}+\pi_{01}^{c,1}}\right)$. Then it follows for identification region $\mathcal{I}_c$ for $rTPR_c$: $\mathcal{I}_c = [LB_c(\boldsymbol{\pi}^{NC}), UB_c(\boldsymbol{\pi}^{NC})]$ At this stage one can consider two types of Bonferroni bounds: The first type derives the distribution of $\widehat{LB}_c = min\left(g_1(\widehat{\boldsymbol{p}}^{NC}), g_2(\widehat{\boldsymbol{p}}^{NC})\right)$ and $\widehat{UB}_c = max\left(g_1(\widehat{\boldsymbol{p}}^{NC}), g_2(\widehat{\boldsymbol{p}}^{NC})\right)$. Before looking into it in more detail, consider that the lack of asymptotic normality of the bounds prevents us from using the approach Imbens and Manski (2004) for a CI of the partially identified parameter. The main problem lies within the fact that the min and max of a bivariate normal vector are generally nonnormal. To illustrate that, consider the simple example of the bivariate vector

$$(X, Y) \sim N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 5^2 & 0.9 \\ 0.9 & 0.2^2 \end{array}\right)\right) \tag{3.9}$$

and $\rho = 0.9$ on Figure 3.1.

On the other hand, an opportunity arises, as asymptotically, the two estimators are jointly normal by the multivariate delta method and one can use results for the distribution of the minimum and maximum of two normal variables to find $f_{\widehat{LB}_c}$ and $f_{\widehat{UB}_c}$. Exact expressions for the density and the moments of the minimum and maximum of bivariate correlated normal variables are well known at least as early as in Clark (1961), and more recently in Nadarajah and Kotz (2008).

Given (X_1, X_2) as a bivariate Gaussian vector with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and

correlation coefficient ρ, the density of, e.g. $X = max(X_1, X_2)$ is expressed as (Nadarajah and Kotz (2008)):

$$\begin{aligned} f_X(x) &= f_1(-x) + f_2(-x) \\ f_1(x) &= \frac{1}{\sigma_1} \cdot \phi\left(\frac{x+\mu_1}{\sigma_1}\right) \cdot \Phi\left(\frac{\rho \cdot (x+\mu_1)}{\sigma_1 \cdot \sqrt{1-\rho^2}} - \frac{x+\mu_2}{\sigma_2\sqrt{1-\rho^2}}\right) \\ f_2(x) &= \frac{1}{\sigma_2} \cdot \phi\left(\frac{x+\mu_2}{\sigma_2}\right) \cdot \Phi\left(\frac{\rho \cdot (x+\mu_2)}{\sigma_2 \cdot \sqrt{1-\rho^2}} - \frac{x+\mu_1}{\sigma_1\sqrt{1-\rho^2}}\right) \end{aligned} \tag{3.10}$$

The density of the minimum is analogous and is omitted here. The reader is referred to Clark (1961), Nadarajah and Kotz (2008) for more details. Hence, an analytic, large sample, Bonferroni-type solution which covers $\mathcal{I}(\theta)$ could be to use the sample equivalent of LB_c and UB_c and construct a CI, such that:

$$P\left(\widehat{LB}_c - LB_c > \delta_l\right) = P\left(UB_c - \widehat{UB}_c > \delta_u\right) = \frac{\alpha}{2} \tag{3.11}$$

and numerically find the cutoff point for δ_l and δ_u which would then be used to construct the confidence interval. While one would still be doing a marginal analysis for $\widehat{LB}_c$ and $\widehat{UB}_c$, the criticism explained in point a) would be somewhat mitigated by the fact that the joint structure of $g_1(\widehat{\boldsymbol{p}}^{NC})$ and $g_2(\widehat{\boldsymbol{p}}^{NC})$ would still be exploited. The problem is, however, that LB_c and UB_c are naturally unknown quantities. To be able to solve for δ_l, one can possibly replace them with the expected value of the minimum (maximum) of two normally correlated normal variables. Indeed, applying the expression from Nadarajah and Kotz (2008) for the maximum of two normally correlated normal variables to our case (in which the estimators $\widehat{g}_1$ and $\widehat{g}_2$ are asymptotically normal) we find the functional form of $E(\widehat{UB}_c)$ to be:

$$E_{max} = g_1 \cdot \Phi\left(\frac{g_1 - g_2}{\theta}\right) + g_2 \cdot \Phi\left(\frac{g_2 - g_1}{\theta}\right) + \theta \cdot \Phi\left(\frac{g_1 - g_2}{\theta}\right)$$

where $\theta = \sqrt{\frac{\left(\sigma^2_{g_1} + \sigma^2_{g_2} - 2 \cdot \sigma_{g_1 g_2}\right)}{n}}$. It can be shown by the continuous mapping theorem that the sample equivalent of E_{max} is a consistent estimator for LB_u and could thus be plugged in

$P\left(UB_c - \widehat{UB}_c > \delta_u\right) = \frac{\alpha}{2}$ to find out δ_u.

Unfortunately, preliminary analysis based on simulations shows that the dependence of $\widehat{E}_{max}$ on n also leads to difficulties in constructing a reasonable (in terms of sample size) asymptotic pivotal quantity for the CI of LB_c and UB_c. The problem is worst in the case of point-identification or near-point-identification, i.e. when $\mathcal{I}_c = 0$ or very small. Further analysis directions would be of future research interest and some related ideas will be discussed in Chapter 5.

The second Bonferroni type CI which is proposed is more simple and does not take into account the joint distribution of $g_1(\widehat{\boldsymbol{p}}^{NC})$ and $g_2(\widehat{\boldsymbol{p}}^{NC})$. Naturally, this means that the CI are wider, however they give at least the desired coverage in reasonable sample sizes as it will be shown in the next Chapter 4. Define the endpoints of the following CI as

$$
\begin{aligned}
g_{1L}(\widehat{\boldsymbol{p}}^{NC}) &= g_1(\widehat{\boldsymbol{p}}^{NC}) - z_{1-\alpha^*/4} \cdot \widehat{\sigma}_{g_1}/\sqrt{n} \\
g_{1U}(\widehat{\boldsymbol{p}}^{NC}) &= g_1(\widehat{\boldsymbol{p}}^{NC}) + z_{1-\alpha^*/4} \cdot \widehat{\sigma}_{g_1}/\sqrt{n} \\
g_{2L}(\widehat{\boldsymbol{p}}^{NC}) &= g_2(\widehat{\boldsymbol{p}}^{NC}) - z_{1-\alpha^*/4} \cdot \widehat{\sigma}_{g_2}/\sqrt{n} \\
g_{2U}(\widehat{\boldsymbol{p}}^{NC}) &= g_2(\widehat{\boldsymbol{p}}^{NC}) + z_{1-\alpha^*/4} \cdot \widehat{\sigma}_{g_2}/\sqrt{n}
\end{aligned}
$$

for some level α^* as lower and upper points of symmetric marginal CI for $g_1(\boldsymbol{\pi}^{obs})$ and $g_2(\boldsymbol{\pi}^{obs})$.

Proposition 3.2.2 *For a given level α^*, such that*

$$
\begin{aligned}
&P\left(g_{1L}(\widehat{\boldsymbol{p}}^{NC}) \geq g_1(\widehat{\boldsymbol{p}}^{NC})\right) = P\left(g_{2L}(\widehat{\boldsymbol{p}}^{NC}) \geq g_2(\widehat{\boldsymbol{p}}^{NC})\right) \\
&= P\left(g_{1U}(\widehat{\boldsymbol{p}}^{NC}) \leq g_1(\widehat{\boldsymbol{p}}^{NC})\right) = P\left(g_{2U}(\widehat{\boldsymbol{p}}^{NC}) \leq g_2(\widehat{\boldsymbol{p}}^{NC})\right) = \frac{\alpha^*}{4}
\end{aligned} \tag{3.12}
$$

for large n it holds that:

$$
\begin{aligned}
P\Big(\Big\{min\left(g_{1L}(\widehat{\boldsymbol{p}}^{NC}), g_{2L}(\widehat{\boldsymbol{p}}^{NC})\right) \leq min\left(g_1(\boldsymbol{p}^{NC}), g_2(\boldsymbol{p}^{NC})\right)\Big\} \cap \\
\Big\{max(g_{1U}(\widehat{\boldsymbol{p}}^{NC}), g_{2U}(\widehat{\boldsymbol{p}}^{NC})) \geq max(g_1(\boldsymbol{p}^{NC}), g_2(\boldsymbol{p}^{NC}))\Big\}\Big) \geq 1 - \alpha^*
\end{aligned}
$$

See the Appendix for a proof.

Finite sample properties of the proposed hypothesis test and confidence interval procedures which are proposed here will be discussed in Chapter 4. Before I proceed with those, in the next section I discuss some considerations related to design and the choice of a one-arm vs. two-arm screening trial which is naturally posed when dealing with nested diagnostic tests.

3.3 Design considerations

Although the decision about the design of a study (in particular, sample size determination) is made before any inference is conducted, it is convenient to situate this section in the current Chapter 3, as the sample size cannot be determined before functional expressions for the variances of the test statistics under the alternative and under the null are known. Indeed, they were derived in the previous sections and from the asymptotic distributions of the test statistics under full compliance $T_1^F(\widehat{\boldsymbol{p}}^F)$, $T_2^F(\widehat{\boldsymbol{p}}^F)$ and their counterparts under post-screening noncompliance in Sections 2.1 and 2.2, one can use the expressions for the asymptotic variances under the null and alternative hypotheses to determine the sample size that achieves a desired level of power.

Moreover, before the screening trial is conducted, it is advisable to compare the power of the two test statistics within each model and use the one which gives the higher power to determine the necessary sample size, or, alternatively, use the one which requires less patients (and thus less money) to achieve the same level of power. These standard techniques for sample size determination are well known and I do not discuss them here further.

However, the specific nature of nested diagnostic tests creates an opportunity for an additional consideration regarding the overall cost of the trial. Consider the full compliance model as in Section 2.1: Due to nestedness, from each administered test kit in arm 1, the researcher also has access to information on e.g. the biomarker for diagnostic test T_0, since $\mathcal{A} \subseteq \mathcal{B}$. Furthermore, the functional form of the T_0 diagnostic test is *known* to the researcher, so that the would-be T_0 diagnostic test recommendation is known for all patients. Hypothetically, instead of doing a two-arm test in which one group receives the T_0 test and the other group, the T_1 test, the T_1 test kit could be administered to *all* patients instead. Thus, in this hypothetical example, if either the

would-be T_0 test result, or the T_1 test results are positive, then the patient could get her D status verified.

This hypothetical one-arm trial would be equivalent to the analysis of a *paired* (as opposed to the unpaired design that has been under discussion for the screening trial outlined in this thesis), screen positive screening trial, which has been studied extensively (e.g. refer to Pepe (2003) for estimators of both $rTPR$ and $rFPR$ in this hypothetical one-arm paired design setting). Note that under both the hypothetical one-arm trial and the two-arm trial, the inference would be conducted for the same $rTPR$ and $rFPR$. Naturally, the variances of the estimators under each trial design will be different. Indeed, it can be shown that, for equal true underlying probabilities $\boldsymbol{\pi}^F$ and total number of patients n, the variance of the estimator $rTPR$ under the one-arm paired trial would be lower than the variance of $rTPR(\widehat{\boldsymbol{p}}^F)$ from the two-arm unpaired trial.

However, the researcher has to also consider the difference in the costs of the two tests. If the T_1 diagnostic test costs much more than the T_0 (possibly routine, thus inexpensive) diagnostic test, then the number of patients that the researcher can afford to put in the hypothetical one-arm trial, might be much lower than the total number of patients in a (potentially unbalanced) two-arm trial. Thus, for the same level of desired power of the hypothesis test $H_0 : rTPR \leq 1$ vs. $H_1 : rTPR > 1$, the two-arm trial may lead to a lower overall cost.

Setting aside noncompliance for a moment, the nested property of the test poses the natural question of whether it would be better (from a cost perspective) to conduct a one-arm study or a two-arm study. Of course, this consideration is in no way restricted to nested tests only: even if T_0 were not nested within T_1, one could ask the same question. Note that, often, unpaired (two-arm) designs are necessary when the two diagnostic tests are mutually exclusive, which clearly is not the case in the nested framework. Hence, when the researcher is confronted with a test nested within another, given the costs of the two tests and a desired level of power, she would need a decision rule for the choice whether to administer a one-arm or a two-arm trial.

To formalize the argument, consider a case as in Section 2.1 with full compliance. Let c_0 and c_1 represent the unit cost for each test kit. Assume that one is interested in testing $H_0 : rTPR \leq 1$ vs. $H_1 : rTPR > 1$ (the argument for $rFPR$ is analagous). In a hypothetical one-arm trial, $\pi_{10}^{c,1}$

and $\pi_{01}^{c,1}$ would be point-identifiable, so one could directly build a test statistic, say $\left(\widehat{\pi}_{10}^{c,1} - \widehat{\pi}_{01}^{c,1}\right)$, to test this hypothesis. For the two-arm trial described in Section 2.1, let n_0 and n_1 be the number of patients in the respective arms. Let $\tilde{n}_1$ be the number of patients in the one-arm hypothetical design. $\tilde{n}_1$ will be determined from the power function of the test statistic $\left(\widehat{\pi}_{10}^{c,1} - \widehat{\pi}_{01}^{c,1}\right)$ and will be the "baseline" number of patients. For simplicity, throughout, I treat the number of patients as a continuous variable when determining the sample size for a fixed power and then round up to the next integer. The cost minimization function for a two-arm trial is then:

$$\underset{n_1,n_0}{\arg\min} \left(c_1 \cdot n_1 + c_0 \cdot n_0\right) \quad \text{s.t.} \quad \gamma(\cdot) = 1 - \beta \tag{3.13}$$

where $\gamma(\cdot)$ is the power function when using either $T_1^F(\widehat{p}^F)$ or $T_2^F(\widehat{p}^F)$ at some desired level $1-\beta$, say 80%. For illustration purposes, I choose to work with the power function of $T_1(\widehat{p}^F)$, but note that all results discussed in this section would also hold for $T_2(\widehat{p}^F)$. Observe that $\tilde{n}_1$ will be determined from the hypothetical one-arm trial to ensure the same desired level power of $1-\beta$. The "baseline" budget from the one-arm trial is $M^{one-arm} = c_1 \cdot \tilde{n}_1$. All feasible solutions for (n_0, n_1) will give an overall cost, $M^{two-arm} = c_1 \cdot n_1 + c_0 \cdot n_0$. I am interested in the optimal pair of (n_0, n_1) corresponding to $\min\{M^{two-arm}(n_0, n_1) : M^{two-arm} \leq M^{one-arm}\}$. Thus, for all solutions of interest one has that $n_1 \leq \tilde{n}_1$. Since I am looking for the minimum under pairs $(n_0, n_1) \in \mathbb{N}^+ \times ([1, \tilde{n}_1] \cap \mathbb{N}^+)$, one could avoid running a (discrete) optimization procedure for the target function in expression (3.13) and simply enumerate all possible pairs. By comparing the total costs, the optimal solution can be found.

A more elegant and computationally faster way is to rewrite the problem from expression (3.13) by fixing n_1 and thus reducing dimensionality:

$$\underset{n_0}{\arg\min} \left(c_1 \cdot n_1 + c_0 \cdot n_0\right) \quad \text{s.t.} \quad \gamma(\cdot) = 1 - \beta \quad \text{for } n_1 = 1, \ldots, \tilde{n}_1 \tag{3.14}$$

Effectively, this way reduces the problem to solving the power function for n_0, given each fixed n_1 (which I treat here as a discrete variable) and subsequently plugging the pair of sample sizes in the total cost function. Fixing the vector of parameters under the desired null and alternative,

the rejection level α, n_1 and the desired power level, one can use, e.g. `uniroot` in R 3.0.2 (2014) to solve the power function for n_0, given each value of n_1. Let us call each such solution (n_0^*). Plugging in all pairs (n_0^*, n_1) in $M^{two-arm}$ then gives us the unique minimum (among the set of pairs which give the desired level of power, provided that this set is not empty for a given n_1), i.e. (n_0^*, n_1^*) which minimizes the target function in (3.13).

A complication arises from the fact that the functional expressions for the (population) variances under the null and the alternative in the power function $\gamma(\cdot)$ all depend on n_1 and n_0. As a consequence, uniqueness of n_0^* may be not guaranteed. From my experience, in the case of multiple solutions one will simply take the first (smallest) one. However, for a large set of suitable alternative hypotheses, n_0^* is indeed unique as the following proposition shows:

Proposition 3.3.1 *For any $n_1 \in \{1, \ldots, \tilde{n}_1\}$ and any $\boldsymbol{\pi}^F$ under the null, if $\left(2 \cdot \pi_{11}^{c,1} + \pi_{10}^{c,1} + \pi_{01}^{c,1}\right) < 1$ under the alternative, then $\exists!$ n_0, s.t. $\gamma(\cdot) = 1 - \beta$.*
For a proof, see the appendix.

Note that since if one is dealing with a rare disease (e.g. cancer), the condition is likely to be met. Intuitively, if this is the case, for each $n_1 \in \{1, \ldots, \tilde{n}_1\}$ by solving the power function $\gamma(\cdot)$ for n_0, an unique answer can be obtained.

For an illustration, consider the following example:

$$c_1 = 80, c_0 = 1$$

$$\alpha = 5\%, \beta = 20\%$$

$$\boldsymbol{\pi}_0^F = (0.03, 0.05, 0.05, 0.007, 0.04, 0.013, 0.013, 0.797)$$

$$\boldsymbol{\pi}_1^F = (0.03, 0.075, 0.025, 0.007, 0.04, 0.013, 0.013, 0.797)$$

for the vector of parameters of the true underlying table under the null under the alternative. One has that $T_1^F(\boldsymbol{p}^F) = \pi_{10}^{c,1} - \pi_{01}^{c,1} = 0.025$. In this example, I find out that $\tilde{n}_1 = 246$. The following Figure 3.2 illustrates the comparison between the hypothetical one-arm trial and the two-arm trial.

Looking at the left y-axis scale and the x-axis, one can see combinations of pairs for (n_0, n_1) which (uniquely) guarantee the desired power level. That is, for each point of n_1 in the grid search,

the power function is inverted in order to obtain an unique n_0, such that the power is 80%. On the right hand y-axis, I plot the corresponding budget, as per the target function in expression (3.13).

In the example where $c_1 = 80$, one can find that the minimum overall cost is the pair $(n_0^*, n_1^*) = (1760, 219)$. They key message here is the following: By removing 26 patients from the expensive arm and putting 1760 patients in the nested, cheaper arm, one is able to achieve the *same* statistical power as the hypothetical one-arm experiment *while reducing* the total cost by 320 monetary units.

In comparison, if, ceteris paribus, I would choose a lower price for the expensive test, say $c_1 = 40$, the corresponding one-arm trial budget will below the one of the two-arm trial. That means that $\{(n_0, n_1) : M^{two-arm} \leq M^{one-arm}\}$ is an empty set, hence the economic choice here would be to run the one-arm trial.

The main purpose of the discussion in this section is to give a decision rule as to whether prefer a one-arm or a two-arm trial, depending on pre-trial variables, such as the costs for each separate test, desired hypothesis and power level. I emphasize that this example was chosen for illustration purposes and note that additional costs, e.g. administrative costs should also be considered. I will elaborate in this point in Chapter 5.

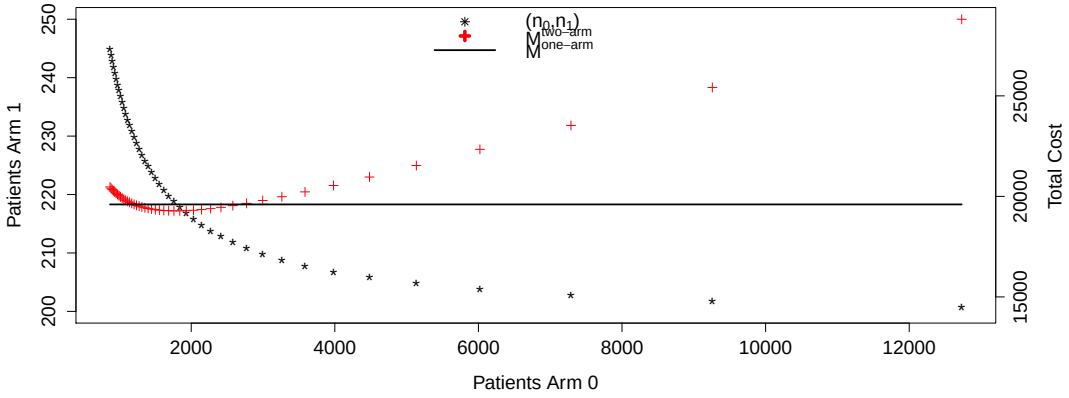


Figure 3.2: Total cost of the hypothetical one-arm trial ($M^{one-arm}$) vs. total cost of a two-arm trial ($M^{two-arm}$) on the right-hand Y-axis, for pairs of (n_0, n_1) (X-axis and left-hand Y-axis) which all achieve the desired statistical power of 80%. See text for details.

Under post-screening noncompliance, the outlined argument for two-arm vs. one-arm trial still holds. Arguably, both in the hypothetical one-arm trial and in the two-arm trial, the set of refusers

will be the same. Thus, in both designs, one would have to do inference for $rTPR_{i+c}$. Working out a simple proposition such as (3.3.1) is however much more difficult and an enumeration of all feasible solutions (n_0, n_1) seems to be preferable. Note that in a two-arm trial with noncompliance, one would have the possibility to also identify some insisters —which would not be possible under a one-arm trial.

Naturally, the whole argument would hold for any two diagnostic tests, in which one is nested within the other. Indeed, it may happen that the nested test is a standard (and essentially free) routine clinical procedure: in this case, putting the additional information to use in the form of a second trial arm can indeed prove to be cost efficient.

In the next Chapter 4, a finite sample evaluation of the relevant inferential procedures discussed in Chapter 3 is presented.

Chapter 4

Finite sample evaluation of the inferential procedures

First, I conduct a finite sample evaluation of the tests related to the hypotheses stated in expression (3.3). Overall, it shows good finite sample properties in terms of empirical size and power. To evaluate the described test procedures for their finite sample size properties, i.i.d data from a multinomial distribution is simulated for different sets of parameters (scenarios) under the null of $rTPR_{i+c} \leq 1$ and the alternative of $rTPR_{i+c} > 1$. Starting from a (realistic) set of parameter vectors $\boldsymbol{\pi}_0^{NC}$ under the null and $\boldsymbol{\pi}_1^{NC}$ under the alternative, I evaluate the performance of the empirical size and power of $\widehat{T}_1^{NC}$ and $\widehat{T}_2^{NC}$.

I look at one scenario for the size and three scenarios for the power. In this set of simulations, I choose for the set of parameters under the null $\boldsymbol{\pi}_0^{NC}$ for $s = c, i.r$:

$$(\pi_{11}^{s,1}, \pi_{10}^{s,1}, \pi_{01}^{s,1}, \pi_{00}^{s,1}) = (0.01, 0.05, 0.05, 0.007)$$
$$(\pi_{11}^{s,0}, \pi_{10}^{s,0}, \pi_{01}^{s,0}, \pi_{00}^{s,0}) = (0.007, 0.05, 0.05, 0.11)$$

For the three power scenarios, most of the elements from $\boldsymbol{\pi}_0$ remain the same but I add an equal number to $\pi_{10}^{c,1}$ and $\pi_{10}^{i,1}$ while removing the same number from $\pi_{01}^{c,1}$ and $\pi_{01}^{i,1}$. Thus

- For scenario 1, $\pi_{10}^{c,1} + \pi_{10}^{i,1} - \pi_{01}^{c,1} - \pi_{01}^{i,1} = 0.010$

- For scenario 2, $\pi_{10}^{c,1} + \pi_{10}^{i,1} - \pi_{01}^{c,1} - \pi_{01}^{i,1} = 0.015$
- For scenario 3, $\pi_{10}^{c,1} + \pi_{10}^{i,1} - \pi_{01}^{c,1} - \pi_{01}^{i,1} = 0.020$

The results are reported in Table 4.1. For the first two rows of the table, $\widehat{\gamma}(\cdot)$ represents empirical size for the two test statistics. The theoretical size is set at $\alpha = 5\%$. In the remaining rows, it represents the empirical power for the corresponding estimators. $\gamma(\cdot)$ represent the theoretical power for each estimator under each scenario. The number of simulations is 50000. The considered sample size ranges between 5000 up to 25000. Both estimators have satisfactory finite sample properties, as the empirical values are close to the theoretical ones. Note that these are not unrealistic sample sizes when evaluating the effect of a screening policy. As an example, consider the multi-centre"European Randomized Study of Screening for Prostate Cancer" which aims to study the effect of PSA screening (vs. no screening) on prostate cancer mortality in the core age group and has over 162000 patient (see Schröder et al. (2014) for more details).

To conduct a finite sample evaluation of the test size and power for the stratum-specific ratios, the same approach is followed and i.i.d data from the underlying multinomial model is generated. For the testing part of the hypothesis $H_0 : rTPR_c \leq 1$, I have considered the following scenarios for $\boldsymbol{\pi^{NC}}$:

(i): In the baseline scenarios for $rTPR_c = 1$ and $rFPR_c = 1$ (i.e. the null hypothesis for equal relative accuracy of both tests), for the first 12 probabilities of $\boldsymbol{\pi^{NC}}$ (i.e. where $D = 1$) I set $(\pi_{11}^{s,1}, \pi_{10}^{s,1}, \pi_{01}^{s,1}, \pi_{00}^{s,1}) = (0.06, 0.05, 0.05, 0.007)$ for $s = c, i, r$. Note that $\pi_{00}^{s,1}$ is realistically chosen to be very small, as there should be a low probability that both tests are negative and the patient is indeed diseased. The numbers are reverse for the elements 13 to 24 of the $\boldsymbol{\pi^{NC}}$ vector, i.e. the cases where the patient does not have cancer (i.e. $D = 0$) and indeed the probability that both diagnostic tests give a positive answer should be small: $(\pi_{11}^{s,0}, \pi_{10}^{s,0}, \pi_{01}^{s,0}, \pi_{00}^{s,0}) = (0.007, 0.05, 0.05, 0.06)$ for $s = c, i, r$. Note that under the null, i.e. under this baseline scenario, $g_1(\boldsymbol{p}^{NC}) = g_2(\boldsymbol{p}^{obs}) = 0$. That is to say, the $rTPR_c$ and $rFPR_c$ are both point-identified from the observed data distribution and the interpretation is that both tests have the same relative accuracy.

(ii): The next three scenarios are simulating data under the alternative (separately for $rTPR_c$ and $rFPR_c$) that $g_2(\boldsymbol{p}^{NC}) > 0$. The way that I construct the alternative distribution similar to

	$n = 5000$	$n = 10000$	$n = 20000$	$n = 25000$
		Size experiment		
$\widehat{\gamma}(T_1^{NC})$	0.0511	0.0469	0.0495	0.0536
$\widehat{\gamma}(T_2^{NC})$	0.0500	0.0495	0.0504	0.0498
		Power experiment 1		
$\widehat{\gamma}(T_1^{NC})$	0.2955	0.4621	0.7059	0.7890
$\gamma(T_1^{NC})$	0.2888	0.4577	0.7024	0.7847
$\widehat{\gamma}(T_2^{NC})$	0.1709	0.2489	0.3938	0.4547
$\gamma(T_2^{NC})$	0.1662	0.2444	0.3820	0.4434
		Power experiment 2		
$\widehat{\gamma}(T_1^{NC})$	0.5138	0.7807	0.9488	0.9782
$\gamma(T_1^{NC})$	0.4949	0.7464	0.9473	0.9775
$\widehat{\gamma}(T_2^{NC})$	0.2642	0.4222	0.6424	0.7364
$\gamma(T_2^{NC})$	0.2632	0.4134	0.6435	0.7272
		Power experiment 3		
$\widehat{\gamma}(T_1^{NC})$	0.7161	0.9365	0.9967	0.9996
$\gamma(T_1^{NC})$	0.7024	0.9241	0.9966	0.9994
$\widehat{\gamma}(T_2^{NC})$	0.3889	0.6064	0.8563	0.9134
$\gamma(T_2^{NC})$	0.3824	0.5998	0.8493	0.9112

Table 4.1: Finite sample evaluation of T_1^{NC} and T_2^{NC} for size and different power experiments. $\widehat{\gamma}(\cdot)$ represents empirical size (in the first two rows) and empirical power for the corresponding estimators. $\gamma(\cdot)$ represent the theoretical power for each estimator, under each scenario - see details in main text.

the approach for the evaluation of the asymptotic test procedures in Table 4.1: I keep all other probabilities as in the baseline case, but for each scenario under the alternative I add a small number to $\pi_{10}^{c,1}$ and subtract the same number from $\pi_{01}^{c,1}$. I have 3 scenarios and the numbers that I add and subtract are respectively:

- 0.011 for scenario 1 (thus $exp(g_2(\boldsymbol{p}^{\boldsymbol{NC}})) = 1.15$)
- 0.018 for scenario 2 (thus $exp(g_2(\boldsymbol{p}^{\boldsymbol{NC}})) = 1.25$)
- 0.034 for scenario 3 (thus $exp(g_2(\boldsymbol{p}^{\boldsymbol{NC}})) = 1.5$)

In this way, the total sum of probabilities in $\boldsymbol{p}^{NC}$ does not change, but the relative accuracy of the two diagnostic tests does (as well as the value of $g_1(\boldsymbol{p}^{NC})$). For the evaluation regarding $rFPR_c$ I proceed in the same manner, keeping all other probabilities except $\pi_{10}^{c,0}$ and $\pi_{01}^{c,0}$ as in the baseline case. For these two, I add a small number to $\pi_{10}^{c,0}$ and subtract the same number from $\pi_{01}^{c,0}$. The 3

numbers which I add and subtract to construct the 3 scenarios under the alternative are respectively

- 0.004 for scenario 1 (thus $exp(g_2^{rFPR}(\boldsymbol{p^{NC}})) = 1.15$)
- 0.007 for scenario 2 (thus $exp(g_2^{rFPR}(\boldsymbol{p^{NC}})) = 1.25$)
- 0.012 for scenario 3 (thus $exp(g_2^{rFPR}(\boldsymbol{p^{NC}})) = 1.5$)

In this way, the total sum of probabilities in $\boldsymbol{p^{NC}}$ does not change, but the relative accuracy of the two diagnostic tests does (as well as the value of $g_1^{rFPR}(\boldsymbol{p}^{NC})$). For both the size and power evaluations in the stratum-specific case, I conducted 20000 simulations with different sample sizes from $n = 2500$ to $n = 22500$ equally spaced by 2500.

To complement the fixed level α cut-off reporting approach for Table 2.6, now I choose to display the results for the size by using a size-discrepancy plot. I display the results for the size by using a size-discrepancy plot (see e.g. Davidson and MacKinnon (1998)), rather than by using a table with fixed cut-off level α. The former approach has the advantage of providing a graphical overview of how well the null distribution is indeed approximated for all quantiles. To see this, consider the inverse CDF of the normal $\Phi^{-1}(\cdot)$ and let M represent the number of simulations. Then, for the distribution of all the transformed test statistics under the null, it must hold $\left(\Phi^{-1}\left(g_2^{H_0,(1)}(\widehat{\boldsymbol{p}}^{NC})\right), \ldots, \Phi^{-1}\left(g_2^{H_0,(M)}(\widehat{\boldsymbol{p}}^{NC})\right)\right) \overset{i.i.d.}{\sim} U(0,1)$.

A size discrepancy plot displays the difference between a given (sorted) sequence

$$\left(\Phi^{-1}\left(g_2^{H_0,(m)}(\widehat{\boldsymbol{p}}^{NC})\right), 1 \leq m \leq M\right)$$

and the 45° line, i.e the nominal size. Obviously, this difference has to be evaluated at $M = 20000$ points. Here, I construct the points of the 45° line at which this difference has to be evaluated, as an equidistant sequence $(x_1, \ldots, x_M) = ((\texttt{0E} - \texttt{16}), \ldots, (1 - (\texttt{0E} - \texttt{16})))$. (For more details on alternative ways to construct this sequence - see Davidson and MacKinnon (1998)).

Since the aim is to determine optimal sample size of patients to detect a given alternative for a given power level, I also display a plot of the empirical power vs. the theoretical power for different n and different alternatives in Figure 4.1. One can see that the asymptotic normal approximation is adequate with reasonable sample sizes (e.g. $n = 10000$). The test is slightly oversized at,

e.g., $\alpha = 0.05$, but this problem seems to disappear with increasing sample size. Theoretical and empirical power seem to be close enough to each other, especially as the alternative and the null get further apart.

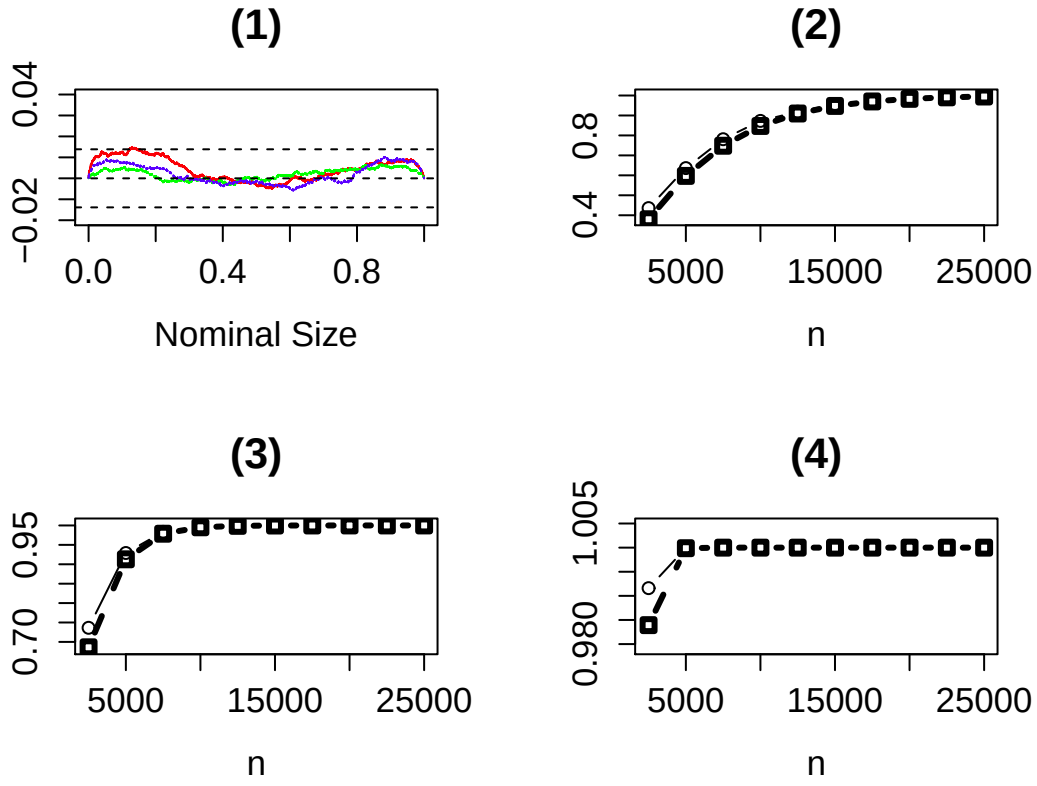


Figure 4.1: Empirical vs. theoretical size and power for $rTPR_c$. 20000 simulations for different sample sizes n, $\alpha = 0.05$ and $P(Z = 0) = 0.5$. Panel (1) represents a size discrepancy plot for the $rTPR_c$ (see text for details). The red line corresponds to $n = 2500$, blue to $n = 10000$ and green to $n = 22500$. The top and the bottom dashed line represent the corresponding Kolmogorov-Smirnov critical value at $\alpha = 0.05$. Panel (2) - (4) represent empirical power $(-\circ-)$ vs. theoretical power $(-\square-)$ for values for n ranging from 2500 to 22500 equidistantly by 2500, in the 3 scenarios described in the text - Panel (2) with $exp(g_2(p^{NC})) = 1.15$, panel (3) with $exp(g_2(p^{NC})) = 1.25$ and panel (4) with $exp(g_2(p^{NC})) = 1.5$

Results for the $rFPR_c$ from Figure 4.2 are similar. In smaller finite sample sizes, ($n = 10000$ vs. $n = 22500$), the approximation of the asymptotic normal distribution seems to be problematic, especially in the tails. Hence, the test is also slightly oversized at $\alpha = 0.05$. However, increasing sample size seems to remedy the problem. The same goes for the comparison between the theoretical

and the empirical power.

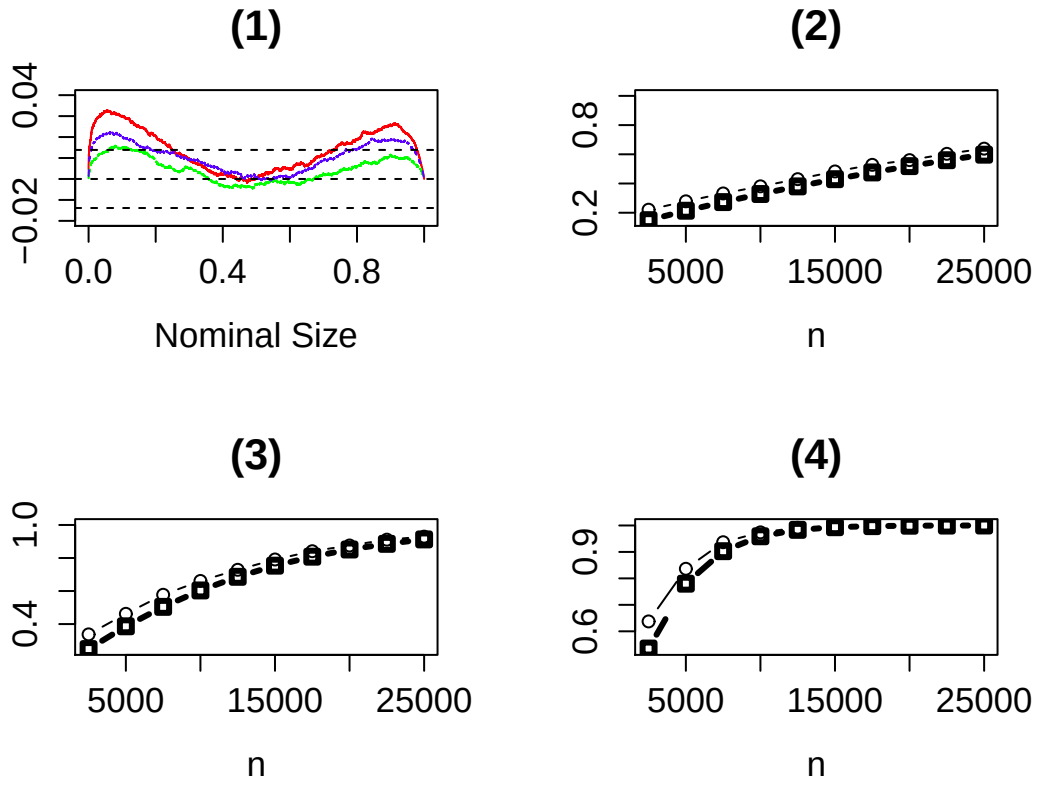


Figure 4.2: Empirical vs. theoretical size and power for $rFPR$. 20000 simulations for different sample sizes n, $\alpha = 0.05$ and $P(Z = 0) = 0.5$. Panel (1) represents a size discrepancy plot for the $rTPR$ (see text for details). The red line corresponds to $n = 2500$, blue to $n = 10000$ and green to $n = 22500$. The top and the bottom dashed line represent the corresponding Kolmogorov-Smirnov critical value at $\alpha = 0.05$. Panel (2) - (4) represent empirical power $(-\circ-)$ vs. theoretical power $(-\square-)$ for values for n ranging from 2500 to 22500 equidistantly by 2500, in the 3 scenarios described in the text - Panel (2) with $exp(g_2(\pi^{NC})) = 1.15$, panel (3) with $exp(g_2(p^{NC})) = 1.25$ and panel (4) with $exp(g_2(p^{NC})) = 1.5$

Finally, for the evaluation of the CI, I choose $\alpha^* = 0.05$, such that the theoretical coverage is 95%. I evaluate the CI for different lengths of $\mathcal{I}_c$ and for cases where $rTPR_c < 1, rTPR_c = 1, rTPR_c > 1$ (analogously for $rFPR_c$). One would expect that the bounds give coverage as per Proposition (3.2.2), i.e. such that empirically the empirical coverage is greater or equal than 95%.

The scenarios differ again in the probabilities of the pairs $(\pi_{10}^{c,1}, \pi_{01}^{c,1})$ and $(\pi_{10}^{c,0}, \pi_{01}^{c,0})$ respectively. All the other probabilities of the vector $\boldsymbol{\pi}$ are left as described above for the simulations evaluating

the finite sample size and power of the test procedure for $rTPR_c$. For $rTPR_c$ I have that

- $\Big((\pi_{10}^{c,1}, \pi_{01}^{c,1}) \quad [LB, UB]\Big) = (0.03, 0.07; [0.43, 0.7])$ for scenario 1
- $\Big((\pi_{10}^{c,1}, \pi_{01}^{c,1}) \quad [LB, UB]\Big) = (0.048, 0.052; [0.92; 0.96])$ for scenario 2
- $\Big((\pi_{10}^{c,1}, \pi_{01}^{c,1}) \quad [LB, UB]\Big) = (0.05, 0.05; [1, 1])$ for scenario 3 (note that this is the point-identified case as the length of the identification region is 0)
- $\Big((\pi_{10}^{c,1}, \pi_{01}^{c,1}) \quad [LB, UB]\Big) = (0.052, 0.048; [1.025, 1.037])$ for scenario 4
- $\Big((\pi_{10}^{c,1}, \pi_{01}^{c,1}) \quad [LB, UB]\Big) = (0.065, 0.035; [1.19, 1.31])$ for scenario 5.

In an analogous manner I proceed (indeed, by using the same numbers for the pairs $(\pi_{10}^{c,0}, \pi_{01}^{c,0})$, as I did for $(\pi_{10}^{c,1}, \pi_{01}^{c,1})$) for the simulation scenarios for $rFPR_c$. I let $P(Z = 0) = 0.05$, $n = 20000$ and have $M = 10000$ repetitions under the 5 different scenarios. Figure 4.3 and Figure 4.4 show boxplots of the lower and upper bounds for $rTPR_c$ and $rFPR_c$ under the different scenarios for the simulated cases. One can see that when the identification region is quite big (scenarios 1 and 5), the minimum and the maximum are quite clearly (and correctly) estimated as $g_1(\widehat{\boldsymbol{p}}^{NC})$ and $g_2(\widehat{\boldsymbol{p}}^{NC})$. Thus, depending on which one of their estimators has the shorter variance, the boxplot shows a narrower or a less narrow concentration around the median/mean.

The empirical coverage for $rTPR_c$ is (for each of the 5 scenarios respectively) (97.4% ; 98%; 97.9%;96.7%; 97.68%), so one can see that the bounds are generally always wider than they should be. Similar values hold for the coverage of the $rFPR_c$. This is understandable, especially for the scenarios where the identification region is very short, since the proposed procedure for the CI construction is very conservative and always picks the lowest (highest) lower point (upper point) of the marginal CIs.

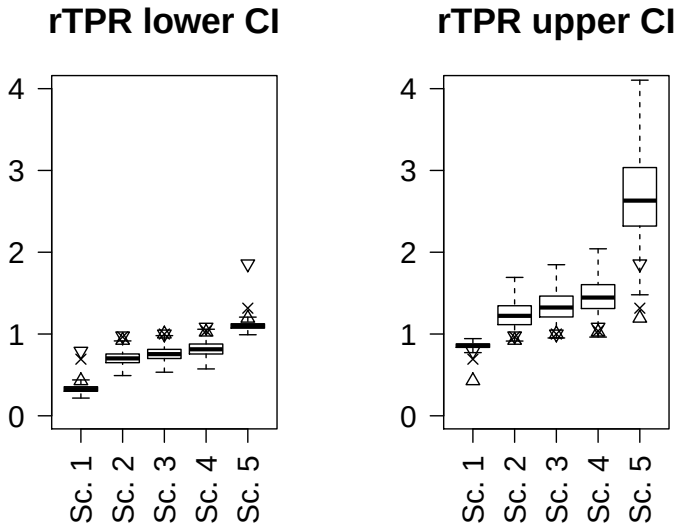


Figure 4.3: Boxplots of simulated lower and upper CI (see details in the text) for $rTPR_c$. The true LB is represented by "△", the true $rTPR_c$ by "×", the true UB by "▽."

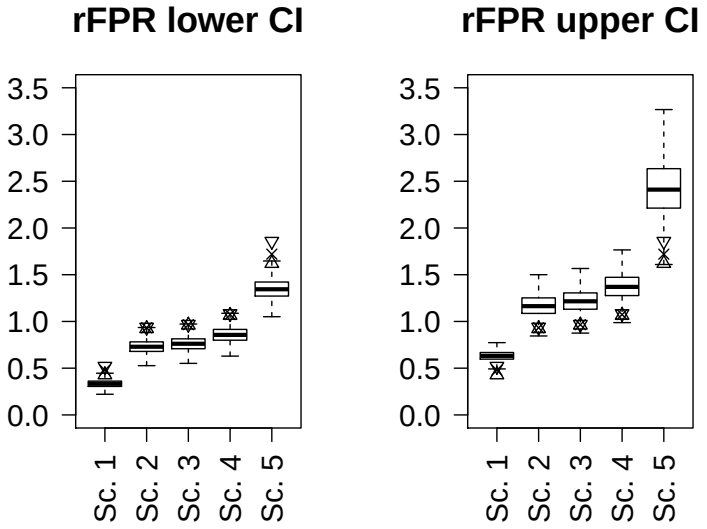


Figure 4.4: Boxplots of simulated lower and upper CI (see details in the text) for $rFPR_c$. The true LB is represented by "△", the true $rFPR_c$ by "×", the true UB by "▽."

Overall, one can see that the asymptotic hypothesis tests show good finite sample size and power properties. As expected from such a conservative approach, the CI are fairly wide, however

they do deliver proper coverage.

Chapter 5

Conclusion

I have defined the notion of a screening trial with a nested diagnostic test and provided an inferential procedure for relevant measures of relative accuracy, both in absence and in presence of post-screening noncompliance. By employing principal stratification, one is able to construct the underlying model under noncompliance. Furthermore, it was shown how to conduct inference for relevant quantities to measure relative accuracy in such a trial. The inferential procedures show good finite sample properties, as shown in Chapter 4.

It was also shown how having a nested diagnostic test can provide the researcher with additional information for the underlying population parameters. In the case of full compliance, this additional information can be used to construct alternative test statistics. In the case of post-screening noncompliance, the added informational value from a nested diagnostic test is twofold: First, as in the full compliance case, one is able to construct additional test statistics. Both in case of noncompliance and full compliance, the researcher can do a pre-trial power analysis and choose the sample size from the statistic ($T_1(\widehat{\boldsymbol{p}}^{NC})$ or $T_2(\widehat{\boldsymbol{p}}^{NC})$) with higher power. I emphasize that these statistics are structurally different, as they are based on a different set of parameters. In the case of $T_2(\widehat{\boldsymbol{p}}^{NC})$, these parameters only become identifiable thanks to the nested structure.

The second advantage comes when one considers stratum-specific versions of $rTPR$. These come out as a natural consequence of the flexible modeling capability that the PS method provides. Indeed, under the noncompliance model, one might define stratum-specific $rFPR_c$, $rTPR_c$ for

compliers and $rTPR_i$, $rFPR_i$ for insisters. It can be argued that a researcher might be interested in, e.g., $rFPR_c$ and not (only) in $rTPR_{i+c}$. This will most likely be the case when a statistical test shows that $rFPR_c$ and $rTPR_i$ are not both bigger than 1, even though $rTPR_{i+c}$ might be (intuitively this can be interpreted a manifestation of Simpson's paradox).

Also, note that having an unpaired design — nested or not — allows some patients to exhibit their underlying compliance status. Indeed, the principal strata here have intuitive interpretation for groups among the population with common background characteristics and thus the usage of the PS method here can be considered as justified in the sense of (Pearl, 2011). While compliers are not identifiable, one can at least identify some of the insisters and refusers after the trial has been conducted. Studying them in more detail may give insight as to why noncompliance behavior was exhibited (e.g. to identify relevant covariates of noncompliance) and would allow for future trials to either model compliance explicitly, and possibly animate patients to comply.

It was shown that all of these stratum-specific metrics are not point-identifiable, but are however still partially identifiable. This follows from the outlined problem of not being able to point-identify e.g. $\pi_{11}^{c,1}$, i.e. not being able to "split" the sum of the probabilities $\left(\pi_{11}^{c,1} + \pi_{11}^{i,1}\right)$ corresponding to the cell of patients with both diagnostic tests positive and $D = 1$. However, using the additional information from a nested diagnostic test allows the partial identification of these stratum-specific ratios. Inference for such partially identified parameters poses challenges, in particular when constructing CI for them.

With regards to the design considerations in Section 3.3, it is worth mentioning that other costs — such as fixed costs and administrative costs — should be taken in consideration when applying the decision rule in practice. Naturally, there will be also some restrictions on the number of patients that can be put in the "cheaper" arm.

Considering the assumptions made in Section 2.2, one might question in particular the assumption that patients do not undergo an additional test if the assigned one has been positive (Assumption 2.2.5). There might indeed exist patients who are "hard-to-convince" (and thus will do both tests, regardless of the outcome of the assigned one). These might be included as an additional, separate stratum. Another issue might arise if the decision $Q(T(Z))$ of whether to comply

is also influenced directly by Z as well (and not through $T(Z)$ alone), i.e. whether one should have $Q(T(Z), Z)$. This, however, should not be the case with blinded, that is, as long as the patient does not know his group assignment.

The inferential procedure for the confidence intervals can be improved upon in future work. The major problem for an analytic solution is the fact that we have non-differentiable, albeit continuous functions for the lower and the upper bound. This is a general problem and prevents the usage of standard procedures such as the delta-method to obtain the asymptotic distributions. Existing procedures such as the ones in Imbens and Manski (2004) and in Stoye (2009) require joint normality, which is clearly not the case for the distribution of $min(\widehat{g}_1, \widehat{g}_2)$ and $max(\widehat{g}_1, \widehat{g}_2)$. Overall, constructing CI for partially identified parameters is difficult and generally one must resort to some sort of a bootstrap procedure or simulation (see Tamer (2010) for a general discussion and Chernozhukov et al. (2007) for a procedure based on simulation). Even simple Bonferroni-type of bounds which try to work with the marginal distribution of the two bounds and use the analytic results for their distributions from Clark (1961), do not give satisfying results in terms of empirical coverage: in fact, some preliminary simulation results show that while the theoretical means of the distributions of the minimum and the maximum of $\left(g_1(\widehat{\boldsymbol{p}}^{NC}), g_2(\widehat{\boldsymbol{p}}^{NC})\right)$ do give the true minimum and maximum as $n \to \infty$, they cannot be used to construct an asymptotically pivotal quantity for the CI of LB_c and UB_c for reasonable sample sizes at all values of $\mathcal{I}_c$.

Naturally, in this case, bootstrap procedures can be also considered. However, a major problem might be that we cannot know a-priori whether the parameter of interest (in our case, the function of interest $rTPR_c$) is actually point-identified or not. Andrews (2000) warns against using bootstrap procedures for cases where the parameter of interest might be on the border of the identification region, which would be the situation if $rTPR_c = 1$. This is an issue which warrants further investigation.

Setting this potential problem aside, an empirical bootstrap procedure (e.g. following van der Vaart (2000) and resampling the multinomial data, calculating the minimum and the maximum for M repetitions and then selecting the desired quantiles) would be the next logical step to enhance the inference section for the CI. A comparison can then be made with the current analytic solution.

Another interesting approach to compare results with, would be a bootstrap procedure developed to deal with nondifferentiable functions of the underlying parameters by Woutersen and Ham (2013).

While bootstrap procedures do exist, an improved analytic solution is still to be developed. The expressions for the expected value of the minimum and the maximum in Clark (1961) could potentially be used, as one can show that the sample version of the function $E(\widehat{LB_c})$ and $E(\widehat{UB_c})$, which I call $\widehat{E}_{min}$ and $\widehat{E}_{max}$ are consistent estimators for LB_c and UB_c.

$$\widehat{E}_{min} = \widehat{g}_1 \cdot \Phi\left(\frac{\widehat{g}_2 - \widehat{g}_1}{\widehat{\theta}}\right) + \widehat{g}_2 \cdot \Phi\left(\frac{\widehat{g}_1 - \widehat{g}_2}{\widehat{\theta}}\right) - \widehat{\theta} \cdot \Phi\left(\frac{\widehat{g}_2 - \widehat{g}_1}{\widehat{\theta}}\right)$$
$$\widehat{E}_{max} = \widehat{g}_1 \cdot \Phi\left(\frac{\widehat{g}_1 - \widehat{g}_2}{\widehat{\theta}}\right) + \widehat{g}_2 \cdot \Phi\left(\frac{\widehat{g}_2 - \widehat{g}_1}{\widehat{\theta}}\right) + \widehat{\theta} \cdot \Phi\left(\frac{\widehat{g}_1 - \widehat{g}_2}{\widehat{\theta}}\right)$$

where $\theta = \sqrt{\frac{\left(\sigma^2_{g_1} + \sigma^2_{g_2} - 2 \cdot \sigma_{g_1 g_2}\right)}{n}}$ and $\widehat{\theta}$ is its sample equivalent.

They can be thus potentially used as (smoother) estimators for $LB_c = min(g_1, g_2)$ and $UB_c = max(g_1, g_2)$ to build a Bonferroni-type CI for $\widehat{E}_{min}$ and $\widehat{E}_{max}$. Indeed, it can be conjected that for any *fixed* n and any given set of underlying parameters in the Multinomial model with noncompliance, $E_{min} \leq min(g_1, g_2)$ and similarly $E_{max} \geq max(g_1, g_2)$. The difficulty lies in deriving the distribution of, e.g. $\widehat{E}_{min}$. However, if this can be done, then one would expect an improvement on the analytic CI procedure outlined in Chapter 3, as here the correlation between g_1 and g_2 would be accounted for in the estimator $\widehat{E}_{min}$.

A Bayesian inferential procedure for calculating credible sets (see, e.g. Gelman et al. (1995)) could also be taken into consideration. In the Bayesian framework, all information necessary for inference is contained in the posterior distribution. A straightforward way of inferring for θ is again via Bonferroni-type of bounds. Thus, as a first step, one could consider a Bayesian inferential procedure for LB_c and UB_c separately. The observed data distribution is a contingency table and can be naturally modeled by a Multinomial with parameter vector $\boldsymbol{p}^{NC}$ as outlined in Section 2.2. A choice for conjugate prior for this model would be the Dirichlet distribution. Denoting the data with $\boldsymbol{y}$, the data distribution as $f(\boldsymbol{y}|\boldsymbol{p}^{NC})$, the prior distribution as $p(\boldsymbol{p}^{NC})$ and the posterior as $h(\boldsymbol{p}^{NC}|\boldsymbol{y})$ by the Bayes theorem we then have for the posterior $h(\boldsymbol{p}^{NC}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{p}^{NC}) \cdot p(\boldsymbol{p}^{NC})$ which would be also Dirichlet. However, this is not the posterior that we need to infer for LB_c and

UB_c. Indeed, by recalling that $LB_c = max\left(g_1(\boldsymbol{p}^{NC}), g_2(\boldsymbol{p}^{NC})\right)$ what needs to be considered is the posterior in

$$\tilde{h}\left(max\left(g_1(\boldsymbol{p}^{NC}), g_2(\boldsymbol{p}^{NC})\right)|\boldsymbol{y}\right) \propto$$
$$f\left(\boldsymbol{y}|max\left(g_1(\boldsymbol{p}^{NC}), g_2(\boldsymbol{p}^{NC})\right)\right) \cdot \tilde{p}\left(\left(max\left(g_1(\boldsymbol{p}^{NC}), g_2(\boldsymbol{p}^{NC})\right)\right)\right)$$

Instead, an alternative could be to use Monte Carlo based methods to estimate, e.g. the LB_c function from the posterior $h(\boldsymbol{p}^{NC}|\boldsymbol{y})$ and then proceed to construct credible sets. A more thorough investigation of such an approach is beyond the scope of this thesis and would be left for future work.

Lastly, a further topic of interest seems to be the inclusion of pre-screening noncompliance (i.e. noncompliance with the assignment) in the analysis. This might take form of an extension of the work done by Baker (2000). One could think of creating "cross-strata" of patients who e.g. comply with the screening diagnostic test assignment (i.e. assignment to a screening arm) and also comply with the post-screening recommendation for disease status verification. Such patients can be called "compliers-compliers". Additional strata may be built using a similar approach as the one outlined in the thesis, creating cross-strata of e.g. "compliers-insisters", "nevertakers-refusers" etc. A big difficulty in that case would be the curse of dimensionality, as even for the simple case of all-or-none compliance in a two-arm trial with binary diagnostic tests and binary disease status, one can potentially have 16 compliance strata, each with two possible disease statuses. Clearly, reasonable assumptions which reduce the dimensionality should be undertaken in that case.

Since, as outlined in Section 1, the initial motivation behind this work came from a planned (prostate) cancer screening trial, it is only natural to consider a cancer screening trial for a possible real-data application of the outlined methodology. This need not be a trial for prostate cancer. Indeed, biomarker-based diagnostic tests for other types of cancer (e.g. colon cancer) are considered as preferable to physical tests as the former pose less inconvenience to the patients by being much less invasive. It is then a logical consequence to think that situation in which diagnostic tests are nested would arise more often in the future, as more and more bio-markers become available.

A unified methodology which possibly takes into account all three outlined types of noncom-

pliance in a typical (cancer) screening trial is desirable, as these different kinds of deviations from the protocol are observed in practice.

Appendix A

Appendix

Proof of Proposition 2.2.3. By using the observed data $R^D, T(Z), D$ and Z, I can estimate $rTPR_{i+c}$:

$$\frac{P(R^D(Z)=1\cap T(Z)=1\cap D=1|Z=1)}{P(R^D(Z)=1\cap T(Z)=1\cap D=1|Z=0)}=$$
$$\frac{P(R^D(Z)=1,T(z)=1,D=1,S=\text{“}i+c\text{”}|Z=1)+P(R^D(Z)=1,T(z)=1,D=1,S=\text{“}r\text{”}|Z=1)}{P(R^D(Z)=1,T(Z)=1,D=1,S=\text{“}i+c\text{”}|Z=1)+P(R^D(Z)=1,T(Z)=1,D=1,S=\text{“}r\text{”}|Z=0)}$$
$$=\frac{P(R^D(z)=1,T(z)=1,D=1,S=\text{“}i+c\text{”}|Z=1)+0}{P(R^D(Z)=1,T(Z)=1,D=1,S=\text{“}i+c\text{”}|Z=0)+0}$$
$$=\frac{P(R^D(Z)=1|T(Z)=1,D=1,S=\text{“}i+c\text{”},Z=1)}{P(R^D(Z)=1|T(Z)=1,D=1,S=\text{“}i+c\text{”},Z=0)}\cdot\frac{P(T(Z)=1,D=1,S=\text{“}i+c\text{”}|Z=1)}{P(T(Z)=1,D=1,S=\text{“}i+c\text{”}|Z=0)}$$
$$=\frac{P(T_1=1,D=1,S=\text{“}i+c\text{”})}{P(T_0=1,D=1,S=\text{“}i+c\text{”})}$$

■

Proof of Proposition 2.2.5:

For the lower bound consider:

$$\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}}\leq 1\Leftrightarrow\pi_{11}^{c,1}+\pi_{10}^{c,1}\leq\pi_{11}^{c,1}+\pi_{01}^{c,1}\Leftrightarrow\frac{\pi_{11}^{c,1}+\pi_{10}^{c,1}}{\pi_{11}^{c,1}+\pi_{01}^{c,1}}\leq 1$$

Consider also

$$\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}} \leq 1 \Leftrightarrow \pi_{10}^{c,1} \cdot \pi_{11}^{c,1} \leq \pi_{01}^{c,1} \cdot \pi_{11}^{c,1}$$
$$\Leftrightarrow \pi_{10}^{c,1} \cdot \pi_{11}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{01}^{c,1} \leq \pi_{01}^{c,1} \cdot \pi_{11}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{01}^{c,1}$$
$$\Leftrightarrow \pi_{10}^{c,1}(\pi_{11}^{c,1} + \pi_{01}^{c,1}) \leq \pi_{01}^{c,1}(\pi_{11}^{c,1} + \pi_{10}^{c,1}) \Leftrightarrow \frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}} \leq \frac{(\pi_{11}^{c,1} + \pi_{10}^{c,1})}{(\pi_{11}^{c,1} + \pi_{01}^{c,1})}$$

and

$$\frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}} \leq 1 \Leftrightarrow \frac{\pi_{10}^{c,1}}{\pi_{01}^{c,1}} \leq \frac{\pi_{11}^{i}}{\pi_{11}^{i}} \Leftrightarrow \pi_{10}^{c,1} \cdot \pi_{11}^{i} \leq \pi_{01}^{c,1} \cdot \pi_{11}^{i}$$
$$\Leftrightarrow \pi_{11}^{c,1} \cdot \pi_{11}^{i} + \pi_{11}^{c,1} \cdot \pi_{11}^{c,1} + \pi_{11}^{c,1} \cdot \pi_{01}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{11}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{01}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{11}^{i} \leq$$
$$\Leftrightarrow \pi_{01}^{c,1} \cdot \pi_{11}^{i} + \pi_{11}^{c,1} \cdot \pi_{11}^{i} + \pi_{11}^{c,1} \cdot \pi_{11}^{c,1} + \pi_{11}^{c,1} \cdot \pi_{01}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{11}^{c,1} + \pi_{10}^{c,1} \cdot \pi_{01}^{c,1}$$
$$\Leftrightarrow \frac{\pi_{11}^{c,1} + \pi_{10}^{c,1}}{\pi_{11}^{c,1} + \pi_{01}^{c,1}} \leq \frac{\pi_{11}^{i} + \pi_{11}^{c,1} + \pi_{10}^{c,1}}{\pi_{11}^{i} + \pi_{11}^{c,1} + \pi_{01}^{c,1}}$$

■

For convenience, in order to find the asymptotic variances, I choose to work with the full multinomial data generating model distribution (i.e. including the assignment Z as a level, alongside other patient features $T(Z)$, $T(1-Z)$, D, R^D and S, R^S). The expressions shown here were partially are derived using `Mathematica 7.0`.

Derivations for expressions in Chapter 3, Section 3.1: inference for the model under full compliance

I now derive the observed data distribution based on a full multinomial model, i.e. including the assignment (Z) as a level. Let $\boldsymbol{Y}$ collect all recorded patient data: $\boldsymbol{Y} = \left(R^D, T(Z), T(1-Z), D, Z\right)$ and assume that for a sample of n patients I have $\boldsymbol{Y}_1 \ldots \boldsymbol{Y}_n \overset{i.i.d}{\sim} f(\boldsymbol{Y})$. Since the patients' features

represent categorical data, the levels of $\boldsymbol{Y}$ define a contingency table. For the observed data distribution it then follows $f(\boldsymbol{Y^{obs}}) = \int f(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}|Z) \cdot P(Z) d\boldsymbol{Y}^{mis}$. Integrating over the missing values results into collapsing of cells of the contingency table defined by the levels of $\boldsymbol{Y}$.

In the following Table A.1 I define each element of the vector of the parameters of the observed data distribution ($\boldsymbol{\tau}^F$) from the full multinomial model and link them to the true underlying probability model $\boldsymbol{\pi}^F$ which I have defined in the main text.

τ_1^F	$= P(T_1 = 1, T_0 = 1, D = 1, Z = 1) =$	$(\pi_{11}^{c,1}) \cdot P(Z = 1)$
τ_2^F	$= P(T_1 = 1, T_0 = 0, D = 1, Z = 1) =$	$(\pi_{10}^{c,1}) \cdot P(Z = 1)$
τ_3^F	$= P(T_1 = 1, T_0 = 1, D = 0, Z = 1) =$	$(\pi_{11}^{c,0}) \cdot P(Z = 1)$
τ_4^F	$= P(T_1 = 1, T_0 = 0, D = 0, Z = 1) =$	$(\pi_{10}^{c,0}) \cdot P(Z = 1)$
τ_5^F	$= P(T_1 = 0, T_0 = 1, Z = 1) =$	$(\pi_{01}^{c,1} + \pi_{01}^{c,0}) \cdot P(Z = 1)$
τ_6^F	$= P(T_1 = 0, T_0 = 0, Z = 1) =$	$(\pi_{00}^{c,1} + \pi_{00}^{c,0}) \cdot P(Z = 1)$
τ_7^F	$= P(T_0 = 1, D = 1, Z = 0) =$	$(\pi_{11}^{c,1} + \pi_{01}^{c,1}) \cdot P(Z = 0)$
τ_8^F	$= P(T_0 = 0, T_0 = 1, Z = 0) =$	$\begin{pmatrix} \pi_{10}^{c,1} + \pi_{10}^{c,0} \\ +\pi_{00}^{c,1} + \pi_{00}^{c,0} \end{pmatrix} \cdot P(Z = 0)$
τ_9^F	$= P(T_0 = 1, D = 0, Z = 0) =$	$(\pi_{11}^{c,0} + \pi_{01}^{c,0}) \cdot P(Z = 0)$

Table A.1: Parametrization $\boldsymbol{\tau}^F$ of the observed full multinomial model as a function of the underlying true population parameter vector $\boldsymbol{\pi}^F$ and of $P(Z)$.

Proof Proof of Proposition 3.1.1

Let $\tau^F = \left(\tau_1^F, \tau_2^F, \tau_3^F, \tau_4^F, \tau_5^F, \tau_6^F, \tau_7^F, \tau_8^F, \tau_9^F\right)$. Letting $\boldsymbol{\Sigma} = diag\left(\boldsymbol{\tau}^F\right) - \boldsymbol{\tau}^F \left(\boldsymbol{\tau}^F\right)^T$, by the multivariate central limit theorem (see e.g. Billingsley (1995)) I have that $\sqrt{n}\left(\widehat{\boldsymbol{\tau}}^F - \boldsymbol{\tau}^F\right) \xrightarrow{D} (\mathbf{0}, \boldsymbol{\Sigma})$. The detailed calculations are shown below.

■

For $rTPR^F$ I use $rTPR^F(\widehat{\boldsymbol{\tau}}^F) = \left(\frac{\widehat{\tau}_1^F + \widehat{\tau}_2^F}{\widehat{P}(Z=1)}\right) / \left(\frac{\widehat{\tau}_7^F}{\widehat{P}(Z=0)}\right)$ as an estimator.

Note that $P(Z = 0) = \tau_7^F + \tau_8^F + \tau_9^F$. I can use the delta method to find the asymptotic distribution of that estimator.

$$\nabla rTPR^F(\widehat{\boldsymbol{\tau}}^F)\Big|_{\widehat{\tau}^F=\tau^F} = \left(\frac{\partial rTPR^F(\cdot)}{\partial \tau_1^F}; \frac{\partial rTPR^F(\cdot)}{\partial \tau_2^F}; \frac{\partial rTPR^F(\cdot)}{\partial \tau_7^F}; \frac{\partial rTPR^F(\cdot)}{\partial \tau_8^F}; \frac{\partial rTPR^F(\cdot)}{\partial \tau_9^F}\right)^T$$

$$\frac{\partial rTPR^F(\cdot)}{\partial \tau_1^F} = \frac{P(Z=0)}{\tau_7^F \cdot P(Z=1)}$$

$$\frac{\partial rTPR^F(\cdot)}{\partial \tau_2^F} = \frac{P(Z=0)}{\tau_7^F \cdot P(Z=1)}$$

$$\frac{\partial rTPR^F(\cdot)}{\partial \tau_7^F} = \frac{\left(\tau_1^F + \tau_2^F\right) \cdot \left(P(Z=0) \cdot P(Z=1) + \tau_7^F\right)}{\left(\tau_7^F\right)^2 \cdot P(Z=1)^2}$$

$$\frac{\partial rTPR^F(\cdot)}{\partial \tau_8^F} = \frac{\left(\tau_1^F + \tau_2^F\right)}{\left(\tau_7^F\right) \cdot P(Z=1)^2}$$

$$\frac{\partial rTPR^F(\cdot)}{\partial \tau_9^F} = \frac{\left(\tau_1^F + \tau_2^F\right)}{\left(\tau_7^F\right) \cdot P(Z=1)^2}$$

For the variance $\left(\sigma_{rTPR}^F/n\right)^2$ I then have

$$\frac{1}{n} \cdot (\sigma_{rTPR^F})^2 = \frac{1}{n} \nabla rTPR^F(\widehat{\boldsymbol{\tau}}^F) \cdot \Sigma \cdot \left(\nabla rTPR^F\left(\widehat{\boldsymbol{\tau}}^F\right)\right)^T$$
$$= \frac{\left(\tau_1^F + \tau_2^F\right)\left(\tau_1^F + \tau_2^F + \tau_7^F - \frac{2\cdot(\tau_1^F+\tau_2^F+\tau_7^F)}{P(Z=1)} + \frac{\tau_1^F+\tau_2^F+(\tau_1^F+\tau_2^F+1)\cdot\tau_7^F}{P(Z=1)^2} - \frac{(\tau_1^F+\tau_2^F)\cdot\tau_7^F}{P(Z=1)^3}\right)}{n \cdot \left(\tau_7^F\right)^3}$$

Next, for the test statistic $\widehat{T}_1^F = \frac{\widehat{\tau}_1}{\widehat{P}(Z=1)} + \frac{\widehat{\tau}_2}{\widehat{P}(Z=1)} - \frac{\widehat{\tau}_7}{\widehat{P}(Z=0)}$, I have the following gradient, evaluated at $\boldsymbol{\tau}^F$:

$$\nabla T_1^F(\widehat{\boldsymbol{\tau}}^F)\Big|_{\widehat{\tau}^F=\tau^F} = \left(\frac{\partial T_1^F(\cdot)}{\partial \tau_1^F}; \frac{\partial T_1^F(\cdot)}{\partial \tau_2^F}; \frac{\partial T_1^F(\cdot)}{\partial \tau_7^F}; \frac{\partial T_1^F(\cdot)}{\partial \tau_8^F}; \frac{\partial T_1^F(\cdot)}{\partial \tau_9^F}\right)^T$$

$$\frac{\partial T_1^F(\cdot)}{\partial \tau_1^F} = \frac{1}{P(Z=1)}$$

$$\frac{\partial T_1^F(\cdot)}{\partial \tau_2^F} = \frac{1}{P(Z=1)}$$

$$\frac{\partial T_1^F(\cdot)}{\partial \tau_7^F} = \frac{\left(\tau_1^F + \tau_2^F\right)}{P(Z=1)^2} - \frac{P(Z=0) - \tau_7^F}{P(Z=0)^2}$$

$$\frac{\partial T_1^F(\cdot)}{\partial \tau_8^F} = \frac{\left(\tau_1^F + \tau_2^F\right)}{P(Z=1)^2} - \frac{\tau_7^F}{P(Z=0)^2}$$

$$\frac{\partial T_1^F(\cdot)}{\partial \tau_9^F} = \frac{\left(\tau_1^F + \tau_2^F\right)}{P(Z=1)^2} - \frac{\tau_7^F}{P(Z=0)^2}$$

Again, using the delta method, I can find that under the null of $T_1^F \leq 0$, I have for its variance:

$$\left(\sigma_{1,H_0}^F/n\right)^2 = \frac{\tau_7^F \cdot (\tau_8 - \tau_9)}{n \cdot P(Z=1) \cdot P(Z=0)^3}$$

and under the alternative of $T_1^F > 0$ the variance of the test statistic is:

$$\left(\sigma_1^F/n\right)^2 = \frac{\tau_7^F(\tau_8^F + \tau_9^F)}{n \cdot P(Z=0)^3} + \frac{(\tau_1^F - \tau_2^F) \cdot \left(P(Z=1) - \tau_1^F - \tau_2^F\right)}{n \cdot P(Z=1)^3}$$

For the test statistic $\widehat{T}_2^F$ the gradient evaluated at $\boldsymbol{\tau}^F$ has the following expression:

$$\nabla T_2^F(\widehat{\boldsymbol{\tau}}^F)\Big|_{\widehat{\tau}^F=\tau^F} = \left(\frac{\partial T_2^F(\cdot)}{\partial \tau_2^F}; \frac{\partial T_2^F(\cdot)}{\partial \tau_3^F}; \frac{\partial T_2^F(\cdot)}{\partial \tau_5^F}; \frac{\partial T_2^F(\cdot)}{\partial \tau_7^F}; \frac{\partial T_2^C(\cdot)}{\partial \tau_8^F}; \frac{\partial T_2^F(\cdot)}{\partial \tau_9^F}\right)^T =$$

$$\begin{pmatrix} \frac{1}{P(Z=1)} \\ -\frac{1}{P(Z=1)} \\ -\frac{1}{P(Z=1)} \\ -\frac{-\tau_2^F+\tau_3^F+\tau_5^F}{P(Z=1)^2} - \frac{\tau_8^F}{P(Z=0)^2} \\ \frac{\tau_7^F+\tau_9^F}{P(Z=0)^2} - \frac{-\tau_2^F+\tau_3^F+\tau_5^F}{P(Z=1)^2} \\ -\frac{-\tau_2^F+\tau_3^F+\tau_5^F}{P(Z=1)^2} - \frac{\tau_8^F}{P(Z=0)^2} \end{pmatrix}$$

Under the null of $T_2^F(\cdot) \leq 0$, I have for its variance

$$\left(\sigma^C_{2,H_0}/n\right)^2 = \frac{2\tau_3^F + 2\tau_5^F + \tau_8^F\left(\frac{-3\cdot(P(Z=0))^2+\tau_8^F\cdot P(Z=0)+\tau_7^F+\tau_9^F}{P(Z=0)^3}+2\right)}{n\cdot P(Z=1)^2}$$

and under the alternative hypothesis $T_2^F(\cdot) > 0$

$$\left(\sigma_2^F/n\right)^2 = \frac{1}{n}\cdot\left(\frac{\tau_8^F(\tau_7^F+\tau_9^F)}{P(Z=0)^3} - \frac{(\tau_2^F)^2+(-2\tau_3^F-2\tau_5^F-P(Z=1))\tau_2^F+(\tau_3^F+\tau_5^F)(\tau_3^F+\tau_5^F-P(Z=1))}{P(Z=1)^3}\right)$$

The derivations for the $rFPR^F$ and the corresponding test statistics follow a completely analogous approach, with the simple adjustment of conditioning on $d=0$ instead of $d=1$. As an example,

$$\widehat{rFPR}^F = \frac{\frac{\widehat{\tau}_3^F+\widehat{\tau}_4^F}{\widehat{P}(Z=1)}}{\frac{\widehat{\tau}_9^F}{\widehat{P}(Z=0)}}$$

The corresponding test statistics (with and without using the nestedness property are)

$$\widehat{T}_1^{F,rFPR} = \Big((\widehat{\tau}_3 + \widehat{\tau}_4^F)/\widehat{P}(Z=1)\Big) - \widehat{\tau}_9^F/\widehat{P}(Z=0)$$
$$\widehat{T}_2^{F,rFPR} = \Big((\widehat{\tau}_4^F - \widehat{\tau}_1^F - \widehat{\tau}_5^F)/\widehat{P}(Z=1)\Big) + \widehat{\tau}_7^F/\widehat{P}(Z=0)$$

Using the delta method as shown above, the large sample variances of these test statistics and estimator can be easily obtained by using analogous steps as for the $rTPR^F$.

Derivations for expressions in Chapter 3, Section 3.2: inference for the model with post-screening noncompliance

To derive the observed data distribution, note that D is missing nonignorably, as R^D is a function of the typically unknown compliance stratum S, which in turn is a function of D. Let $\boldsymbol{Y} = (Y^{obs}, Y^{mis}) = (T(Z), T(1-Z), D, S, Z)$ collect the relevant features of the patients and let $\boldsymbol{R} = (R^{T(Z)}, R^{T(1-Z)}, R^D, R^S)$ be a vector of corresponding indicators for missingness. Each element of $\boldsymbol{R}$ has the value of 1 if the value of the corresponding variable in $\boldsymbol{Y}$ is observed.

I follow Little and Rubin (2002) and consider the joint p.m.f. of the patient features data $\boldsymbol{Y}$ and the vector of the missingness indicators $\boldsymbol{R}$, i.e. $f(\boldsymbol{R}, \boldsymbol{Y})$. Now suppose that an i.i.d sample of n patients is available the data analyst, such that $(\boldsymbol{R}, \boldsymbol{Y})_1, \ldots, (\boldsymbol{R}, \boldsymbol{Y})_n \overset{i.i.d}{\sim} f_{\boldsymbol{R},\boldsymbol{Y}}$. I integrate over the missing data

$$f_{(\boldsymbol{R},\boldsymbol{Y}^{obs})_1,\ldots,(\boldsymbol{R},\boldsymbol{Y}^{obs})_n}(\cdot) = \int\limits_{\boldsymbol{Y}^{mis}} \prod_{j=1}^{n} f(\boldsymbol{r}_j, \boldsymbol{y}_j) d\boldsymbol{Y}^{mis} = \iiint \prod_{j=1}^{n} f(\mathbf{Y}, \mathbf{R})_j d\mathbf{Y}_\mathbf{1}^\mathbf{mis} \ldots d\mathbf{Y}_\mathbf{n}^\mathbf{mis}$$

It can be shown that the resulting expression is proportional to the likilehood of a

multinomial distribution with a set of parameters $(\boldsymbol{\tau}^{NC})$ which are linear combinations of the true underlying set of parameters $\boldsymbol{\pi}^{NC}$ and of $P(Z)$. Without loss of generality, I give the expressions for the elements of $(\boldsymbol{\tau}^{NC}) = (\tau_1^{NC}, \ldots, \tau_{16}^{NC})^T$ in Table A.2 and link them to the parameters of the true underlying probability model $(\boldsymbol{\pi}^{NC})$.

$\tau_1^{NC} = \left(\pi_{11}^{c,1} + \pi_{11}^{i,1}\right) \cdot P(Z=1)$
$\tau_2^{NC} = \left(\pi_{10}^{c,1} + \pi_{10}^{i,1}\right) \cdot P(Z=1)$
$\tau_3^{NC} = \left(\pi_{01}^{i,1}\right) \cdot P(Z=1)$
$\tau_4^{NC} = \left(\pi_{11}^{c,0} + \pi_{11}^{i,0}\right) \cdot P(Z=1)$
$\tau_5^{NC} = \left(\pi_{10}^{c,0} + \pi_{10}^{i,0}\right) \cdot P(Z=1)$
$\tau_6^{NC} = \left(\pi_{01}^{i,0}\right) \cdot P(Z=1)$
$\tau_7^{NC} = \left(\pi_{11}^{r,1} + \pi_{11}^{r,0}\right) \cdot P(Z=1)$
$\tau_8^{NC} = \left(\tau_{10}^{r,1} + \tau_{10}^{r,0}\right) \cdot P(Z=1)$
$\tau_9^{NC} = (\pi_{01}^{c,1} + \pi_{01}^{c,0} + \pi_{01}^{r,1} + \pi_{01}^{r,0}) \cdot P(Z=1)$
$\tau_{10}^{NC} = (\pi_{00}^{c,1} + \pi_{00}^{c,0} + \pi_{00}^{r,1} + \pi_{00}^{r,0} + \pi_{00}^{i,1} + \pi_{00}^{i,0}) \cdot P(Z=1)$
$\tau_{11}^{NC} = (\pi_{11}^{c,1} + \pi_{11}^{i,1} + \pi_{01}^{c,1} + \pi_{01}^{i,1}) \cdot P(Z=0)$
$\tau_{12}^{NC} = (\pi_{10}^{i,1}) \cdot P(Z=0)$
$\tau_{13}^{NC} = \left(\pi_{10}^{c,1} + \pi_{10}^{c,0} + \pi_{10}^{r,1} + \pi_{10}^{r,0} + \pi_{00}^{c,1} + \pi_{00}^{c,0} + \pi_{00}^{r,1} + \tau_{00}^{r,0} + \tau_{00}^{i,0} + \tau_{00}^{i,1}\right) \cdot P(Z=0)$
$\pi_{14}^{NC} = \left(\pi_{11}^{r,1} + \pi_{11}^{r,0} + \pi_{01}^{r,1} + \pi_{01}^{r,0}\right) \cdot P(Z=0)$
$\tau_{15}^{NC} = \left(\pi_{10}^{i,0}\right) \cdot P(Z=0)$
$\tau_{16}^{NC} = \left(\pi_{11}^{c,0} + \pi_{01}^{c,0} + \pi_{11}^{i,0} + \pi_{01}^{i,0}\right) \cdot P(Z=0)$

Table A.2: Parametrization $\boldsymbol{\tau}^{NC}$ of the observed full multinomial as a function the underlying true population parameter vector $\boldsymbol{\pi}^{NC}$ and $P(Z)$.

Proof of Proposition 3.2.1

In the case of post-screening noncompliance, let $\tau^{NC} = \left(\tau_1^{NC}, \ldots, \tau_{16}^{NC}\right)$. Letting $\boldsymbol{\Sigma} = diag\left(\boldsymbol{\tau}^{NC}\right) - \boldsymbol{\tau}^{NC}\left(\boldsymbol{\tau}^{NC}\right)^T$, again by the multivariate central limit theorem I have that

$$\sqrt{n}\left(\widehat{\boldsymbol{\tau}}^{NC} - \boldsymbol{\tau}^{NC}\right) \xrightarrow{D} (\mathbf{0}, \boldsymbol{\Sigma})$$

The detailed calculations for the variances and the test statistics are shown below.

■

For $rTPR_{i+c}$ I use $rTPR^{NC}_{i+c}(\widehat{\boldsymbol{\tau}}^{\boldsymbol{NC}}) = \left(\frac{\widehat{\tau}^{NC}_1+\widehat{\tau}^{NC}_2}{\widehat{P}(Z=1)}\right) / \left(\frac{\widehat{\tau}^{NC}_{11}}{\widehat{P}(Z=0)}\right)$ as an estimator. Note that when I have noncompliance: $P(Z=0) = \tau^{NC}_{11} + \tau^{NC}_{12} + \tau^{NC}_{13} + \tau^{NC}_{14} + \tau^{NC}_{15} + \tau^{NC}_{16}$. I again use the delta method to find the asymptotic variances of the relevant estimators and test statistics. The gradient of the estimator's parameters, evaluated at the true values of $\boldsymbol{\tau}^{NC}$ is:

$$\nabla rTPR^{NC}_{i+c}(\widehat{\boldsymbol{\tau}}^{NC})\Big|_{\widehat{\tau}^{NC}=\tau^{NC}} =$$

$$= \left(\frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_1}; \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_2}; \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_{11}}; \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_{12}}; \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_{13}}; \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_{14}}, \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_{15}}, \frac{\partial rTPR^{NC}_{i+c}(\cdot)}{\partial \tau^{NC}_{16}}\right)^T$$

$$= \begin{pmatrix} \frac{P(Z=0)}{\tau^{NC}_{11}P(Z=1)} \\ \frac{P(Z=0)}{\tau^{NC}_{11}P(Z=1)} \\ \frac{((\tau^{NC}_{11})^2+2(P(Z=0)-\tau^{NC}_{11})\tau^{NC}_{11}-(P(Z=1)+\tau^{NC}_{11})(P(Z=0)-\tau^{NC}_{11}))(\tau^{NC}_1+\tau^{NC}_2)}{(\tau^{NC}_{11})^2P(Z=1)^2} \\ \frac{\tau^{NC}_1+\tau^{NC}_2}{\tau^{NC}_{11}P(Z=1)^2} \\ \frac{\tau^{NC}_1+\tau^{NC}_2}{\tau^{NC}_{11}P(Z=1)^2} \\ \frac{\tau^{NC}_1+\tau^{NC}_2}{\tau^{NC}_{11}P(Z=1)^2} \\ \frac{\tau^{NC}_1+\tau^{NC}_2}{\tau^{NC}_{11}P(Z=1)^2} \\ \frac{\tau^{NC}_1+\tau^{NC}_2}{\tau^{NC}_{11}P(Z=1)^2} \end{pmatrix}$$

For the asymptotic variance of the estimator I then have:

$$\begin{aligned} \frac{\left(\sigma^{NC}_{rTPR}\right)^2}{n} &= \frac{(P(Z=0))}{n\cdot(\tau^{NC}_{11})^3(P(Z=0)-1)^3}\left((\tau^{NC}_{11})^3+\right. \\ &(\tau^{NC}_1+2\tau^{NC}_{12}+2\tau^{NC}_{13}+2\tau^{NC}_{14}+2\tau^{NC}_{15}+2\tau^{NC}_{16}+ \\ &+\tau^{NC}_2-1)(\tau^{NC}_{11})^2+(P(Z=0)-\tau^{NC}_{11})\cdot \\ &\cdot(2\tau^{NC}_1+\tau^{NC}_{12}+\tau^{NC}_{13}+\tau^{NC}_{14}+\tau^{NC}_{15}+\tau^{NC}_{16}+2\tau^{NC}_2-1)\cdot\tau^{NC}_{11}+ \\ &+(P(Z=0)-\tau^{NC}_{11}-1)\cdot \\ &\left.\cdot(P(Z=0)-\tau^{NC}_{11})(\tau^{NC}_1+\tau^{NC}_2)\right) \end{aligned}$$

The gradient of the test statistic $\widehat{T}^{NC}_1$ evaluated at $\boldsymbol{\tau}^{NC}$is:

$$\nabla T_1^{NC}(\widehat{\boldsymbol{\tau}}^{NC})\Big|_{\widehat{\tau}^{NC}=\tau^{NC}} = \left(\frac{\partial T_1^{NC}(\cdot)}{\partial \tau_1^{NC}}; \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_2^{NC}}; \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_{11}^{NC}}; \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_{12}^{NC}}; \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_{13}^{NC}}; \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_{14}^{NC}}, \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_{15}^{NC}}, \frac{\partial T_1^{NC}(\cdot)}{\partial \tau_{16}^{NC}}\right)^T$$

$$= \begin{pmatrix} \frac{1}{P(Z=1)} \\ \frac{1}{P(Z=1)} \\ \frac{\tau_1^{NC}+\tau_2^{NC}}{P(Z=1)^2} - \frac{P(Z=0)-\tau_{11}^{NC}}{P(Z=0)^2} \\ \frac{\tau_{11}^{NC}}{P(Z=0)^2} + \frac{\tau_1^{NC}+\tau_2^{NC}}{P(Z=1)^2} \\ \frac{\tau_{11}^{NC}}{P(Z=0)^2} + \frac{\tau_1^{NC}+\tau_2^{NC}}{P(Z=1)^2} \\ \frac{\tau_{11}^{NC}}{P(Z=0)^2} + \frac{\tau_1^{NC}+\tau_2^{NC}}{P(Z=1)^2} \\ \frac{\tau_{11}^{NC}}{P(Z=0)^2} + \frac{\tau_1^{NC}+\tau_2^{NC}}{P(Z=1)^2} \\ \frac{\tau_{11}^{NC}}{P(Z=0)^2} + \frac{\tau_1^{NC}+\tau_2^{NC}}{P(Z=1)^2} \end{pmatrix}$$

For the asymptotic variance of $\widehat{T}_1^{NC}$ under the null of $T_1^{NC} \leq 0$ I have:

$$\frac{\left(\sigma_{1,H_0}^{NC}\right)^2}{n} = \frac{\tau_{11}^{NC} \cdot \left(P(Z=0) - \tau_{11}^{NC}\right)}{n \cdot P(Z=1) \cdot P(Z=0)^3} \tag{A.1}$$

For the asymptotic variance of $\widehat{T}_1^{NC}$ under the alternative I have

$$\frac{\left(\sigma_1^{NC}\right)^2}{n} = \frac{1}{n} \cdot \left(\frac{(\tau_1^{NC})^2}{(P(Z=0)-1)^3} + \frac{(P(Z=0)+2\tau_2^{NC}-1)\tau_1^{NC}}{(P(Z=0)-1)^3} + \frac{\tau_2^{NC}(P(Z=0)+\tau_2^{NC}-1)}{(P(Z=0)-1)^3} + \right. \tag{A.2}$$

$$\left. \frac{\tau_{11}^{NC}(P(Z=0)-\tau_{11}^{NC})^2}{(P(Z=0))^4} + \frac{(\tau_{11}^{NC})^2(P(Z=0)-\tau_{11}^{NC})}{(P(Z=0))^4} \right) \tag{A.3}$$

The expressions for the derivatives and variances under the null and the alternative for $T_2^{NC}(\cdot)$ are too long to be reported here, but the approach to their derivation is analogous to the one described above.

The derivations for the $rFPR_{i+c}^{NC}$ and the corresponding test statistics follow a completely analogous approach, with the simple adjustment of conditioning on $D = 0$ instead of $D = 1$.

$$\widehat{rFPR}^{NC} = \frac{\frac{\widehat{\tau}_4^{NC}+\widehat{\tau}_5^{NC}}{\widehat{P}(Z=1)}}{\frac{\widehat{\tau}_{16}^{NC}}{\widehat{P}(Z=0)}}$$

The corresponding test statistics (with and without using the nestedness property are)

$$\widehat{T}_1^{NC,rFPR} = \left((\widehat{\tau}_4^{NC} + \widehat{\tau}_5^{NC})/\widehat{P}(Z=1)\right) - \widehat{\tau}_{16}^{NC}/\widehat{P}(Z=0)$$
$$\widehat{T}_2^{NC,rFPR} = \left((\widehat{\tau}_5^{NC} - \widehat{\tau}_9^{NC} - \widehat{\tau}_1^{NC} - \widehat{\tau}_3^{NC} - \widehat{\tau}_7^{NC} - \widehat{\tau}_6^{NC})/\widehat{P}(Z=1)\right) + \left(\widehat{\tau}_{14}^{NC} + \widehat{\tau}_{11}^{NC}\right)/\widehat{P}(Z=0)$$

Using the delta method as described above, the large sample variances of these test statistics and estimator can be easily obtained using the analogous steps as for the $rTPR_{i+c}^{NC}$.

Proof (sketch) of Proposition 3.3.1 Considering the power function of $T_1^F(\cdot)$, we have that

$$\gamma(\cdot) = 1 - \left(\frac{\frac{\left(\sigma^2_{1,H_0}(\boldsymbol{\pi}_0^F, P(Z=0)\right)}{n_1+n_0} \cdot z_{1-\alpha} - \pi_{10}^{c,1} + \pi_{01}^{c,1}}{\frac{\left(\sigma_1^2(\boldsymbol{\pi}_1^F, P(Z=0)\right)}{(n_1+n_0)}} \right) = 1 - \beta$$

It is clear from the elements in Table A.1, that the asymptotic variances of $\widehat{T}_1^F$ under the null and the alternative, depend both on $P(Z=0)$ and on the vector of true probabilities (respectively under the null and the alternative).

Clearly, $P(Z=0) = n_0/(n_1+n_0)$. It is thus not clear from the expression for the power, whether the function would be monotonically increasing in n_0 for all possible values of $\boldsymbol{\pi_1^F}$ and $\boldsymbol{\pi_0^F}$, given a fixed n_1.

Following the approach in the text and fixing n_1, I treat $\gamma(\cdot)$ as a (continuous) function of n_0 only. We can take its derivative towards n_0 using `Mathematica 7.0`.

The result is a long expression and hence not reported here, but available upon request. It can be shown that the derivative is positive (hence the function monotonically increasing) only when the condition in Proposition 3.3.1 holds. By Bolzano's theorem, a monotonically increasing continuous function defined on $(\beta-1, \beta)$ has only one root. Hence, the n_0 which would give the desired level of power would be unique. Practically, one would look for this

root only in realistic ranges using the `uniroot` function of R 3.0.2 (2014).

■

Proof of Proposition 3.2.2.

First, consider showing

$$\{\{g_{1L} \leq g_1\} \cap \{g_{2L} \leq g_2\}\} \Rightarrow \{min(g_{1L}, g_{2L}) \leq min(g_1, g_2)\}$$

Look at the following 3 cases.

- Case 1: The left hand side holds and $g_1 < g_2$. Then, regardless i $min(g_{1L}, g_{2L}) = g_{1L}$ or $min(g_{1L}, g_{2L}) = g_{2L}$; $min(g_{1L}, g_{2L}) \leq min(g_1, g_2) = g_1$ holds.
- Case 2: $g_1 = g_2$ holds. The same logic applies. In fact, the CI point build around the point estimator with the higher variance would be the minimum of g_{1L} and g_{2L}
- Case 3: $g_1 > g_2$ holds. Then, either $g_{1L} \leq g_{2L} \leq g_2$ or $g_{2L} \leq g_{1L} \leq g_2$. Either way, $min(g_{1L}, g_{2L}) \leq min(g_1, g_2)$ holds. Thus, in terms of sets, every point from the set on the left hand side, is a point in the set on the right hand side.

In a similar fashion, it is easy to show that

$$\{\{g_{1U} \geq g_1\} \cap \{g_{2U} \geq g_2\}\} \Rightarrow \{max(g_{1U}, g_{2U}) \geq max(g_1, g_2)\}$$

holds. By the assumption in Proposition 3.2.2 and using Bonferroni inequality we have

$$P(g_{1L} \geq g_1) + P(g_{2L} \geq g_2) + P(g_{1U} \leq g_1) + P(g_{2U} \leq g_2) \geq \tag{A.4}$$
$$P\left(\{g_{1L} \geq g_1\} \cup \{g_{2L} \geq g_2\} \cup \{g_{1U} \leq g_1\} \cup \{g_{2U} \leq g_2\}\right)$$

and by taking the complement we have

$$P\left(\{g_{1L} \leq g_1 \leq g_{1U}\} \cap \{g_{2L} \leq g_2 \leq g_{2U}\}\right) \geq 1 - \alpha$$

$$\Rightarrow$$

$$P\left(\{min(g_{1L}, g_{2L}) \leq min(g_1, g_2)\} \cap \{max(g_{1U}, g_{2U}) \geq max(g_1, g_2)\}\right) \geq 1 - \alpha$$

■

Bibliography

Agresti, A. *Categorical data analysis*, volume 359. John Wiley & Sons, 2002.

Alonzo, T. A. Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests. *Statistics in medicine*, 24(3):403–417, 2005.

Alonzo, T. A. and Kittelson, J. M. A novel design for estimating relative accuracy of screening tests when complete disease verification is not feasible. *Biometrics*, 62(2):605–612, 2006.

Alonzo, T. A. and Pepe, M. S. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1): 173–190, 2005.

Alonzo, T. A., Pepe, M. S., and Moskowitz, C. S. Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in medicine*, 21(6):835–852, 2002.

Alonzo, T. A., Braun, T. M., and Moskowitz, C. S. Small sample estimation of relative accuracy for binary screening tests. *Statistics in medicine*, 23(1):21–34, 2004.

Andrews, D. W. K. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Baker, S. G. Compliance, All-or-None. *Encyclopedia of Statistical Sciences*, 1997.

Baker, S. G. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association*, 93(443):929–934, 1998.

Baker, S. G. Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance. *Journal of the American Statistical Association*, 95(449):43–50, 2000.

Baker, S. G. and Lindeman, K. S. The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statistics in medicine*, 13(21):2269–2278, 1994.

Baker, S. G., Kramer, B. S., and Lindeman, K. S. Latent class instrumental variables: a clinical and biostatistical perspective. *Statistics in medicine*, 2015.

Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, 98(462):299–323, 2003.

Bartolucci, F. and Farcomeni, A. Causal inference in paired two-arm experimental studies under noncompliance with application to prognosis of myocardial infarction. *Statistics in medicine*, 2013.

Billingsley, P. Probability and measure. *Wiley series in probability and mathematical statistics*, 1995.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. *Discrete multivariate analysis: theory and practice*. Springer, 2007.

Broemeling, L. D. *Bayesian biostatistics and diagnostic medicine*. CRC Press, 2007.

Casella, G. and Berger, R. L. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Chen, H., Geng, Z., and Zhou, X.-H. Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics*, 65(3):675–682, 2009.

Cheng, H. and Macaluso, M. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. *Epidemiology*, 8(1):104–106, 1997.

Cheng, J. and Small, D. S. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836, 2006.

Chernozhukov, V., Hong, H., and Tamer, E. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007.

Clark, C. E. The greatest of a finite set of random variables. *Operations Research*, 9(2):145–162, 1961.

Cuzick, J., Edwards, R., and Segnan, N. Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in medicine*, 16(9):1017–1029, 1997.

Cuzick, J., Sasieni, P., Myles, J., and Tyrer, J. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):565–588, 2007.

Davidson, R. and MacKinnon, J. G. Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, 66(1):1–26, 1998.

Frangakis, C. E. and Rubin, D. B. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379, 1999.

Frangakis, C. E. and Rubin, D. B. Principal stratification in causal inference. *Biometrics*, 58(1): 21–29, 2002.

Gareen, I. F. Noncompliance in cancer screening trials. *Clinical Trials*, 4(4):341–349, 2007.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. Bayesian Data Analysis. 1995.

Heckman, J. Shadow prices, market wages, and labor supply. *Econometrica: journal of the econometric society*, pages 679–694, 1974.

Horowitz, J. L. and Manski, C. F. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449):77–84, 2000.

Imbens, G. W. and Angrist, J. D. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Imbens, G. W. and Manski, C. F. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Imbens, G. W. and Rubin, D. B. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997.

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, 2002.

Lui, K.-J. *Binary Data Analysis of Randomized Clinical Trials with Noncompliance*. John Wiley & Sons, 2011.

Manski, C. F. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990.

Manski, C. F. *Partial identification of probability distributions*. Springer, 2003.

Mealli, F. and Mattei, A. A refreshing account of principal stratification. *The international journal of biostatistics*, 8(1), 2012.

Mealli, F. and Pacini, B. Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. *Computational Statistics & Data Analysis*, 53 (2):507–516, 2008.

Mealli, F. and Pacini, B. Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance. *Journal of the American Statistical Association*, 108(503):1120–1131, 2013.

Mealli, F. and Rubin, D. B. Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services and Outcomes Research Methodology*, 3(3-4): 225–232, 2002.

Mealli, F., Imbens, G. W., Ferro, S., and Biggeri, A. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*, 5(2):207–222, 2004.

Nadarajah, S. and Kotz, S. Exact distribution of the max/min of two gaussian random variables. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(2):210–212, 2008.

O'Malley, A. J. and Normand, S.-L. T. Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics*, 61(2):325–334, 2005.

Pearl, J. Principal stratification, a goal or a tool? *The International Journal of Biostatistics*, 7(1), 2011.

Pepe, M. S. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press Oxford, 2003.

Pepe, M. S. and Alonzo, T. A. Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics*, 2(3):249–260, 2001.

Permutt, T. and Hebel, J. R. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, pages 619–622, 1989.

R 3.0.2. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Rao, C. R. Linear statistical inference and its applications. *Wiley series in probability and mathematical statistics*, 1973.

Rosenbaum, P. R. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, pages 656–666, 1984.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

Schatzkin, A., Connor, R. J., Taylor, P. R., and Bunnag, B. Comparing new and old screening tests when a reference procedure cannot be performed on all screenes: example of automated cytometry for early detection of cervical cancer. *American Journal of Epidemiology*, 125(4): 672–678, 1987.

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L. J., Zappa, M., Nelen, V., Kwiatkowski, M., Lujan, M., Määttänen, L., Lilja, H., and Others. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *The Lancet*, 384(9959):2027–2035, 2014.

Schwartz, S. L., Li, F., and Mealli, F. A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, 106(496):1331–1344, 2011.

Stoye, J. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4): 1299–1315, 2009.

Tamer, E. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, 2000.

Woutersen, T. and Ham, J. C. Calculating confidence intervals for continuous and discontinuous functions of parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2013.