NORTHWESTERN UNIVERSITY

Essays in Identification and Inference

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Filip Obradović

EVANSTON, ILLINOIS

June 2025

ABSTRACT

Essays in Identification and Inference

Filip Obradović

This dissertation consists of three essays studying identification and inference in settings pertaining to economics and medicine. The emphasis is on developing and utilizing frameworks that incorporate relevant features of the underlying empirical context which may invalidate frequently imposed assumptions.

The first chapter studies identification of long-term treatment effects. Since long-term experimentation is frequently infeasible, a large body of recent work proposes appending a short-term experimental dataset to longitudinal observational data to provide more credible alternatives to observational studies. As the first contribution, I show that justifiable modeling assumptions remain central for plausible inference despite the addition of the experiment, just as in conventional observational studies. Experimental data bring no identifying power on their own; they serve only to amplify the identifying power of assumptions restricting conditional means of long-term potential outcomes conditional on short-term potential outcomes—temporal link functions. However, existing research argues that previously proposed modeling assumptions may frequently be challenging to

justify. This motivates the second contribution. I introduce the use of treatment response assumptions that only restrict the shape of temporal link functions and may thus be defensible based on economic intuition or theory. As the third contribution, I introduce a novel identification framework which yields the smallest possible, i.e., sharp, bounds on the long-term average treatment effect under a broad class of restrictions on temporal link functions and imperfect compliance in the experiment. The framework produces the bounds via solutions to generalized bilinear problems. It thus: 1) enables the use of the proposed treatment response assumptions; 2) facilitates the development of new justifiable assumptions by removing the need to prove sharpness; 3) extends existing methods to account for imperfect compliance.

The second chapter studies the measurement of misclassification rates of diagnostic tests and general binary classifiers. The rates are of great interest to regulators and clinicians. However, their identification requires knowledge of the underlying ground truth, which is often measured by an imperfect reference test or classifier. The common practice is thus to report misclassification rates with respect to the reference—"apparent" misclassification rates—which do not measure true performance. The first contribution are the sharp bounds on the measures of true performance—sensitivity (true positive rate) and specificity (true negative rate), or equivalently false positive and negative rates, under standard assumptions in performance studies. The second contribution is the construction of uniformly consistent confidence sets in level over a relevant family of data distributions. This allows researchers to account for statistical imprecision, which is typically recommended by relevant regulatory guidelines. As the third contribution, I revisit the performance of the ubiquitous BinaxNOW COVID-19 antigen test, based on

Emergency Use Authorization and independent study data. The analysis reveals that the estimated false negative rates for symptomatic and asymptomatic patients are, respectively, up to 3.17 and 4.59 times higher than the frequently cited "apparent" false negative rate. This finding brings into question whether the test would have met the contemporaneous threshold for Emergency Use Authorization once the imperfections of the reference test are taken into account.

The third chapter (joint work with Gabriel Ziegler) shows that dilation is a real-world phenomenon that may be induced by diagnostic tests under established clinical practice. Clinicians often seek to determine the probability that a patient has a suspected illness conditional on a test result. Dilation entails that conditioning on *any* test result *only* introduces uncertainty about the patient's health status and has been mostly considered a theoretical curiosity. As the first contribution, we show that dilation may be induced by conditioning on diagnostic tests whose misclassification rates are evaluated with respect to an imperfect reference, which is widespread. Moreover, dilation may occur even when tests are approved or recommended based on satisfactory "apparent" misclassification rates, often used as the primary criterion for evaluation. This motivates the second contribution. We enable decision-makers to identify such diagnostic tests by equivalently characterizing when dilation is induced, and providing a statistical testing procedure that is uniformly consistent in level for a large family of relevant data-generating processes. For the third contribution, we study computed tomography (CT) chest scans for detecting COVID-19 infection that were recommended as a primary detection tool in epidemic areas based on conventional "apparent" performance measures. We find that they induced dilation and thus only introduced uncertainty about the patient's health status.

Acknowledgements

I am forever indebted to my mentors, Charles Manski, Ivan Canay, and Federico Bugni, for their selfless support and guidance. Chuck and Ivan have been deeply devoted advisors, persistently pushing me to be an unrelenting skeptic and not to shy away from shooting for the stars. Fede has shown me what true resilience looks like and has taught me not to fear failure. To them, I owe being the researcher that I am.

I was fortunate to be surrounded by an incredible group of academics, faculty and students alike. Joel Horowitz and Eric Auerbach have graciously shared advice and encouragement. Amilcar Velez, beyond being a friend and coauthor, has served as my informal advisor. I have greatly benefited from interacting with many Northwestern econometrics students. To my delight, the group we formed may even be numerous and cohesive enough to render individual acknowledgments impertinent. I am likewise beholden to Lori Beaman, Chris Udry, Seema Jayachandran, and the clinicians and surgeons at the Feinberg School of Medicine, led by Daniela Ladner, for their direction and perspective. It has been a privilege to learn from all of them.

I am thankful that the space here may be too limited to express my appreciation to each dear friend I have made along the way. You have made graduate school a memorable and gratifying experience. I gratefully acknowledge the financial support from Robert Eisner Memorial, Unicredit Crivelli, Northwestern Graduate and Dissertation Year Fellowships.

Lastly, I owe an eternal debt of gratitude to my family and friends in Serbia. My parents, Sandra and Goran, have—against all odds—succeeded in providing me with the opportunity to pursue doing what I love. The difficulty of understanding their sacrifices is surpassed only by the hardships they had to overcome. My wife, Jana, has been a source of unconditional love and encouragement through the most trying times. Without their unwavering support, this dissertation would not have been written. For this, I dedicate it to them.

Dedication

To my parents, Sandra and Goran, and my wife, Jana.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Identification of Long-Term Treatment Effects via Temporal Links, Observational, and Experimental Data

Section 1.1. Introduction

Identification of the long-term average treatment effect (henceforth LTE) is an important goal in economics and various other fields of science. For example, one may be interested in the effects of childhood intervention on outcomes in adulthood; the impact of conditional cash transfers early in life on employment prospects; or the adverse/protective effects of vaccination years after administration. Gupta et al. (2019) explain that identifying the LTE is also recognized as an important challenge by researchers in the private sector.

Point identification of the LTE is commonly done using observational data (for examples, see Currie and Almond (2011), Hoynes and Schanzenbach (2018)). However, observational studies critically rely on modeling or identifying assumptions that may often be deemed implausible. While randomized controlled trials (RCTs) eliminate the need for such assumptions, long-term experiments may be prohibitively costly or infeasible.[1] Short-term RCTs may be more feasible, but they do not reveal the long-term outcomes and hence the LTE. Nevertheless, short-term RCTs may complement observational data.

Motivated by this, a large body of recent work following Athey, Chetty, and Imbens (2020) and Athey et al. (2024) aims to provide credible alternatives to conventional

1. Institutions supporting RCTs in development economics frequently require phase-in designs with staggered rollout of treatment to the whole sample. This limits follow-up for the control group.

observational studies that rely on a combination of: 1) a long-term observational dataset with non-randomized treatment assignment; 2) a short-term experimental dataset with unobserved long-term outcomes.[2] Pursuing point identification, this literature commonly imposes assumptions on the selection mechanism in the observational data, mirroring conventional observational studies. Ghassami et al. (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023) and Imbens et al. (2024) argue that frequently used assumptions may fail in contexts of economic interest; Park and Sasaki (2024a) show that they are incompatible with common selection models, including the Roy model. It is acknowledged that selection assumptions may broadly be challenging to justify based on economic theory. However, existing results do not reveal whether the addition of experimental data provides identifying power in the absence of such assumptions, or whether it can yield results closer to the truth should the assumptions fail.

This paper makes three main contributions. First, it uncovers the roles of the experimental data and modeling assumptions by showing that neither of the two previous claims is true. The addition of experimental data brings no identifying power in the absence of modeling assumptions. Moreover, it may only lead to results that are farther from the ground truth if the modeling assumptions fail. Hence, data combination may provide credible alternatives to observational studies only under plausible modeling assumptions. Motivated by this, the paper introduces two assumptions on *treatment response* that may be defensible based on economic theory or intuition, as the second contribution. Third, it develops a novel two-step identification approach which enables the utilization of such

2. Structural modeling with experimental data predates this work (Todd and Wolpin (2006), Attanasio, Meghir, and Santiago (2012), García et al. (2020), Todd and Wolpin (2023)). The focus here is on "reduced form" methods.

assumptions. The approach bounds latent *temporal link functions* – means of long-term potential outcomes conditional on short-term potential outcome – as an intermediate step.

The first contribution of the paper is that it uncovers the roles of the experimental data and modeling assumptions, which relates to ongoing discussions (see Remark 1.4.1 and Park and Sasaki (2024b)). I show that experimental data provide no identifying power, per se; the identified sets for the LTE obtained from combined and solely observational data are equal in the absence of modeling assumptions. Modeling assumptions are thus *necessary* to leverage the experimental data and to identify the sign of the LTE. When the assumptions are imposed, the identified set based on combined data is a subset of the one that uses only observational data. It need not be a strict subset. Hence, the experimental data serve to potentially, but not necessarily, *amplify* the identifying power of the modeling assumptions. These observations reveal the *auxiliary* role of experimental data. Assumptions in the observational data remain *central* under data combination, mirroring their prominence in observational studies.

I illustrate this point via a selection assumption that is often used in the literature relying on data combination – latent unconfoundedness (LUC), introduced by Athey, Chetty, and Imbens (2020). I show that LUC may commonly have identifying power for the LTE using observational data alone. When experimental data are added, LUC has more identifying power, and point identifies the LTE. The experimental data thus may amplify the identifying power of LUC. In extreme cases, observational data point identify the LTE under LUC even without experimental data. Then, the experimental data do not provide additional identifying power. Therefore, the experimental data potentially, but not necessarily, amplify the identifying power of LUC.

The amplifying role of the experimental data highlights the importance of modeling assumptions. If the imposed assumptions are difficult to justify based on economic substantives, it may be preferable to discard the experimental data. Under misspecified assumptions, the identified set obtained using combined data can never be closer to the true LTE than the set obtained using only observational data. This finding is an application of a more general lemma. The lemma states that whenever two misspecified identified sets are nested, the smaller one must be at least as far from the truth as the larger one. This result is strikingly simple, but it appears that it was not formalized before.

The central role of modeling assumptions motivates the main identification results. To summarize them, let $Y(d) \in \mathcal{Y}$ and $S(d) \in \mathcal{S}$ denote long- and short-term potential outcomes under treatment $d \in \{0, 1\}$ and let:

$$m_d(s) := E[Y(d)|S(d) = s]$$

$$\gamma_d := P(S(d)).$$

I refer to $m_d(s)$ as the *temporal link functions*, while γ_d are the distributions of short-term potential outcomes $S(d)$. We can then write the LTE using the identity:

$$LTE := E[Y(1) - Y(0)] = \underbrace{\int_{\mathcal{S}} m_1(s) d\gamma_1(s)}_{E[Y(1)]} - \underbrace{\int_{\mathcal{S}} m_0(s) d\gamma_0(s)}_{E[Y(0)]}, \tag{1.1}$$

For the second contribution, I introduce two assumptions on temporal link functions (m_0, m_1) which are defensible based on economic theory or intuition – *latent monotone instrumental variable (LIV)* and *treatment invariance (TI)*. LIV asserts that functions m_d

are non-decreasing for any d. That is, means of long-term potential outcomes are non-decreasing in short-term potential outcomes. It is related to the *monotone instrumental variable* assumption of Manski and Pepper (2000a). LIV may be interpreted as maintaining that the latent potential outcome $S(d)$ is itself a monotone instrumental variable. TI posits that the temporal link functions are invariant to the treatment – $m_1 = m_0$. In other words, TI states that the relationship between the short-term potential outcome and the mean long-term potential outcome is unaffected by the treatment.

While LIV may be justifiable based on intuition, TI is implied by a model. For example, TI would hold under the model proposed by García et al. (2020) in the context of early childhood intervention. LIV and TI do not impose any restrictions on the selection mechanism and represent assumptions on treatment response. In contrast, existing work primarily imposes restrictions on the selection mechanism. As mentioned previously, the literature has argued that frequently used assumptions in the context of LTE identification via data combination may also fail in economic settings of interest, or may be incompatible with common selection models. Manski (1997) note that convincing behavioral arguments are generally often lacking for such assumptions.

The third contribution is a novel two-step identification approach that enables the use of the proposed assumptions. In the first step, I find all $(m_0, m_1, \gamma_0, \gamma_1)$ compatible with the data and assumptions. In the second step, I collect all values for the LTE that possible $(m_0, m_1, \gamma_0, \gamma_1)$ produce via (1.1), which yields the identified set for the LTE. The LTE is point identified if this set is a singleton. If it is not a singleton, the LTE is partially identified. Either is permitted and which occurs depends on the imposed assumptions and

the observed data distributions. It should be emphasized that in either case, all produced bounds are the smallest possible under the maintained assumptions, or *sharp*.

In the first step, I find the set of all possible $(m_0, m_1, \gamma_0, \gamma_1)$ under a generic restriction on (m_0, m_1), which embodies modeling assumptions maintained by the researcher. This is done via appropriately defined random sets, extending the arguments in Beresteanu, Molchanov, and Molinari (2012). To operationalize the result, I combine two concepts from random set theory – Artstein's inequalities and the conditional Aumann expectation. This characterizes the restrictions on $(m_0, m_1, \gamma_0, \gamma_1)$ via a collection of moment conditions. The two concepts are commonly used in isolation, or combined when the conditioning variable in the Aumann expectation is observed (Chesher and Rosen (2017), Chesher and Rosen (2020)). A distinguishing feature of the setting here is that the conditioning variable is itself a measurable selection of a random set.

In the second step, I collect all values that possible $(m_0, m_1, \gamma_0, \gamma_1)$ produce via (1.1), which yields the identified set for the LTE. I show that this set can be characterized as an interval bounded by solutions to two *generalized bilinear programs* (Al-Khayyal (1992)). Bilinear programs are computationally more demanding than linear programs that are commonly used to characterize identified sets. To alleviate the computational burden, I show that the bilinear programs can be restated as bilevel (nested) programs and prove that the inner optimization problems have closed-form solutions under LIV, TI, and existing assumptions. I further reduce the number of constraints in the optimization problems via the concept of *core determining classes* (Galichon and Henry (2011)). This characterization leads to tractable plug-in estimators based on results in Shi and Shum (2015) and Russell (2021).

The two-step identification approach has additional appealing features, beyond providing bounds under LIV and TI. Concretely, it can computationally produce sharp bounds under modeling assumptions representable as restrictions on (m_0, m_1). This can be achieved by introducing new modeling assumptions as constraints in the optimization problems. The identification results thus provide a tool for researchers to characterize identified sets under new assumptions tailored to their empirical setting, without requiring proofs of sharpness. For similar results in different settings, see Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019), Russell (2021), Kamat (2024) and references therein. The approach developed here may also be of independent interest in other settings as it facilitates tractable utilization of assumptions restricting latent conditional means.

One additional important advantage of the approach is that it can accommodate imperfect compliance in the experimental data by allowing partial identification of the subvector (γ_0, γ_1). Accommodating imperfect compliance is of great practical relevance. Compliance issues are prevalent in RCTs, and especially in the experiments previously used in this context.[3] Moreover, often considered alternative parameters under non-compliance, such as the intent-to-treat effect (ITT) and local average treatment effect (LATE) of Imbens and Angrist (1994) are unidentified in this setting because the long-term outcomes are never observed in the experimental data.[4] However, despite its practical relevance and dearth of identified alternative target parameters, related literature did not consider experiments with imperfect compliance, to the extent of my knowledge. Since existing

3. Athey, Chetty, and Imbens (2020) and Park and Sasaki (2024a) use the Project STAR and Aizer et al. (2024) the Job Corps RCT. Both had significant reassignment/compliance issues (e.g. see Chen, Flores, and Flores-Lagunes (2018) and Russell (2021)).

4. Even when identified, ITT or LATE may or may not be of interest, depending on the research question. For more details see discussions in Deaton (2009), Heckman and Urzua (2010) and Imbens (2010).

identification strategies may be represented as restrictions on (m_0, m_1), the approach also extends existing identification strategies by allowing them to account for imperfect compliance.

Section 1.2 introduces the setting, summarizes the roles of the experimental data and modeling assumptions, and introduces LIV and TI. Section 1.3 characterizes the identified set and provides a consistent estimator. Section 1.4 discusses the roles of experimental data and modeling assumptions in detail. Section 1.5 concludes. Appendix A.1 contains the extensions of the findings, and Appendix A.2 collects the proofs.

Section 1.2. Setting and Assumptions

I formalize the problem using the standard potential outcomes model. Let $Y(d) \in \mathcal{Y}$ and $S(d) \in \mathcal{S}$ denote the long-term and short-term potential outcomes under some binary treatment $d \in \{0, 1\}$, respectively.[5] Denote the realized treatment by $D \in \{0, 1\}$. The observed outcomes are:

$$\begin{aligned} Y &= DY(1) + (1 - D)Y(0) \\ S &= DS(1) + (1 - D)S(0). \end{aligned} \tag{1.2}$$

Let $X \in \mathcal{X}$ be a vector of observed covariates. Define the conditional *long-term* average treatment effect (CLTE) $\tau(x)$:

$$\tau(x) = E[Y(1) - Y(0)|X = x]. \tag{1.3}$$

5. Supports are invariant to the treatment. This can be relaxed at the expense of more complicated notation.

The parameter of interest can be the CLTE itself or its weighted averages, such as the average *long-term* treatment effect (LTE). I focus on the former for generality noting that it is sufficient for identification of the latter when the weights are identified or given. Throughout the paper, I assume $E[|Y(d)|] < \infty$ for $d \in \{0, 1\}$, which ensures that the parameters are well defined.

Example 1.2.1. *(Head Start Participation)* For illustration, D is an indicator for Head Start participation, $S(d)$ are cognitive test scores in childhood, and $Y(d)$ are outcomes in adulthood, such as earnings, under treatment d.

Subsection 1.2.1. Observed Data

As in Athey, Chetty, and Imbens (2020), I maintain the existence of a population divided into two subpopulations from which the two datasets are randomly drawn: a short-term experimental and a long-term observational dataset. Let $G \in \{O, E\}$ be the indicator for the subpopulation, where $G = O$ generates the observational and $G = E$ the experimental dataset.[6] Let $Z \in \mathcal{Z}$ be an exogenous (i.e. randomly assigned) instrument in the experiment, inducing individuals into treatment. In the experimental dataset, the researcher observes (S, D, X, Z), but not Y. In the observational dataset, (Y, S, D, X) are observed, but Z is absent as there is no instrument in the observational data.

Usually, $Z \in \{0, 1\}$, representing random assignment to the treatment or the control group. The identification analysis can accommodate bounded $\mathcal{Z}$ with multiple or even a continuum of points $\mathcal{Z} = [0, 1]$, as in Heckman and Vytlacil (1999). For expositional

6. This setting has become common. See also García et al. (2020), Athey, Chetty, and Imbens (2020), Ghassami et al. (2022), Hu, Zhou, and Wu (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023), Chen and Ritzwoller (2023), Park and Sasaki (2024a), Aizer et al. (2024) and Imbens et al. (2024).

simplicity, I refer to experiments with $P(D = Z|G = E) < 1$ as having imperfect compliance, as opposed to perfect compliance when $P(D = Z|G = E) = 1$. I thus also refer to Z as treatment assignment regardless of its support, keeping in mind that the $\mathcal{Z}$ may contain points beyond $\{0, 1\}$.

The main purpose of Z is to allow for imperfect experimental compliance, which is practically relevant. This represents a critical distinction between the setting of this paper and related existing work. Researchers often obviate compliance issues by focusing on parameters such as the ITT and LATE. Identification of ITT and LATE requires jointly observing treatment *assignment* Z and the long-term outcomes Y. Since Z is never jointly observed with Y in this setting, both parameters are unidentified.

Example 1.2.1 (continued). The observational dataset is the National Longitudinal Survey of Youth (NLSY), and the experimental dataset is the Head Start Impact Study (HSIS). In the HSIS, $Z = 1$ if the individual is assigned to participation in Head Start and $Z = 0$ if assigned to non-participation. D is the indicator for true participation. Kline and Walters (2016) explain that some individuals may have $D \neq Z$.

I maintain the following assumptions throughout the paper.

Assumption RA. *(Random Assignment)* $Z \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = E$

Assumption EV. *(Experimental External Validity)* $G \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X$.

PAssumption RA holds if Z in the experimental data is randomly assigned. It is a standard assumption in the program evaluation literature. $D \not\perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = g$ is permitted for any $g \in \{O, E\}$. This is expected in the

observational dataset, and in the experimental data under imperfect compliance. When compliance is perfect, Assumption RA implies $D \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = E$. I do not assume that $P(D = 1|G = g) \in (0, 1)$ for any $g \in \{0, 1\}$. Instead, $P(D = 1|G = g) \in [0, 1]$ which may be relevant for $g = O$ when a certain treatment is only available in the experiment. This is the case with some early childhood intervention programs or novel vaccines.

Assumption EV is a standard assumption in the data combination literature, linking the two datasets. It states that the subpopulations generating them do not differ in terms of counterfactual distributions (conditional on X). It holds when participants are randomly recruited into the datasets from the same population (conditional on X).

Under Assumption EV, CLTE is invariant to G, $E[Y(1) - Y(0)|X = x, G] = E[Y(1) - Y(0)|X = x] = \tau(x)$. Henceforth, I keep conditioning on X implicit. The following analysis should be understood as conditional-on-X, and I write the parameter of interest $\tau(x)$ as:

$$\tau = E[Y(1) - Y(0)] \tag{1.4}$$

and I continue referring to it as the LTE, with the understanding that it represents the CLTE.

Notation: I denote laws of random elements using subscripts when the element needs to be specified (e.g. $P_{S(d)}$ is the law of $S(d)$). If the random element is clear from the context, I write laws conditional on an event $\mathcal{E}$, $P(\cdot|\mathcal{E}, G = g)$, as $P_g(\cdot|\mathcal{E})$ for $g \in \{O, E\}$. Whenever $P_E(\cdot|\mathcal{E}) = P_O(\cdot|\mathcal{E})$, I omit the subscript g. This is inherited by their features $E[\cdot|\mathcal{E}, G = g] = E_g[\cdot|\mathcal{E}]$ and $V[\cdot|\mathcal{E}, G = g] = V_g[\cdot|\mathcal{E}]$.

Subsection 1.2.2. Identification Preliminaries

This paper proposes a novel identification approach. To introduce it, recall that for $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$m_d(s) := E[Y(d)|S(d) = s] \tag{1.5}$$

$$\gamma_d := P_{S(d)}. \tag{1.6}$$

I refer to $m_d(s)$ as *temporal link functions*, since they "link" the short-term and long-term potential outcomes in a way that is meaningful for identification of τ. We can write the parameter of interest as:

$$\tau = E[Y(1) - Y(0)] = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{1.7}$$

Denote the pair of temporal link functions $m := (m_0, m_1)$, and the pair of short-term potential outcome distribution functions by $\gamma := (\gamma_0, \gamma_1) = (P_{S(0)}, P_{S(1)})$. Observe that γ consists of the marginal distributions $P_{S(d)}$, and is not the joint-distribution function $P(S(0), S(1))$. Given functions (m, γ), the corresponding value of τ follows by (1.7). Relying on this, the approach identifies (m, γ) as an intermediate step towards identifying τ.

As mentioned in the introduction, this approach two benefits. First, it will computationally produce sharp bounds for a broad class of modeling assumptions, removing the need for proving sharpness for each assumption. Second, it allows one to account for imperfect compliance in the experiment by permitting partial identification of γ, which is of great practical relevance. To formalize the class of modeling assumptions, let $\mathcal{M}$ be the

set of all temporal link functions, i.e. measurable functions mapping $\mathcal{S} \times \mathcal{S} \to \mathcal{Y} \times \mathcal{Y}$.[7] I assume that the researcher knows or can identify the subset $\mathcal{M}^A \subseteq \mathcal{M}$ to which m belongs, which represents a generic modeling assumption. I will provide the identified set for τ for any modeling assumption in this form.

Assumption MA. *(Modeling Assumption)* $m \in \mathcal{M}^A \subseteq \mathcal{M}$ *for a known or identified set* $\mathcal{M}^A$.

Modeling assumptions may be classified as: *selection assumptions*, restricting the relationship between $(Y(1), Y(0), S(1), S(0))$ and D; and *treatment response assumptions*, restricting how $(Y(1), Y(0), S(1), S(0))$ are related to each other.

Assumption MA can accommodate both treatment response and selection assumptions. I will introduce two treatment response assumptions in Section 1.2.3. Remark 1.2.1 explains that Assumption MA nests existing selection assumptions and approaches. Thus, the identification framework will directly extend previously proposed approaches by allowing them to account for imperfect compliance.

Let $\mathcal{H}(\cdot)$ be the identified set for a specified parameter. Finding all (m, γ) consistent with the data and maintained assumptions, including any restriction in the form of Assumption MA, yields $\mathcal{H}(m, \gamma)$. In turn, by the identity (1.7), $\mathcal{H}(\tau)$ follows directly. To this end, define the functional $T : \mathcal{M} \times \mathcal{P}^{\mathcal{S}} \times \mathcal{P}^{\mathcal{S}} \to \bar{\mathbb{R}}$, where $\mathcal{P}^{\mathcal{S}}$ collects distribution functions supported on $\mathcal{S}$:

$$T(m, \gamma) = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{1.8}$$

7. More precisely, $\mathcal{M}$ is the set of Borel-measurable functions $\mu : \mathcal{S} \times \mathcal{S} \to \mathcal{Y} \times \mathcal{Y}$ such that $\mu \circ \varsigma$ is P-integrable for some $\mathcal{F}/\mathcal{B}(\mathcal{S} \times \mathcal{S})$-measurable function $\varsigma : \Omega \to \mathcal{S} \times \mathcal{S}$.

By definition, the identified set $\mathcal{H}(\tau)$ is then equivalent to the set of values T can produce over the identified set $\mathcal{H}(m, \gamma)$:

$$\mathcal{H}(\tau) := \{T(m, \gamma) : (m, \gamma) \in \mathcal{H}(m, \gamma)\}. \tag{1.9}$$

Section 1.3 constructs $\mathcal{H}(m, \gamma)$, and develops a tractable characterization and estimators of $\mathcal{H}(\tau)$.

Subsection 1.2.3. Modeling Assumptions

Section 1.4 provides a detailed discussion on the roles of experimental data and modeling assumptions. It reveals that modeling assumptions remain *central* under data combination, as in observational studies. Experimental data bring *no identifying power* on their own, and serve only to *amplify* the identifying power of the modeling assumption. Hence, if a modeling assumption fails, bounds on τ obtained using just observational data may only be closer to the truth than the bounds obtained using combined data. Therefore, plausible inference hinges on plausible modeling assumptions, despite the use of experimental data.

As mentioned previously, Ghassami et al. (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023) and Imbens et al. (2024) argue that frequently used assumptions may fail in contexts of economic interest; Park and Sasaki (2024a) indicate that they are incompatible with standard models of selection. I thus propose treatment response assumptions that may be defensible based on economic theory or intuition. These assumptions rely on the identification approach for implementability.

Assumption LIV. *(Latent Monotone Instrumental Variable) For any $m \in \mathcal{M}^A$ and $s, s' \in \mathcal{S}$ such that $s < s'$ it holds that $m_d(s) \leq m_d(s')$ for $d \in \{0, 1\}$.*

Assumption LIV has an intuitive interpretation. It posits that the mean of the long-term *potential outcome* $Y(d)$ is non-decreasing conditional on the short-term *potential outcome* $S(d)$. One can symmetrically assume that $E[Y(d)|S(d) = s]$ is non-increasing in s. Results follow directly by defining $\tilde{S}(d) = -S(d)$ and observing that $E[Y(d)|\tilde{S}(d) = s]$ satisfies LIV.

Example 1.2.2. *(LIV and Head Start)* LIV means that people with higher *potential* childhood test scores $S(d)$ under Head Start participation d, on average, also have weakly higher *potential* earnings in adulthood $Y(d)$ under d.

LIV is related to the monotone instrumental variable (MIV) assumption of Manski and Pepper (2000a) (see also Manski and Pepper (2009)). MIV maintains that there exists a variable $V \in \mathcal{V}$ such that $E[Y(d)|V = v]$ is non-decreasing in $v \in \mathcal{V}$, which is observed for *all* individuals. The critical distinction is that the conditioning variable in Assumption LIV is a latent counterfactual. This introduces further complexity, which will be addressed by the identification approach.

Assumption TI. *(Treatment Invariance - TI) For all* $m \in \mathcal{M}^A$ *and* $s \in \mathcal{S}$, $m_1(s) = m_0(s)$.

The assumption intuitively states that the relationship between the *potential* outcomes $S(d)$ and mean long-term *potential* outcomes $Y(d)$ does not vary with the underlying treatment d.

Example 1.2.3. *(TI and Head Start)* TI follows from previously used models in the context of early childhood intervention. Consider the following separable model of

potential earnings:

$$Y(d) = \phi_d(S(d)) + \varepsilon_d = \phi(S(d)) + \varepsilon_d, \;\; \varepsilon_1 \sim \varepsilon_0, \;\; \varepsilon_{d'} \perp\!\!\!\perp S(d), \forall d, d' \in \{0,1\}. \tag{1.10}$$

$S(d)$ is a vector of short-term potential outcomes including test scores and measures of non-cognitive skills. $S(d)$ represents inputs in the production function ϕ_d for $Y(d)$. The production function ϕ_d and the distributions of unobservables ε_d do not depend on Head Start participation d. Therefore, $E[Y(d)|S(d) = s] = \phi(s) + E[\varepsilon]$ which is invariant to d, so TI is implied by the model.

Researchers may utilize TI whenever they find the model from Example 1.2.3 to be plausible. For example, based on mediation results in Heckman, Pinto, and Savelyev (2013) and extensive falsification testing, García et al. (2020) argue the plausibility of a similar model when the treatment is an early childhood intervention. They then identify τ by combining observational and experimental data in the special case where $P_O(D = 0) = 1$ and compliance is perfect, i.e. when there is no selection in either dataset. This paper extends their findings by demonstrating that one may use implications of the same model to provide informative inference on τ when there is selection in either dataset.

In the special case of perfect compliance, TI is implied by the statistical surrogacy assumption of Prentice (1989) – $Y \perp\!\!\!\perp S|D, G = E$. Appendix A.1.1.4 explains the differences. However, researchers may still wish to assign the informal interpretation of the surrogacy assumption to Assumption TI. Intuitively, one may choose to say that the treatment affects the mean long-term outcome only through the short-term outcomes.

To connect the findings of this paper with previous results, I will refer to a widely used selection assumption introduced by Athey, Chetty, and Imbens (2020).

Assumption LUC. *(Latent Unconfoundedness) For all* $d \in \{0,1\} : Y(d) \perp\!\!\!\perp D|S(d), G = O.$

According to Chen and Ritzwoller (2023): "Informally, LUC states that all unobserved confounding in the observational sample is mediated through the short-term outcomes". Park and Sasaki (2024a) describe it as a "statistical assumption" and indicate that it is difficult to interpret economically outside of restricted non-parametric selection models.

Previous work notes that Assumption LUC may be untenable in economic contexts such as early childhood interventions and job-training programs. In the former case, parental interference and the child's inherent ability may be confounding factors for $(D, S(d), Y(d))$, invalidating the assumption. In the latter, the confounding factors may be worker's innate motivation and resourcefulness. For more details see Ghassami et al. (2022) and Imbens et al. (2024). For examples of its use, see Hu, Zhou, and Wu (2022), Park and Sasaki (2024b), Aizer et al. (2024).

Remark 1.2.1. Existing approaches are subsumed under Assumption MA. For example, relevant restrictions for identification of τ under Assumption LUC can be restated as $\mathcal{M}^{LUC} = \{m \in \mathcal{M} : m_d(s) = E_O[Y|S = s, D = d], \forall s \in \mathcal{S}\}$. One can do the same for the outcome bridge function approach of Imbens et al. (2024, Theorem 1). Let S_t for $t \in \{1,2,3\}$ be subvectors of S. Then under the corresponding assumptions: $\mathcal{M}^{Bridge} = \{m \in \mathcal{M} : m_d(s_3, s_2) = h(s_3, s_2, d),\ h \text{ solves } E_O[Y|S_2, S_1, D] = E_O[h(S_3, S_2, D)|S_2, S_1, D]\}$.

Section 1.3. Main Results

Section 1.3.1 summarizes the main identification results and the underlying intuition behind the two-step identification approach; technical discussions follow. Section 1.3.2 characterizes $\mathcal{H}(m, \gamma)$. Section 1.3.3 provides a tractable implementation of $\mathcal{H}(\tau)$ based on this characterization. Section 1.3.4 proposes a consistent estimator for $\mathcal{H}(\tau)$.

Subsection 1.3.1. Identification Intuition

The identification approach aims to operationalize the proposed modeling assumptions. However, it also presents a novel challenge; it necessitates finding $\mathcal{H}(m, \gamma)$ as an intermediate step. Both $m_d(s) = E[Y(d)|S(d) = s]$ and $\gamma_d = P_{S(d)}$ are features of *latent* random variables $Y(d)$ and $S(d)$, and thus (m, γ) are not directly revealed by the data. The identification results exploit this apparent complexity to construct $\mathcal{H}(m, \gamma)$. By characterizing the feasible potential outcomes, the corresponding (m, γ) follow by definition.

There will exist a set of random variables $(S(0), S(1), Y(0), Y(1))$ that are consistent with the data and maintained assumptions because the potential outcomes are latent. Concretely, let $\mathcal{Q}$ be the set of all $(S(0), S(1), Y(0), Y(1))$ that are consistent with the data, and Assumptions RA and EV. The researcher can determine $\mathcal{Q}$. To find $\mathcal{H}(m, \gamma)$, one then only needs to collect all corresponding (m, γ) such that they additionally satisfy the modeling assumption $m \in \mathcal{M}^A$. By definition:

$$(1.11)\quad \mathcal{H}(m,\gamma) = \left\{ (m,\gamma) : \begin{array}{c} \overbrace{m \in \mathcal{M}^A,}^{Modeling\ assumption} \quad \overbrace{\exists (S(0),S(1),Y(0),Y(1)) \in \mathcal{Q}}^{Data\ +\ Assumptions\ RA/EV} \\ \underbrace{\forall d \in \{0,1\} : \gamma_d \overset{d}{=} S(d),\ m_d(S(d)) = E[Y(d)|S(d)] \text{ a.s.}}_{(m,\gamma)\ correspond\ to\ S(d)\ and\ Y(d)} \end{array} \right\}.$$

Then, again definitionally, $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$. These expressions demonstrate that one may use the information on the potential outcomes to relate (m,γ) and τ to observed data and assumptions. However, while intuitive, the definitions are intractable.

The first main identification result utilizes (1.11) to provide a general equivalent characterization of $\mathcal{H}(m,\gamma)$ in terms of moment restrictions. Let $Y_L := \inf \mathcal{Y}$ and $Y_U := \sup \mathcal{Y}$. When Y_L, Y_U, $\mathcal{S}$ and $\mathcal{Z}$ are finite with $|\mathcal{S}| = k$, the general characterization of $\mathcal{H}(m,\gamma)$ simplifies to:[8]

(1.12)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z \in \mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s, D=d)\right), \\ (m_d(s) - Y_L)\,\gamma_d(s) \geq (E_O[Y|S=s, D=d] - Y_L)\,P_O(S=s, D=d)\ , \\ (Y_U - m_d(s))\,\gamma_d(s) \geq (Y_U - E_O[Y|S=s, D=d])\,P_O(S=s, D=d) \end{array} \right\}$$

8. Conceptually, the identification results do not require $\mathcal{S}$ or $\mathcal{Z}$ to be finite. The former allows one to computationally characterize the set in full. Both will be assumed to construct the consistent estimator.

where $\Delta(k)$ denotes the k-dimensional simplex. In turn, this yields the second main identification result – characterization of $\mathcal{H}(\tau)$ using optimization problems:

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right]. \tag{1.13}$$

where the moment conditions defininining $\mathcal{H}(m,\gamma)$ take the role of the constraint set.

Therefore, the researcher may determine $\mathcal{H}(\tau)$ by solving two constrained optimization problems. Beyond producing sharp bounds under the previously introduced assumptions, (1.13) also provides a tool for researchers to computationally obtain sharp bounds on τ under tailor-made modeling assumptions. To do so, it is sufficient to solve the optimization problems with appropriately defined constraints based on $\mathcal{M}^A$. Computational characterizations of identified sets have been exploited previously to obviate the need to prove sharpness for each set of assumptions; for recent examples in different settings, see Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019), Russell (2021), Kamat (2024) and references therein. Commonly, the corresponding optimization problems are linear or have linear equivalents.

Both the objective T and the constraints imposed by $\mathcal{H}(m,\gamma)$ are bilinear in (m,γ) under previously discussed assumptions. Hence, the optimization problems in (1.13) are *generalized bilinear programs* (see Al-Khayyal (1992)). Such programs are nonlinear and computationally demanding, in general. Section 1.3.3 thus proposes further simplifications that exploit the structure of the identified sets. I use the concept of *core-determining* classes to remove redundant restrictions on γ. Additionally, I restate the problems as

bilevel programs:

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \min_{\tilde{m}\in\mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \max_{\tilde{m}\in\mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma})\right]. \tag{1.14}$$

and show that inner problems have closed-form solutions for the proposed and existing assumptions. Consistent plug-in estimators for $\mathcal{H}(\tau)$ can be constructed based on these representations.

Section 1.3.2 presents the general characterization of $\mathcal{H}(m,\gamma)$ that leads to (1.12), and discusses its derivation. Section 1.3.3 then provides representations of $\mathcal{H}(\tau)$ found in (1.13) and (1.14). Section 1.3.4 develops the consistent estimator.

Subsection 1.3.2. Identification of (m,γ)

This section represents $\mathcal{H}(m,\gamma)$ in terms of moment restrictions. This set is conducive to tractable implementation of $\mathcal{H}(\tau)$. To present the result, I introduce the necessary basic definitions from random set theory specialized to finite-dimensional Euclidean spaces. Appendix A.2.2.1 contains a more complete, but brief, overview of the results used in the proofs. I henceforth maintain that all random elements are defined on a common non-atomic probability space $(\Omega,\mathcal{F},P)$.[9]

Notation: A, B and K represent sets. $\mathcal{K}(A)$, $\mathcal{C}(A)$, and $\mathcal{B}(A)$ are the families of all compact, closed, and Borel subsets of the set A, respectively. $co(A)$ is the closed convex hull of the set A. I write random sets using boldface letters (e.g. $\mathbf{Y}$), and $\mathbf{Y}\times\mathbf{X}$ as $(\mathbf{Y},\mathbf{X})$.

9. That is, for any $A\in\mathcal{F}$ with positive measure there exists a measurable $B\subset A$ such that $0<P(B)<P(A)$.

Definition 1.3.1. A measurable map $\mathbf{R} : \Omega \to \mathcal{C}(\mathbb{R}^d)$ is called a *random (closed) set*.[10]

Definition 1.3.2. A random variable $R : \Omega \to \mathbb{R}^d$ such that $R \in \mathbf{R}$ a.s. is called a *(measurable) selection* of $\mathbf{R}$. $Sel(\mathbf{R})$ and $Sel^1(\mathbf{R})$ are the sets of all selections, and all integrable selections of $\mathbf{R}$, respectively.

Definition 1.3.3. If the random variable $||\mathbf{R}|| = sup\{||R|| : R \in Sel(\mathbf{R})\}$ is integrable $E[||\mathbf{R}||] < \infty$, then the random set $\mathbf{R}$ is said to be *integrably bounded*.

Define the following closed random sets for $d \in \{0, 1\}$:

(1.15)

$$\mathbf{Y}(d) := \begin{cases} \{Y\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}, \quad \mathbf{S}(d) := \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases}.$$

$\mathbf{Y}(d)$ and $\mathbf{S}(d)$ serve to summarize information on the counterfactuals $(S(0), S(1), Y(0), Y(1))$, and thus (m, γ), contained in the data and assumptions. Their properties lead to the following result.

10. $\mathbf{R}$ is measurable if for every compact set $K \in \mathcal{K}(\mathbb{R}^d)$: $\{\omega \in \Omega : \mathbf{R}(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$. The codomain $\mathcal{C}(\mathbb{R}^d)$ is equipped by the $\sigma-$algebra generated by the families of sets $\{B \in \mathcal{C}(\mathbb{R}^d) : B \cap K \neq \emptyset\}$ over $K \in \mathcal{K}(\mathbb{R}^d)$.

Theorem 1.3.1. *Let Assumptions RA, EV and MA hold. If* $\mathbf{Y}(d)$ *is integrably bounded, the identified set for* (m, γ) *is:*

(1.16)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}\text{: } um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s))\ \gamma_d\text{–}a.e. \end{array} \right\}$$

where $h_{co(\mathcal{Y})}(u) = \sup_{y \in co(\mathcal{Y})} uy$, $\mu_d(s) = E_O[Y|S = s, D = d]$, *and* $\pi_{\gamma_d} = dP_O(S, D = d)/d\gamma_d$. *If a collection of sets* $\mathfrak{C}$ *is a core determining class for the containment functional of* $\mathbf{S}(d)$, *then the condition* $\forall B \in \mathcal{C}(\mathcal{S})$ *can be replaced with* $\forall B \in \mathfrak{C}$.

Theorem 1.3.1 equivalently characterizes $\mathcal{H}(m, \gamma)$ for any modeling restriction in the form $m \in \mathcal{M}^A$ via moment restrictions that are identified by the data. This includes, but is not limited to, assumptions and approaches in Section 1.2.3 and Remark 1.2.1. It also offers computational simplifications via the concept of *core-determining classes* – sub-families of $\mathcal{C}(\mathcal{S})$ which are sufficient to completely characterize γ_d (Galichon and Henry (2011)). Informally, a core determining class allows one to remove redundant restrictions on each γ_d, without any loss of information, which will be beneficial for tractability. Theorem 1.3.1 will be used to provide a tractable implementation of $\mathcal{H}(\tau)$ in Section 1.3.3.

The technical contribution of the theorem lies in jointly identifying conditional means and corresponding distributions of latent random variables via moment restrictions. This necessitates novel arguments that may be of independent interest. Namely, the proof combines *Artstein's theorem* (Artstein (1983, Theorem 2.1)) and the *conditional Aumann expectation* when the relevant conditioning σ-algebra is generated by a selection of a

random set, i.e. a latent variable. Section 1.3.2.1 thus sketches how Theorem 1.3.1 is obtained. The main results in Section 1.3.3 do not require these discussions.

1.3.2.1. Mechanics behind Theorem 1. Recalling the intuition, to find all feasible (m, γ), one may first summarize the information about the counterfactuals $Y(d)$ and $S(d)$. By definition, random sets $\mathbf{Y}(d)$ and $\mathbf{S}(d)$ express all information on $Y(d)$ and $S(d)$ contained in *the data*, respectively. As Beresteanu, Molchanov, and Molinari (2012) explain, *all* information in the data about $S(d)$ and $Y(d)$ can be expressed as $(S(0), S(1), Y(0), Y(1)) \in Sel((\mathbf{S}(0), \mathbf{S}(1), \mathbf{Y}(0), \mathbf{Y}(1)))$. Intuitively, we can think about random sets $\mathbf{S}(d)$ and $\mathbf{Y}(d)$ as bundles of random variables. The data only reveal that $S(d)$ and $Y(d)$ are elements of these bundles, but not which ones exactly.

Assumptions RA, EV and $E[|Y(d)|] < \infty$ further restrict which elements the potential outcomes may be. The counterfactuals are consistent with the three assumptions if and only if $(S(0), S(1), Y(0), Y(1)) \in \mathcal{I}$, where $\mathcal{I}$ is the set of random elements (E_1, E_2, E_3, E_4) such that $(E_1, E_2, E_3, E_4) \perp\!\!\!\perp Z | G = E$, $(E_1, E_2, E_3, E_4) \perp\!\!\!\perp G$ and $E[|E_3|], E[|E_4|] < \infty$. Therefore, all information about the counterfactuals in the *data* and the *three assumptions* can be expressed by:

$$(S(0), S(1), Y(0), Y(1)) \in Sel((\mathbf{S}(0), \mathbf{S}(1), \mathbf{Y}(0), \mathbf{Y}(1))) \cap \mathcal{I} := \mathcal{Q}.$$

The identified set for $\mathcal{H}(m, \gamma)$ follows by definition as all corresponding (m, γ) that additionally satisfy the modeling assumption:

$$\mathcal{H}(m, \gamma) := \left\{ \begin{array}{l} (m, \gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \exists (S(0), S(1), Y(0), Y(1)) \in \mathcal{Q}, \\ \forall d \in \{0, 1\}, \ \ \gamma_d \overset{d}{=} S(d), \ \ m_d(S(d)) = E[Y(d)|S(d)] \text{ a.s.} \end{array} \right\}. \tag{1.17}$$

The definition imposes redundant restrictions on (m, γ), which preclude the use of appropriate tools needed to obtain moment conditions. The following lemma disposes of such restrictions and is important for explaining how Theorem 1.3.1 is obtained.

Lemma 1.3.1. *Let Assumptions RA, EV, and MA hold. The identified set for* (m, γ) *is:*

(1.18)
$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \;\; \exists S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}, \\ \exists Y(d) \in Sel^1(\mathbf{Y}(d)), \;\; \gamma_d \overset{d}{=} S(d), \;\; m_d(S(d)) = E_O[Y(d)|S(d)] \;\; a.s. \end{array} \right\}.$$

where $\bar{I}$ *is the set of random elements* $E_1 \in \mathcal{S}$ *such that* $E_1 \perp\!\!\!\perp G$ *and* $E_1 \perp\!\!\!\perp Z|G = E$.

The lemma indicates that, for identification of (m, γ) it is sufficient to: 1) consider restrictions on m_d imposed by $Y(d)$ and $S(d)$ only conditional on $G = O$, reflected by $m_d(S(d)) = E_O[Y(d)|S(d)]$; 2) only impose marginal independence conditions $S(d) \perp\!\!\!\perp G$ and $S(d) \perp\!\!\!\perp Z|G = E$, instead of the full joint independence as in Assumption RA and EV.

With Lemma 1.3.1, the characterization in Theorem 1.3.1 can be constructed. First, for any $S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}$, collect all conditional expectations $E_O[Y(d)|S(d)]$ over $Y(d) \in Sel^1(\mathbf{Y}(d))$ into a set. This yields the *random* set $\{E_O[Y(d)|S(d)] : \;\; Y(d) \in Sel^1(\mathbf{Y}(d))\}$. When $\mathbf{Y}(d)$ is integrably bounded, Li and Ogura (1998, Theorem 1) show that this random set is equivalent to the *conditional Aumann expectation* denoted by $\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$.[11] Then,

11. The conditional Aumann expectation is defined with respect to any conditioning sub-σ-algebra $\mathcal{F}_0 \subsetneq \mathcal{F}$. Here, this is the σ-algebra generated by events $\{\{S(d) \in B\} \cap \{G = O\} : B \in \mathcal{B}(\mathcal{S})\}$, which I keep implicit for ease of notation. See Appendix A.2.2.1 for a formal definition.

it is easy to see that for a given $S(d)$:

(1.19)

$$\exists Y(d) \in Sel^1(\mathbf{Y}(d)): \quad m_d(S(d)) = E_O[Y(d)|S(d)] \text{ a.s.} \Leftrightarrow m_d(S(d)) \in \mathbb{E}_O[\mathbf{Y}(d)|S(d)] \text{ a.s.}$$

$\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$ is convex on non-atomic probability spaces. Therefore, it can be represented using its support function, which equivalently characterizes the condition $m_d(S(d)) \in \mathbb{E}_O[\mathbf{Y}(d)|S(d)]$ as a set of moment restrictions. For a given $S(d)$ with a distribution γ_d, (1.19) holds if and only if:

(1.20)

$$\forall u \in \{-1, 1\}: \; u m_d(s) \leq u E_O[Y|S = s, D = d]\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)) \; \gamma_d\text{−a.e.}$$

recalling that $h_{co(\mathcal{Y})}(u) = \sup_{y \in co(\mathcal{Y})} uy$.

All restrictions on m_d depend on the selection $S(d)$ *only* up to its distribution γ_d, which will be essential in the next step. Observe that all elements on the right-hand side of (1.20) are either known or identified from the data, given γ_d. Notably, π_{γ_d} is identified given γ_d, which is evident when γ_d is discretely supported. Then, $\pi_{\gamma_d}(s) = \frac{P_O(S=s, D=d)}{\gamma_d(s)}$ for $s \in \mathcal{S}$ such that $\gamma_d(s) > 0$. This intuition extends to the case when $\mathcal{S}$ is not discrete. Note that $P_O(D = d|S'(d) = s) = \pi_{\gamma_d}(s)$, so π_{γ_d} can be interpreted as the *latent* propensity score, conditioning on a latent variable $S'(d) \overset{d}{=} \gamma_d$ (for other uses see Masten and Poirier (2023)).

The result in (1.20) removes the need to search over $Y(d) \in Sel^1(\mathbf{Y}(d))$, but the need to search over $S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}$ remains. However, by Artstein's theorem:

$$\begin{aligned} &\exists S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I} \text{ such that } \gamma_d \overset{d}{=} S(d) \\ &\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d) \right). \end{aligned} \tag{1.21}$$

This characterizes the set of distributions γ_d such that they are "rationalized" by a selection $S(d)$ satisfying conditions of Lemma 1.3.1; (1.20) characterizes the set of link functions m_d such that they "rationalized" by a selection $Y(d)$ satisfying conditions of Lemma 1.3.1, *given* a distribution γ_d. By putting the two results together, Theorem 1.3.1 follows. Hence, (m, γ) can be characterized using only moment conditions.

The theorem provides an additional important simplification that reduces the number of conditions for γ_d imposed by (1.21). This is done via a *core-determining class* – a subfamily $\mathfrak{C} \subseteq \mathcal{C}(\mathcal{S})$ sufficient to summarize all restrictions on γ_d in (1.21). More precisely, $\mathfrak{C}$ is a core determining class when:

$$\begin{aligned} &\forall B \in \mathfrak{C} : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d) \right) \\ &\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d) \right). \end{aligned} \tag{1.22}$$

If a core-determining class exists, using it can substantially reduce the number of constraints on γ_d. This is instrumental in reducing computational burden when determining $\mathcal{H}(\tau)$ in the next step.

Subsection 1.3.3. Tractable Characterization of $\mathcal{H}(\tau)$

Recall that $\mathcal{H}(m,\gamma)$ yields the identified set $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$, where $T(m,\gamma) = \int_{\mathcal{S}} m_1(s)d\gamma_1(s) - \int_{\mathcal{S}} m_0(s)d\gamma_0(s)$. This section operationalizes $\mathcal{H}(\tau)$ using Theorem 1.3.1.

To verify if a candidate (m,γ) is in the identified set, one must establish that each γ_d satisfies an inequality condition for each closed subset $B \in \mathcal{C}(\mathcal{S})$. If $\mathcal{S}$ is infinite, then so is $\mathcal{C}(\mathcal{S})$. Some of these restrictions may be redundant. However, even the smallest set of non-redundant restrictions on γ_d, i.e. the smallest core-determining class $\mathfrak{C}$, will contain infinitely many sets (Ponomarev (2024, Theorem 1)). It is thus generally computationally infeasible to fully characterize $\mathcal{H}(m,\gamma)$ and $\mathcal{H}(\tau)$ when the relevant outcome space is infinite. This is a well-known issue with identified sets that follow from Artstein's theorem. A common way of addressing it is to discretize the relevant variables or focus on settings where they are finitely supported (Galichon and Henry (2011), Russell (2021), Ponomarev (2024)). Here, I do the same. A computationally tractable characterization of $\mathcal{H}(\tau)$ that minimizes the loss of information with infinitely supported S is an interesting avenue for future research.

Henceforth, I maintain that $S(d) \in \mathcal{S} = \{1, 2, \ldots, k\}$, either by definition or following discretization performed by the researcher. Subtleties related to the interpretation of results under discretization are discussed in Appendix A.1.2. I do not require the long-term outcome support $\mathcal{Y}$ to be a finite or discrete set, but I maintain $\mathcal{Y}$ is a bounded set for some known finite $Y_L := \inf \mathcal{Y}$ and $Y_U := \sup \mathcal{Y}$. This is commonly required for informative inference under nonparametric treatment response assumptions. The support restriction may be natural for various $Y(d)$ such as binary indicators, or discrete and

continuous variables that are logically bounded. For some $Y(d)$, it may be restrictive. When, $|\mathcal{S}| < \infty$, one can represent γ_d as an element of a $k-$dimensional simplex $\Delta(k)$, and $\gamma \in \Delta(k) \times \Delta(k)$. Similarly, $m \in \mathcal{M} = \mathcal{Y}^k \times \mathcal{Y}^k$, and the modeling assumption can be represented as $\mathcal{M}^A \subseteq \mathcal{Y}^k \times \mathcal{Y}^k$. Let $\gamma_d(s)$ and $m_d(s)$ denote the $s-$th element of the corresponding vectors. This leads to the following characterization result.

Theorem 1.3.2. *Let Assumptions RA, EV, and MA hold. Suppose $\mathcal{S}$ is a finite set, and that $\mathcal{M}^A$ is closed and convex. Then:*

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right] \tag{1.23}$$

where:

(1.24)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_{z \in \mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s, D=d)\right), \\ (m_d(s) - Y_L)\,\gamma_d(s) \geq (E_O[Y|S=s, D=d] - Y_L)\,P_O(S=s, D=d)\ , \\ (Y_U - m_d(s))\,\gamma_d(s) \geq (Y_U - E_O[Y|S=s, D=d])\,P_O(S=s, D=d) \end{array} \right\}$$

By the theorem, $\mathcal{H}(\tau)$ can be equivalently represented as an interval bounded by solutions to two optimization problems where $\mathcal{H}(m,\gamma)$ represents the constraint set. The characterization follows under easily verifiable high-level conditions on $\mathcal{M}^A$. Remark 1.3.1 explains that these conditions are satisfied by the proposed and existing assumptions.

Using optimization problems to characterize identified sets has become common in partial identification analyses. Such representations usually follow directly from the convexity of the constraint set and linearity of the objective function. Theorem 1.3.2

requires a different argument since T is a difference of two Riemann-Stieltjes integrals, thus bilinear and hence separately continuous in m and γ. The proof shows that T is jointly continuous, and that Theorem 1.3.1 yields $\mathcal{H}(m, \gamma)$ which is compact and convex under the assumptions of the theorem. Then $\mathcal{H}(\tau)$ is a continuous image of a compact and convex set, hence a compact connected set, i.e., a closed interval.

Remark 1.3.1. The assumptions considered here are representable via linear equality and inequality restrictions on m. Therefore, the resulting $\mathcal{M}^A$ are polytopes when $|\mathcal{S}| < \infty$, and thus closed and convex. Assumption LIV states that vectors $m_d \in \mathcal{Y}^k$ have non-decreasing components $m_d(s) \leq m_d(s+1)$; Assumption TI maintains that $m_1(s) = m_0(s)$. Moreover, whenever m is identified by the data, such as under LUC, $\mathcal{M}^A$ is a singleton and hence closed and convex.

The constraint set in Theorem 1.3.2 utilizes the fact that the family of sets $\{\{s\} : s \in \mathcal{S}\}$ represents a *core-determining class* (CDC henceforth) for $\mathbf{S}(d)$ when $|\mathcal{S}| < \infty$. The CDC removes redundant constraints on $\mathcal{H}(m, \gamma)$ in the optimization problems without any loss of information.

The reduction in the number of constraints depends on the size of $|\mathcal{S}|$, but it is sizeable even for relatively few support points. Without the CDC, there would be 2^{k-1} inequality conditions for each for each γ_d, one for each nontrivial proper subset in $\mathcal{C}(\mathcal{S})$. With the CDC, the number of constraints for each γ_d is reduced to $k-1$. Table 1.1 depicts the magnitude of this reduction, showing the total number of constraints on γ with and without the CDC in a single optimization problem with respect to $|\mathcal{S}|$.[12] If $S(d)$ represents

12. The number of constraints on m imposed by the data given γ is $4k$; the total number depends on the modeling assumptions.

percentiles, then not using the CDC results in a prohibitively complex constraint set. The number of constraints may potentially be further reduced by adapting methods in Luo and Wang (2018) and Ponomarev (2024), as $\{\{s\} : s \in \mathcal{S}\}$ is not necessarily the smallest CDC.

	$\lvert\mathcal{S}\rvert$				
Constraint # for γ	2	5	10	20	100
Without CDC	4	32	1024	1048576	$> 10^{30}$
With CDC $\{\{s\} : s \in \mathcal{S}\}$	2	8	18	38	198

Table 1.1. Number of Constraints on γ in $\mathcal{H}(m, \gamma)$.

1.3.3.1. Reducing Computational Complexity of Bilinear Programming. It is immediate that constraints imposed by $\mathcal{H}(m, \gamma)$ are bilinear in (m, γ). For example, consider the constraint:

$$(m_d(s) - Y_L)\gamma_d(s) \geq (E_O[Y|S = s, D = d] - Y_L)\, P_O(S = s, D = d).$$

It is separately linear in $m_d(s)$ and $\gamma_d(s)$, and thus bilinear. Constraints imposed on γ_d by Artstein's theorem are linear. Remark 1.3.1 explains that $\mathcal{M}^A$ imposes only linear restrictions on m under Assumptions LIV, TI and LUC. Coupling this with the fact that T is bilinear, the optimization problems in Theorem 1.3.2 represent *generalized bilinear programs* (see Al-Khayyal (1992)) under the considered assumptions. Such programs are non-convex and computationally demanding in general. I thus propose further simplifications that exploit the structure of the identified set $\mathcal{H}(m, \gamma)$ and the objective T. I restate the problems as bilevel programs and show that inner problems may have closed-form solutions.[13]

First, decompose $\mathcal{H}(m, \gamma)$ into its projection $\mathcal{H}(\gamma) := \{\gamma' : \exists m' \text{ s.t. } (m', \gamma') \in \mathcal{H}(m, \gamma)\}$ and corresponding fibers $\mathcal{H}(m|\gamma') := \{m' : (m', \gamma') \in \mathcal{H}(m, \gamma)\}$ at each $\gamma' \in \mathcal{H}(\gamma)$. The

13. Bilinear programs have previously been used for identification by Dutz et al. (2021) and Shea (2022). An example of using bilevel optimization problems for identification can be found in Moon (2024).

fibers form a correspondence $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$. The identified set can then be written as:

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} \min_{\tilde{m} \in \mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}(\gamma)} \max_{\tilde{m} \in \mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}) \right]. \tag{1.25}$$

The inner optimization problems may have known closed-form solutions given by some selectors L_γ and U_γ of the correspondence $\mathcal{H}(m|\cdot)$. This is formalized by the following definition.

Definition 1.3.4. (Minimal and Maximal Selectors) Let $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$ be a correspondence defined by fibers of $\mathcal{H}(m, \gamma)$ over its projection $\mathcal{H}(\gamma)$. L_γ is a *minimal selector with respect to* T if for any $\gamma \in \mathcal{H}(\gamma)$: $T(L_\gamma, \gamma) \leq T(m, \gamma)$ for all $m \in \mathcal{H}(m|\gamma)$. U_γ is a *maximal selector with respect to* T if for any $\gamma \in \mathcal{H}(\gamma)$: $T(U_\gamma, \gamma) \geq T(m, \gamma)$ for all $m \in \mathcal{H}(m|\gamma)$.

Corollary 1.3.1. *Let conditions of Theorem 1.3.2 hold. If $\mathcal{H}(m|\cdot)$ has minimal and maximal selectors with respect to T, then:*

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(L_{\tilde{\gamma}}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(U_{\tilde{\gamma}}, \tilde{\gamma}) \right].$$

Lemma A.2.10 shows that Assumptions LIV and TI produce known minimal and maximal selectors. Moreover, whenever m is identified, such as under LUC, minimal and maximal selectors exist and coincide by definition.

Remark 1.3.2. Optimization problems $\max / \min_{(\tilde{m}, \tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma})$ become linear and simple to solve in special cases. This happens whenever either $\mathcal{H}(m, \gamma) = \{m\} \times \mathcal{H}(\gamma)$; or $\mathcal{H}(m, \gamma) = \mathcal{H}(m) \times \{\gamma\}$ and $\mathcal{M}^A$ can be expressed using linear constraints. Assumptions

that point identify m independently of γ, such as Assumption LUC, yield $\mathcal{H}(m,\gamma) = \{m\} \times \mathcal{H}(\gamma)$. The latter case would occur for Assumptions LIV and TI under perfect compliance. Note that then Lemma A.2.10 would yield a closed-form expression for $\mathcal{H}(\tau) = [T(L_\gamma,\gamma), T(U_\gamma,\gamma)]$ where γ takes a single value.

Subsection 1.3.4. Estimation

The analysis thus far has focused on identification. I now propose a consistent estimation procedure via sample analogues of $\mathcal{H}(m,\gamma)$, and $\mathcal{H}(\tau)$. For this, suppose that the researcher observes experimental and observational samples $\{(S_j, D_j, Z_j)\}_{j=1}^{n_E}$ and $\{(Y_i, S_i, D_i)\}_{i=1}^{n_O}$, respectively. To establish consistency, I pose the following assumption.

Assumption E. *(Estimation)*

i) $\{(S_j, D_j, Z_j)\}_{j=1}^{n_E}$ *and* $\{(Y_i, S_i, D_i)\}_{i=1}^{n_O}$ *are i.i.d. samples;*

ii) $|\mathcal{S}|, |\mathcal{Z}| < \infty$*;*

iii) $\mathcal{M}^A$ *is defined through finitely many linear equality and inequality restrictions. The Jacobian of linear equality constraints, if imposed, has full row rank.*

iv) $int(\mathcal{H}(m,\gamma)) \neq \emptyset$*.*

Assumption E *i)* is standard under random sampling, *ii)* maintains that short-term potential outcomes and the instrument Z are finitely supported. Assumption E *iii)* defines the class of modeling assumptions that are compatible with the estimation procedure. If necessary, it may be further weakened to allow for continuously differentiable restrictions on m, but it is sufficiently general to encompass all previously stated modeling assumptions. Assumption LIV can be represented only using linear inequality constraints $m_d(s) \leq$

$m_d(s+1)$ for $s \in \{1, \ldots, k-1\}$ and $d \in \{0,1\}$ so *iii)* holds directly. Assumption TI involves only linear constraints $m_1(s) = m_0(s)$ for $s \in \mathcal{S}$. Since each constraint restricts a different s, it is immediate that the constraints will be linearly independent, and the Jacobian matrix will have full row rank. Similar arguments apply to assumptions that use equality restrictions involving population parameters, such as Assumption LUC.

Condition E *iv)* is sufficient for a condition of Shi and Shum (2015, Theorem 2.1) which provides a consistent estimator without requiring a tuning parameter. It holds when components in (m, γ) are partially identified. It may thus not be restrictive whenever treatment response assumptions are maintained and there is imperfect experimental compliance. It is restrictive when the assumption point identifies the link function, as is the case with Assumption LUC. It may be relaxed at the expense of introducing tuning parameters, as explained by Shi and Shum (2015, Section 2). The estimator that follows requires no tuning parameters.

Let $n = \min\{n_O, n_E\}$, and denote by $P_{E,n}(S \in A, D = d, Z = z) = \frac{1}{n_E} \sum_{j=1}^{n_E} \mathbb{1}\{S_j \in A, D_j = d, Z_j = z\}$ and $P_{O,n}(S \in A, D = d) = \frac{1}{n_O} \sum_{i=1}^{n_O} \mathbb{1}\{S_i \in A, D_i = d\}$ be standard empirical measures. Denote by $E_{E,n}$ and $E_{O,n}$ the corresponding empirical expectations. Note that their population counterparts, along with $\mathcal{M}^A$, fully characterize $\mathcal{H}(m, \gamma)$ and

thus $\mathcal{H}(\tau)$. Define the empirical analog of $\mathcal{H}(m,\gamma)$:

(1.26)

$$\mathcal{H}_n(m,\gamma) := \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}_n^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z\in\mathcal{Z}} P_{E,n}(S=s,D=d|Z=z), P_{O,n}(S=s,D=d)\right), \\ (m_d(s)-Y_L)\gamma_d(s) \geq \left(E_{O,n}[Y|S=s,D=d]-Y_L\right) P_{O,n}(S=s,D=d), \\ \left(Y_U - m_d(s)\right)\gamma_d(s) \geq \left(Y_U - E_{O,n}[Y|S=s,D=d]\right) P_{O,n}(S=s,D=d) \end{array} \right\}$$

where $\mathcal{M}_n^A$ highlights that restrictions imposed by the modeling assumptions may depend on estimated population parameters. For example, this would happen with Assumption LUC which imposes that $m_d(s) = E_O[Y|S=s,D=d]$, but not with Assumptions LIV and TI since they only relate components of m to each other. Let $\mathcal{H}_n(\gamma)$ and $\mathcal{H}_n(m|\gamma)$ be the projection and fibers of $\mathcal{H}_n(m,\gamma)$, respectively. I propose to estimate $\mathcal{H}(\tau)$ using:

$$\mathcal{H}_n(\tau) := \left[\min_{\tilde{\gamma}\in\mathcal{H}_n(\gamma)} \min_{\tilde{m}\in\mathcal{H}_n(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}_n(\gamma)} \max_{\tilde{m}\in\mathcal{H}_n(\tilde{m}|\tilde{\gamma})} T(m,\tilde{\gamma})\right]. \tag{1.27}$$

If maximal and minimal selectors of $\mathcal{H}(m|\cdot)$ exist, let $U_{n,\gamma}$ and $L_{n,\gamma}$ be their plug-in sample analogs. Then, the following equivalence holds:

$$\mathcal{H}_n(\tau) = \left[\min_{\tilde{\gamma}\in\mathcal{H}_n(\gamma)} T(L_{\tilde{\gamma}},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}_n(\gamma)} T(U_{\tilde{\gamma}},\tilde{\gamma})\right] \tag{1.28}$$

The next theorem proves that (1.27) and (1.28) are consistent in the Hausdorff distance.

Theorem 1.3.3. *Let Assumptions RA, EV, MA, and E hold. Then as* $n \to \infty$*:*

$$d_H(\mathcal{H}_n(\tau), \mathcal{H}(\tau)) := max \left\{ \sup_{\tau_0 \in \mathcal{H}(\tau)} \inf_{\hat{\tau} \in \mathcal{H}_n(\tau)} ||\tau_0 - \hat{\tau}||, \sup_{\hat{\tau} \in \mathcal{H}_n(\tau)} \inf_{\tau_0 \in \mathcal{H}(\tau)} ||\tau_0 - \hat{\tau}|| \right\} \xrightarrow{p} 0.$$

The proof relies on the fact that T is a continuous functional in finite-dimensional spaces which implies that it is sufficient to show that $d_H(\mathcal{H}_n(m,\gamma), \mathcal{H}(m,\gamma)) \xrightarrow{p} 0$ to ensure $d_H(\mathcal{H}_n(\tau), \mathcal{H}(\tau)) \xrightarrow{p} 0$. I do so by applying arguments of Russell (2021, Theorem 2) to verify conditions of Shi and Shum (2015, Theorem 2.1). $\mathcal{H}_n(m,\gamma)$ is numerically equivalent to the consistent criterion-based estimator from Shi and Shum (2015, Theorem 2.1) whenever $\mathcal{H}_n(m,\gamma) \neq \emptyset$. This happens with probability approaching 1 for large n, yielding consistency of $\mathcal{H}_n(m,\gamma)$ and thus the plug-in procedure.

Remark 1.3.3. $\mathcal{H}_n(m,\gamma)$ and hence $\mathcal{H}_n(\tau)$ may be empty in finite samples even when $\mathcal{H}(m,\gamma)$ and $\mathcal{H}(\tau)$ are not. In that case, one may consistently estimate $\mathcal{H}(m,\gamma)$ using the estimator of Shi and Shum (2015) which will always be nonempty. In turn, this will yield a nonempty estimate of $\mathcal{H}(\tau)$. However, doing so may be computationally demanding as it involves finding minimizers of a function that may be very high dimensional depending on k. The plug-in procedure proposed here is computationally parsimonious and numerically equivalent when it produces a non-empty set. Hence, researchers may prefer to first attempt plug-in estimation and resort to the criterion approach should the plug-in yield an empty set.

Section 1.4. Results on the Roles of Assumptions and Data

Since S is observed in both datasets, the sole purpose of experimental data lies in providing *additional* information on γ. This section examines how this can be beneficial.

Let $\mathcal{H}^O(m,\gamma)$ be the identified set for (m,γ) if only observational data are used. Continue denoting by $\mathcal{H}(m,\gamma)$ the identified set for (m,γ) when both datasets are used. If (m,γ) are consistent with both datasets, they must be consistent with just one dataset under the same assumptions. Thus, $\mathcal{H}(m,\gamma) \subseteq \mathcal{H}^O(m,\gamma)$. Usually $\mathcal{H}(m,\gamma) \subsetneq \mathcal{H}^O(m,\gamma)$ with or without modeling assumptions in the observational data because experimental data provide additional information on γ. By definition (1.9), the corresponding identified sets for τ are:

$$\begin{aligned} \mathcal{H}^O(\tau) &= \left\{T(m,\gamma): \ \ (m,\gamma) \in \mathcal{H}^O(m,\gamma)\right\} \\ \mathcal{H}(\tau) &= \left\{T(m,\gamma): \ \ (m,\gamma) \in \mathcal{H}(m,\gamma)\right\} \end{aligned} \tag{1.29}$$

recalling that $T(m,\gamma) = \int_{\mathcal{S}} m_1(s)d\gamma_1(s) - \int_{\mathcal{S}} m_0(s)d\gamma_0(s)$. By definition $\mathcal{H}(\tau) \subseteq \mathcal{H}^O(\tau)$. Similarly, let $\mathcal{H}^O(P_{Y(0),Y(1)})$ and $\mathcal{H}(P_{Y(0),Y(1)})$ be the corresponding identified sets for the distribution function $P_{Y(0),Y(1)}$ and observe that $\mathcal{H}(P_{Y(0),Y(1)}) \subseteq \mathcal{H}^O(P_{Y(0),Y(1)})$

Subsection 1.4.1. Central Role of Modeling Assumptions

I first ask whether it is possible to have $\mathcal{H}(\tau) \subsetneq \mathcal{H}^O(\tau)$ if no modeling assumptions are imposed. This would be desirable as then the additional identifying power would solely be the result of the experimental assignment under exprerimental external validity. However, this is not the case.

Proposition 1.4.1. *Suppose Assumptions RA and EV hold. Then:*

i) $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$;

ii) $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$.

On their own, the experimental data bring *no identifying power for* τ or any functional of $P_{Y(0),Y(1)}$. Modeling assumptions in the observational data are *central* in the identification argument for τ, mirroring their importance in conventional observational studies. They are *necessary* to benefit from the existence of the short-term experiment in terms of identifying power. Corollary A.1.1 in Appendix A.1.1 further proves that without such assumptions: 1) τ is unindentified so $\mathcal{H}(\tau) = \mathbb{R}$ when $\mathcal{Y} = \mathbb{R}$; 2) $\mathcal{H}(\tau)$ is equivalent to the bounds of Manski (1990) when the support of $Y(d)$ is bounded. Modeling assumptions in the observational dataset are thus necessary to identify at least the sign of τ.

The intuition behind the result is simple. Since $S(d)$ is revealed whenever $D = d$, experimental data only provide more information on the distribution of $S(d)$ for individuals who choose $D \neq d$. However, for them, no data restrict the relationship between $Y(d)$ and $S(d)$. If this relationship is left unrestricted, then additional information on $S(d)$ does not yield more information on the distribution of $Y(d)$.

Remark 1.4.1. The roles of datasets and modeling assumptions are a topic of ongoing discussion. A seemingly similar analysis can be found in Park and Sasaki (2024b); however, the conclusion is fundamentally different. They find that observational data alone yield worst-case bounds on the treatment effects on treated survivors (ATETS) from Vikström, Ridder, and Weidner (2018), and demonstrate that combined data may be more informative under Assumption LUC. They thus do not uncover the central role of modeling assumptions.

Athey, Chetty, and Imbens (2020, Lemma 2) show that the addition of experimental data is not sufficient to point identify τ in the absence of modeling assumptions, but note that it has identifying power. I further clarify that it may have identifying power for functionals of distributions pertaining to short-term potential outcomes $S(d)$. However,

the addition of experimental data provides no identifying power for any functional of $P_{Y(0),Y(1)}$, in the absence of modeling assumptions.

Subsection 1.4.2. Auxiliary Amplifying Role of Experimental Data

Since modeling assumptions are central, experimental data have an *auxiliary role*. To make the role precise, continue to denote by $\mathcal{H}^O(\tau)$ the identified set for τ when only observational data are used without modeling assumptions, and let $\mathcal{H}^{O/A}(\tau)$ be the identified set when a modeling assumption is added. Finally, denote by $\mathcal{H}(\tau)$ the identified set from combined data under the modeling assumption. It is easy to see that by definition $\mathcal{H}(\tau) \subseteq \mathcal{H}^{O/A}(\tau) \subseteq \mathcal{H}^O(\tau)$.

By Proposition 1.4.1, more information on γ does not result in tighter bounds on τ alone. Any modeling assumption that only restricts γ thus cannot provide more information on τ. Therefore, any set of assumptions that has identifying power for τ must also restrict m, so $\mathcal{M}^A \subsetneq \mathcal{M}$. This yields the following observations.

First, modeling assumptions restricting m may be informative of τ even in the absence of experimental data, since some information on γ is available in both datasets. It is possible that $\mathcal{H}^{O/A}(\tau) \subsetneq \mathcal{H}^O(\tau)$. Second, more information on γ may make assumptions restricting m more informative. Experimental data may thus *amplify* the identifying power of modeling assumptions so $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/A}(\tau)$. Third, $\mathcal{H}(\tau) = \mathcal{H}^{O/A}(\tau)$ is possible. So experimental data do not *necessarily* amplify the identifying power of modeling assumptions. The following remark illustrates these three points using Assumption LUC. Similar results can be derived for other selection assumptions.

Remark 1.4.2. Proposition A.1.1 in Appendix A.1.1 demonstrates that: 1) LUC provides identifying power for τ without experimental data for common data distributions, so $\mathcal{H}^{O/A}(\tau) \subsetneq \mathcal{H}^{O}(\tau)$ is possible; 2) Since LUC point identifies τ with combined data, usually $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/A}$; 3) there exist data distributions for which LUC point identifies τ without experimental data, so $\mathcal{H}(\tau) = \mathcal{H}^{O/A}(\tau)$ is possible.

1.4.2.1. The Importance of Plausible Modeling Assumptions. In terms of the importance of modeling assumptions, approaches that rely on data combination effectively conduct observational studies. The amplifying role of the experimental data emphasizes the importance of plausible modeling assumptions. If the assumptions fail, adding experimental data may be *detrimental*. To see this, suppose a modeling assumption fails and let $\tilde{\mathcal{H}}$ be the misspecified identified set for τ following from combined data. Similarly, let $\tilde{\mathcal{H}}^{O/A}$ be the misspecified set that follows from observational data under the same assumptions. Any value consistent with both datasets must be consistent with just one dataset, so $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$.

Lemma 1.4.1. *(Nested Misspecification) Let $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$ be misspecified identified sets for some parameter τ. Let d be the point-to-set distance defined as $d(A,t) := \inf\{\|t-a\| : a \in A\}$ for $A \subseteq \mathbb{R}$ and $t \in \mathbb{R}$. Then:*

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) \leq d(\tilde{\mathcal{H}}, \tau)$$

Lemma 1.4.1 states that further reducing the size of any misspecified identified set *necessarily* produces results that are weakly farther away from the truth. Thus, adding experimental data can only move the resulting identified set farther away from the true τ

when a modeling assumption fails. In that case, the researcher may only obtain results closer to the ground truth by discarding the available experimental data, and these results may be informative.

Example A.1.1 in Appendix A.1.1 shows that $\tilde{\mathcal{H}}^{O/A}$ can be informative of the sign of τ and strictly, not only weakly, closer to τ than $\tilde{\mathcal{H}}$ when the modeling assumption fails. It relies on a non-pathological data-generating process and standard assumptions. It also demonstrates that adding experimental data may lead the researcher to incorrectly dismiss the true value of τ. We may have $\tau \notin \tilde{\mathcal{H}}$ and $\tau \in \tilde{\mathcal{H}}^{O/A}$, but never the converse. If the modeling assumption holds, τ is in both identified sets, and adding experimental data cannot produce results farther away from the truth.

Remark 1.4.3. Lemma 1.4.1 is a general misspecification result. It implies that reducing the size of the identified set can never result in the set being closer to the truth. Here, the reduction may happen through the addition of data. More commonly, it is a result of layering additional assumptions.

Section 1.5. Conclusion

Recent literature proposes augmenting long-term observational studies with short-term experiments to provide more credible alternatives to conventional long-term observational studies. This paper shows that data combination is not a replacement for tenable modeling assumptions. However, it remains appealing for the purpose.

Assumptions relating short-term to long-term potential outcomes may be defensible based on economic theory or intuition, and thus conducive to plausible inference. Data

combination may be used to amplify the identifying power of such assumptions and thereby yield more informative plausible inference than observational data alone.

This paper introduces two assumptions that utilize this aspect of data combination. It also provides a general identification approach that enables computational derivation of bounds under new modeling assumptions, facilitating further developments. Tailor-made assumptions that are plausible in specific empirical settings are an interesting topic for future research which may benefit from these results.

CHAPTER 2

Measuring Diagnostic Test Performance Using Imperfect Reference Tests: A Partial Identification Approach

Section 2.1. Introduction

Diagnostic tests are indispensable in modern clinical decision making. As they are almost never perfect, evaluation of test performance is a common research goal. Test performance studies seek to quantify test accuracy predominantly in the form of sensitivity and specificity, also referred to as performance measures or operating characteristics. Sensitivity (true positive rate) is the probability that a test will return a positive result for an individual who truly has the underlying condition, while specificity (true negative rate) is the probability that a test will produce a negative result for an individual who does not have the underlying condition. Equivalently, one can measure false positive and false negative rates. False negative rate and sensitivity sum to unity, as do specificity and the false positive rate.

Determining sensitivity and specificity for a diagnostic test of interest, referred to as an index test, requires knowledge of the true health status for all participants in the study. The true health status is most often unobservable, so a reference test is commonly used in lieu of it. However, such tests are rarely perfect themselves. When the reference is imperfect, conventional studies only identify "apparent" sensitivity and specificity, or

the so-called rates of positive and negative agreement with the reference.[1] They measure performance with respect to the reference test and not true performance. Hence, "apparent" parameters are typically not of interest. Moreover, the true performance measures are usually only partially identified, as shown in Section 2.2. In other words, there exists a set of parameter values that are consistent with the observed data, called the identified set. The smallest such set under maintained assumptions, or the set that exhausts all information from the data, is known as the sharp identified set. This paper addresses the issue of finding, estimating and doing inference on the points in the sharp identified set for sensitivity and specificity, or equivalently false negative and positive rates, under standard assumptions used in the literature.

I first characterize the sharp joint identified set for the true performance measures without imposing any assumptions on the latent statistical dependence between the index and reference tests conditional on health status, assuming exact or approximate knowledge of the reference test characteristics. The set is a line segment or a union of line segments in $[0,1]^2$, in contrast to rectangular sets following from comparable existing bounds (Thibodeau (1981), Emerson et al. (2018)). The proposed framework allows researchers to layer on additional assumptions regarding the latent dependence to further reduce the size of the set. This is demonstrated through a formalization of an informally stated assumption from the literature which is plausible when the two tests are physiologically related.

Sensitivity and specificity are frequently used to obtain other policy-relevant parameters. I present how the derived identified sets may be utilized to sharply bound prevalence, or

1. "Apparent" false negative rate and "apparent" sensitivity sum to unity. Similarly, "apparent" false positive rate and "apparent" specificity sum to unity.

the population rate of illness, in populations screened by the index test. Bounds may be markedly narrower than those implied by existing comparable methods, owing to the specific shape of the identified sets. Implied bounds on predictive values, i.e. probabilities that a patient is sick conditional on observing a test result, are discussed in Appendix B.1.

The FDA Statistical Guidance on Reporting Results Evaluating Diagnostic Tests[2] requires diagnostic performance studies to report confidence intervals for index test sensitivity and specificity to quantify the statistical uncertainty in the estimates. To conform to the practice, I construct confidence sets for points in the identified set using an inference method based on moment inequalities (Romano, Shaikh, and Wolf (2014)). The confidence sets are uniformly consistent in level over a large family of permissible distributions relevant in the application. Namely, they asymptotically cover all points in the identified set uniformly over the family of population distributions with probability of at least $1 - \alpha$, where α is the chosen significance level.

Diagnostic test performance studies for rapid COVID-19 tests have a mandated RT-PCR reference test which is known to produce false negative results, and thus pertain to the setting analyzed in the paper.[3] Fitzpatrick et al. (2021) emphasize that the false negative rate of the ubiquitous *Abbott BinaxNOW COVID-19 Ag2 CARD* rapid antigen test may be substantially understated by the reported "apparent" analog due to imperfect reference tests.[4]

Application of the method to the data from the original Emergency Use Authorization (EUA) performance study, as well as an independent study by Shah et al. (2021) bolsters

2. Link: https://www.fda.gov/media/71147/download (Last accessed: 12/25/2022)
3. Link: https://www.fda.gov/media/137907/download (Last accessed: 12/25/2022)
4. The test held 75% of the COVID-19 antigen test market share in the United States, according to Abbott Laboratories CEO Robert Ford on Q3 2021 Results - Earnings Call Transcript.

this claim and reveals that bounds can be very informative. Depending on interpretation, the results from both studies suggest that the test may not satisfy the initial FDA requirement for EUA of at least 80% estimated sensitivity, despite fulfilling the criterion of high "apparent" sensitivity, implying the need for alternative testing protocols. Moreover, the estimated false negative rates for symptomatic and asymptomatic patients are up to 3.17 and 4.59 times higher than the frequently cited "apparent" false negative rate, warranting further attention. Comparison with existing bounds reveals that the proposed method can provide significant reductions in the size of the identified set for operating characteristics, and consequently in the width of implied bounds on prevalence when the test is used for screening.

The methodological framework developed in the paper offers solutions to two issues in the current research practice guidelines set forth by the FDA Statistical Guidance, as explained by Remark 2.2.4: 1) Inability to measure true test performance in common settings; 2) Inability to compare index and reference test performance. It also addresses the concerns raised in Boyko, Alderman, and Baron (1988): *"When two tests are strongly suspected of being conditionally dependent, then the performance of one of these tests should probably not be compared with that of the other, unless better methods are developed to sort out the degree of bias caused by reference test errors in the presence of conditional dependence."*

Provided replication files allow researchers to directly utilize the findings of the paper to obtain estimates and confidence sets in their own work.[5] Since the method does not require any changes to the data-collection process of standard studies, it can also be readily

5. Available from: https://github.com/obradovicfilip/bounding_test_performance

applied to estimate test performance based on published data, as demonstrated by the application section of the paper.

Broader applicability of the approach is discussed in Section 2.6. It can be used to study features of the joint distribution between a binary outcome and a binary latent variable measured with an imperfect proxy. Illustrative examples include variables such as program participation as indicated by a survey response, or race, imputed using the Bayesian Improved Surname Geocoding (BISG) algorithm. The method is appealing when validation studies measuring proxy misclassification rates exist. The bounds also readily apply whenever one wishes to learn performance of a binary classifier by comparing it to another imperfect classifier or label, rather than the ground truth, which is common in satellite imagery and other remote sensing applications.

Subsection 2.1.1. Related Literature

Gart and Buck (1966), Staquet et al. (1981), and Zhou, McClish, and Obuchowski (2009) show that if the reference and index tests are statistically independent conditional on the true health status, index test sensitivity and specificity are point identified, assuming exactly known reference test performance measures. (see also Hui and Walter (1980)) However, Vacek (1985), Valenstein (1990), Hui and Zhou (1998) and Emerson et al. (2018) elaborate that conditional independence may frequently be untenable. A salient case is when the two tests are physiologically related, such as when they rely on the same type of sample (e.g. nasal swab or capillary blood) or measure the same quantities (e.g. antibody reaction to tuberculin).

Tests are generally expected to be dependent, but the dependence structure is latent since true status is unobservable. Several authors explore how dependence between the tests may affect the direction of gold standard bias, defined as the difference between the "apparent" and true performance measures (Deneef (1987), Boyko, Alderman, and Baron (1988)). Valenstein (1990) concludes that when errors committed by index and reference tests are highly correlated, the "apparent" measures may overstate the true parameters. However, they do not precisely define highly correlated errors, prompting the formalization of the assumption in this paper. Authors focus on the direction of the effects of the conditional dependence, rather than on the magnitude. The purpose is to allow researchers to determine whether their estimates are biased upwards or downwards. However, since the dependence cannot be measured, the practical relevance of these findings is diminished. Additionally, one could argue that the magnitude is perhaps even more important than the direction of the bias.

A formal approach to the issue of unknown bias magnitude is found in Thibodeau (1981). Assuming that the reference and index tests are positively correlated, and that the reference is at least as accurate as the index, the author bounds the bias. The framework presented below does not require such assumptions. More recently, Emerson et al. (2018) sketch an argument for individual bounds on sensitivity and specificity under similar assumptions used in this paper. This study contributes to the literature on gold standard bias by presenting the sharp joint identified set for test performance measures, formalizing and incorporating existing dependence assumptions to further reduce its size, bounding derived parameters of interest, and providing an appropriate uniform inference procedure.

In doing so, it builds upon on existing work on partial identification (Manski (2003), Manski (2007)). Proposition 2.2.1 revitalizes the general analysis of Cross and Manski (2002) in the context of diagnostic test performance measurements. Technical contributions over their work primarily lie within the novel identification findings in Propositions 2.2.2 and 2.3.1, as well as the inference procedure for the points in the identified set that is uniformly consistent in level. The procedure assumes a mild and easily interpretable restriction on the family of population distributions, and relies on existing methods for inference in moment inequality models (Andrews and Soares (2010), Andrews and Barwick (2012), Chernozhukov, Lee, and Rosen (2013), Romano, Shaikh, and Wolf (2014), Canay and Shaikh (2017), Bugni, Canay, and Shi (2017), Chernozhukov, Chetverikov, and Kato (2019), Kaido, Molinari, and Stoye (2019), Bai, Santos, and Shaikh (2021)).

This paper aligns with a growing body of literature concerning partial identification in medical and epidemiological research (Bhattacharya, Shaikh, and Vytlacil (2012), Manski (2020), Toulis (2021), Manski and Molinari (2021), Ziegler (2021), Stoye (2022)). It also aims to contribute to the corpus of COVID-19 test performance studies by estimating the true sensitivity and specificity for COVID-19 antigen tests despite reference test imperfections under plausible assumptions (Shah et al. (2021), Pollock et al. (2021), Siddiqui et al. (2021)).

The remainder of the paper is organized as follows. Section 2.2 provides the identification argument. Section 2.3 discusses identification of prevalence. Section 2.4 explains estimation and inference. Section 2.5 presents confidence and estimated identified sets for the operating characteristics of the COVID-19 antigen test. Section 2.6 indicates uses of the results beyond the context of diagnostic test performance studies. Section 2.7

concludes. Appendix B.1 derives bounds on predictive values. All proofs are collected in Appendix B.5.

Section 2.2. Identification

Studies quantifying the performance of a test of interest, also known as an index test, require knowledge of the true health status. Health status is usually unobservable, so it is determined by an alternative test, called the reference test. Even though the reference test should be the best available test for the underlying condition, it is almost always imperfect in practice, giving rise to identification issues. Let $t = 1$ and $r = 1$ if the index and reference tests, respectively, yield positive results and $t = 0$, $r = 0$ otherwise. Let $y = 1$ denote the existence of the underlying condition we are testing for and $y = 0$ the absence of it.[6]

We are interested in learning the sensitivity and specificity of the index test:

$$\text{Sensitivity: } \theta_1 = P(t = 1|y = 1) \tag{2.1}$$

$$\text{Specificity: } \theta_0 = P(t = 0|y = 0) \tag{2.2}$$

which are defined when $P(y = 1) \in (0, 1)$ in the study population. Equivalently, one can study the false negative and false positive rates, $1-\theta_1$ and $1-\theta_0$. Similarly, define reference test sensitivity $s_1 = P(r = 1|y = 1)$ and specificity $s_0 = P(r = 0|y = 0)$. Data collection in test performance studies is commonly done by testing all participants with both the reference and index tests. The observed outcome for each participant is $(t, r) \in \{0, 1\}^2$.

6. I interchangeably say that the person is ill when $y = 1$ and when $y = 0$, that they are healthy. This can be extended to encompass antibody tests with minor semantic changes, since they can also measure if a person has been ill.

The data identify the joint probability distribution $P(t, r)$. When $P(r = 1) \in (0, 1)$, "apparent" sensitivity $\tilde{\theta}_1 = P(t = 1|r = 1)$, and "apparent" specificity $\tilde{\theta}_0 = P(t = 0|r = 0)$ are also identified.

It is typically assumed that the reference test is perfect, so that $r = y$. Then $(\tilde{\theta}_1, \tilde{\theta}_0) = (\theta_1, \theta_0)$. This is rarely the case in practice. Generally, $\tilde{\theta}_j \neq \theta_j$ for some $j = 0, 1$, which referred to as *gold standard bias*. Interpreting $(\tilde{\theta}_1, \tilde{\theta}_0)$ as true performance measures can lead to severely misleading conclusions due to the bias. Alternatively, researchers may explicitly study $(\tilde{\theta}_1, \tilde{\theta}_0)$. However, they only measure performance of t with respect to r, and not y. If one wishes to learn about true performance (θ_1, θ_0), then these parameters are not of interest.

Remark 2.2.1. Index test t is usually a novel test, evaluated against the best currently available test r. In some settings, use of r may not be practical outside the performance study due to high costs, long turnaround time or invasiveness. For example, a reference test for some types of dementia is a postmortem neuropathological report which is not helpful for diagnosis. Viral antigen tests may be preferred over reference RT-PCR tests for screening purposes due to lower resource requirements.

Focusing the analysis on binary tests and binary health statuses is standard procedure. FDA Statistical Guidance on Reporting Results Evaluating Diagnostic Tests recognizes only binary reference tests and health statuses, explicitly stating: *"A reference standard ... divides the intended use population into only two groups (condition present or absent)."* Many tests that yield discrete or continuous test results, such as RT-PCR tests, are reduced to binary tests by thresholding in practice. While the results of this paper can be extended

to cases in which ranges of t and r are finite sets, I limit the analysis to the binary setting to conform to current research practice.

The section begins by outlining the formal assumptions used. I then provide the set of parameter values (θ_1, θ_0) consistent with the observed data, also known as the identified set, without imposing any assumptions on the statistical dependence between t and r. The set is sharp, or the smallest possible under maintained assumptions. For simplicity of exposition, this is first done when (s_1, s_0) are known. I show how an additional assumption on the dependence structure between the two tests can be used to further reduce the size of the sharp identified set. Finally, I allow (s_1, s_0) to be approximately known by assuming $(s_1, s_0) \in \mathcal{S}$, where $\mathcal{S}$ is some known set.

Subsection 2.2.1. Assumptions

The framework in this paper relies on common assumptions maintained in the literature.

Assumption 2.2.1. *(Random Sampling) The study sample is a sequence of i.i.d random vectors $W_i = (t_i, r_i)$, where each W_i follows a categorical distribution $P(t, r)$ for $(t, r) \in \{0, 1\}^2$ and $i = 1, \ldots, n$.*

The distribution $P(t, r)$ is a marginal of the joint distribution $P(t, r, y)$. Since y is not observable, $P(t, r, y)$ is not point identified.

Assumption 2.2.2. *(Reference Performance) Sensitivity and specificity of the reference test $s_1 = P(r = 1|y = 1)$ and $s_0 = P(r = 0|y = 0)$ are known, and $s_1 > 1 - s_0$.*

The analysis is first done for the simple case when (s_1, s_0) are known exactly. The approach is then generalized in Section 2.2.4 by assuming $(s_1, s_0) \in \mathcal{S}$, where $\mathcal{S}$ is some

known set. Thus, reference test performance needs to be known only approximately. The generalization can also be used to perform sensitivity analyses. Knowledge of (s_1, s_0) or $\mathcal{S}$ is commonly assumed in work dealing with gold standard bias correction, such as Gart and Buck (1966), Thibodeau (1981), Staquet et al. (1981), and Emerson et al. (2018). The current norm of relying on the assumption that the reference test is perfect means that researchers regularly maintain $(s_1, s_0) = (1, 1)$, which is a stronger condition since it implies Assumption 2.2.2. The framework hence weakens standard assumptions. However, knowledge of (s_1, s_0) or $\mathcal{S}$ is a crucial identifying assumption and warrants further discussion.

Assumption 2.2.2 is particularly appealing when one has access to a study that identifies performance of r. Mathews et al. (2010) and Matos et al. (2011) justify assumed (s_1, s_0) based on such studies. In specific settings, y may indeed be plausibly observable, allowing for identification of (s_1, s_0). However, the procedure needed to observe y may be exceedingly costly, invasive, or have an unsuitably long turnaround time for widespread use in performance measurement. Thus, r is commonly used as a reference for t instead. For example, a neuropathological autopsy report is the only way to pose a definitive diagnosis of Alzheimer's disease, i.e. observe y (Suemoto and Leite (2023)). Studies of novel tests t for Alzheimer's disease frequently use an amyloid positron emission tomography (PET) scan or a clinical diagnosis as r, since a neuropathological report may be unobtainable (Budelier and Bateman (2020), Wang et al. (2023)). Autopsy performance studies for r in which y is directly observed post-mortem are available, identifying performance of r (Patwardhan et al. (2004)).

Unfortunately, in many cases such a performance study for r does not exist. Nevertheless, tests are often expected to have precisely measured analytical performance measures. Analytical specificity and sensitivity are defined as performance measures obtained based on contrived, rather than clinical samples. This may provide some information on how r will perform in clinical settings. For example, Kucirka et al. (2020) and Kanji et al. (2021) maintain that COVID-19 RT-PCR tests are perfectly specific owing to the absence of cross-reactivity with other pathogens, that is, due to its perfect analytical specificity.[7]

The commonly maintained assumption $(s_1, s_0) = (1, 1)$ has been disputed for a plethora of reference tests. This fact indicates that at least a set $\mathcal{S}$ of more credible values (s_1, s_0) exists for a variety of tests used as r. In these cases, the method will yield sharp bounds on (θ_1, θ_0). However, if nothing can be credibly assumed about the performance of r, one cannot reasonably use it to identify performance of t. This is not a novel observation, but it highlights the importance of the assumption. Emerson et al. (2018) state: "*If very little is known about the reference test performance, then it is clear that a comparison to such a reference test is a futile exercise and can provide no information about a new test.*" Gart and Buck (1966) similarly note that when (s_1, s_0) are unknown, both $t \perp\!\!\!\perp y$ or $t \not\perp\!\!\!\perp y$ will generally be consistent with $P(t, r)$. The choice of (s_1, s_0) is context-specific, and should be carefully considered in each study.

I further assume that $s_1 > 1 - s_0$, or that the reference test is reasonable.[8] If $s_1 = 1 - s_0$, one can show that $r \perp\!\!\!\perp y$, so the test provides no information on y. Tests are costly,

7. Specificity on contrived laboratory samples containing other pathogens, but not SARS-CoV-2.

8. The assumption does not require that both s_1 and s_0 are high. Indeed, it is possible that either s_1 or s_0 are close to 0, but that their sum is higher than 1.

and any use of such test is not rational. If $s_1 < 1 - s_0$, it would be possible to redefine $r^* = 1 - r$, so that $s_1^* = 1 - s_1$ and $s_0^* = 1 - s_0$. Now $s_1^* > 1 - s_0^*$, since $1 - s_1 > s_0$.

Assumption 2.2.3. *(Bounded Prevalence) Population prevalence* $P(y = 1)$ *satisfies* $0 < P(y = 1) < 1$.

The assumption is implicitly found in all diagnostic test performance studies measuring sensitivity and specificity, since it is necessary for them to be defined. Assumptions 2.2.2 and 2.2.3 imply that $P(r = 1) \in (1 - s_0, s_1)$. If the condition fails, at least one of the two assumptions are refuted.

If one additionally maintains that $t \perp\!\!\!\perp r|y$, then (θ_1, θ_0) are point identified (Gart and Buck (1966), Staquet et al. (1981), and Zhou, McClish, and Obuchowski (2009)). However, it is well established that conditional independence is generally untenable (Valenstein (1990), Hui and Zhou (1998) and Emerson et al. (2018)). Dependence may arise t and r are physiologically related, such as when they rely on the same type of sample or measure the same quantities. For example, tine and Mantoux tests may be dependent since they both rely on the antibody reaction to tuberculin (Vacek (1985)), and direct immunoassay and culture swab tests for *Group A streptococci* may be related since they rely on the same type of sample (Valenstein (1990)). Since y is unobserved, the dependence structure is latent, and multiple structures may be consistent with the data distribution $P(t, r)$. This leads to a possibly non-singleton set of values (θ_1, θ_0) that are consistent with the data, called the identified set. We would first like to learn this set without imposing any assumptions on the statistical dependence structure between t and r conditional on y.

Additional assumptions on the possible dependence structures may then be used to reduce the size of the identified set, as shown in Section 2.2.3.

Subsection 2.2.2. Identified Set for (θ_1, θ_0)

The data reveal $P(t, r)$, while probability distributions involving y are not directly observable. Still, $P(r, y)$ can be determined using (s_1, s_0) and $P(t, r)$. I henceforth use P_{s_1,s_0} to denote probability distributions that are derived from observable distributions given (s_1, s_0). All directly observable distributions, such as $P(t, r)$, do not have the subscript. By the law of total probability and $s_1 \neq 1 - s_0$ from Assumption 2.2.2:

(2.3)
$$P(r=1) = s_1 P_{s_1,s_0}(y=1) + (1-s_0) P_{s_1,s_0}(y=0) \Leftrightarrow P_{s_1,s_0}(y=1) = \frac{P(r=1) + s_0 - 1}{s_1 + s_0 - 1}.$$

$P_{s_1,s_0}(r, y)$ is then known from $P_{s_1,s_0}(r, y) = P_{s_1,s_0}(r|y) P_{s_1,s_0}(y)$, since (s_1, s_0) fully characterize $P_{s_1,s_0}(r|y)$. To outline the idea of finding the identified set, first note that for $j = 0, 1$:

$$\theta_j = P_{s_1,s_0}(t=j|y=j) = \frac{P_{s_1,s_0}(t=j, r=0, y=j) + P_{s_1,s_0}(t=j, r=1, y=j)}{P_{s_1,s_0}(y=j)}. \tag{2.4}$$

Probabilities $P_{s_1,s_0}(t = j, r = k, y = j)$ for $k = 0, 1$ are unobservable. However, they can be bounded using the knowledge of $P(t, r)$ and $P_{s_1,s_0}(r, y)$. By the properties of probability measures, an upper bound on $P_{s_1,s_0}(t = j, r = k, y = j)$ is $min\Big(P(t = j, r = k), P_{s_1,s_0}(r = k, y = j)\Big)$. To form a lower bound, one can similarly find that

$P_{s_1,s_0}(t=j,r=k,y=1-j) \leq min\Big(P(t=j,r=k), P_{s_1,s_0}(r=k,y=1-j)\Big)$ and use:

$$\begin{aligned} P_{s_1,s_0}(t=j,r=k,y=j) &= P(t=j,r=k) - P_{s_1,s_0}(t=j,r=k,y=1-j) \\ &\geq max\Big(0, P(t=j,r=k) - P_{s_1,s_0}(r=k,y=1-j)\Big). \end{aligned} \tag{2.5}$$

Note that these coincide with Fréchet-Hoeffding bounds for $P(t=j,y=j|r=k)$ multiplied by $P(r=k)$. Proof of Proposition 2.2.1 demonstrates that any pair of values for $P_{s_1,s_0}(t=j,r=0,y=j)$ and $P_{s_1,s_0}(t=j,r=1,y=j)$ within their respective bounds is consistent with the observed data. Hence, sharp bounds on $P_{s_1,s_0}(t=j,y=j)$ are obtained by summing the two individual set of bounds. The sharp bounds for θ_j then follow directly from (2.3) and (2.4). Finally, the sharp joint identified set for (θ_1,θ_0) is derived using $P(t=1) = \theta_1 P_{s_1,s_0}(y=1) + (1-\theta_0)P_{s_1,s_0}(y=0)$. Observe that no restrictions beyond those set by the distribution $P(t,r)$ are imposed on the latent dependence structure of t and r conditional on y.

Proposition 2.2.1. *The sharp identified set $\mathcal{H}_{(\theta_1,\theta_0)}(s_1,s_0)$ for (θ_1,θ_0) given reference test sensitivity s_1 and specificity s_0 is:*

$$\mathcal{H}_{(\theta_1,\theta_0)}(s_1,s_0) = \left\{ (t_1,t_0) : t_0 = t_1 \frac{P_{s_1,s_0}(y=1)}{P_{s_1,s_0}(y=0)} + 1 - \frac{P(t=1)}{P_{s_1,s_0}(y=0)}, t_j \in \mathcal{H}_{\theta_j}(s_1,s_0) \right\} \tag{2.6}$$

where $\mathcal{H}_{\theta_j}(s_1, s_0) = [\theta_j^L, \theta_j^U]$ *is the sharp bound on* θ_j *defined as:*

$$\begin{aligned}
\theta_j^L = &\Bigg[max\Big(0, P(t=j, r=j) - P_{s_1,s_0}(r=j, y=1-j)\Big) \\
&+ max\Big(0, P_{s_1,s_0}(r=1-j, y=j) - P(t=1-j, r=1-j)\Big)\Bigg] \frac{1}{P_{s_1,s_0}(y=j)} \\
\theta_j^U = &\Bigg[min\Big(P(t=j, r=1-j), P_{s_1,s_0}(r=1-j, y=j)\Big) \\
&+ min\Big(P(t=j, r=j), P_{s_1,s_0}(r=j, y=j)\Big)\Bigg] \frac{1}{P_{s_1,s_0}(y=j)}.
\end{aligned} \tag{2.7}$$

Proposition 2.2.1 revitalizes the general identification result of Cross and Manski (2002). Arguments therein can also be used to derive the identified set for (θ_1, θ_0).[9] Namely, $P(t|r)$ and $P_{s_1,s_0}(y|r)$ are identified from $P(t, r)$ and (s_1, s_0), yielding the identified set for the vector $(E[t|r=j, y=k])_{(j,k)\in\{0,1\}^2}$.[10] Sharp bounds on $(\theta_1, \theta_0) = (E[t|y=1], 1-E[t|y=0])$ follow. This paper relies on a constructive proof approach which can be adapted to accommodate assumptions on the latent dependence structure of t and r conditional on y. Section 2.2.3 introduces a novel formalization of an informally stated assumption from the literature and exploits this feature to further tighten the bounds.

The set $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ is a line segment on $[0, 1]^2$ for a given value of reference test operating characteristics s_1 and s_0. It collapses to a point $(\tilde{\theta}_1, \tilde{\theta}_0)$ when $(s_1, s_0) = (1, 1)$. Emerson et al. (2018) sketch an argument for individual bounds on θ_j as in (2.7) and do not discuss sharpness or the joint identified set. Proposition 2.2.1 goes further by proving that both individual bounds and the joint identified sets are the smallest possible under the

9. I thank an anonymous referee for bringing this to my attention.

10. $P_{s_1,s_0}(y=j|r=1-j) = 0$ when $s_j = 1$ for some j, which violates the assumptions of Cross and Manski (2002). It is still possible to utilize their results by bounding only $(E[t|r=j, y=1], E[t|r=j, y=0])$.

assumptions. Section 2.3 shows that the linear structure of the set $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ is crucial for sharpness of bounds on certain derived policy-relevant parameters, such as the illness rate in populations screened with the test t, otherwise known as prevalence. Bounds on prevalence are unnecessarily wide if any pair of values (θ_1, θ_0) from their individual bounds is considered feasible, so that the joint identified set is a rectangle $\mathcal{H}_{\theta_1}(s_1, s_0) \times \mathcal{H}_{\theta_0}(s_1, s_0)$. The sharp joint identified set for false negative and false positive rates $(1 - \theta_1, 1 - \theta_0)$ also directly follows from $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$. The same will hold for other identified sets for (θ_1, θ_0) below.

Example 2.2.1. Consider a study in which $(s_1, s_0) = (0.9, 0.9)$, $P(t = j, r = j) = 0.45$ and $P(t = j, r = 1 - j) = 0.05$ for $j = 0, 1$. $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ is a line segment connecting $(0.8, 0.8)$ and $(1, 1)$.

The identified set for (θ_1, θ_0) is sharp. Encountering wide bounds on sensitivity and specificity implies that is not possible to learn the operating characteristics more precisely without additional assumptions that may be untenable, or without changing the reference test. Since the reference test is supposed to be the best available test, researchers and practitioners may have to embrace the ambiguity regarding the index test performance.

Remark 2.2.2. Conventional studies maintain that the reference test is perfect so $(s_1, s_0) = (1, 1) \geq (\theta_1, \theta_0)$ component-wise, and t is irrefutably assumed to perform at most as well as r in both dimensions. The bounds allow one to empirically compare index and reference test performance for certain $P(t, r)$ and (s_1, s_0). This is possible in the following cases:

(1) $(\theta_1, \theta_0) < (s_1, s_0)$ component-wise for all $(\theta_1, \theta_0) \in \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$, demonstrating that r outperforms t in both dimensions;

(2) $\theta_j > s_j$ for a single j and all $(\theta_1, \theta_0) \in \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$, demonstrating that t outperforms r in one dimension.

Lemma B.4.3 in Appendix B.4 shows that $(\theta_1, \theta_0) > (s_1, s_0)$ component-wise is impossible for all $(\theta_1, \theta_0) \in \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$. That is, a single study cannot show that t outperforms r in both dimensions. As illustrated by Example 2.2.1, it is also possible that there exist $(\theta_1, \theta_0), (\theta_1', \theta_0') \in \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ such that $(\theta_1, \theta_0) > (s_1, s_0)$ and $(\theta_1', \theta_0') \leq (s_1, s_0)$ component-wise. This indicates that t or r may outperform the other test in both dimensions, but it is inconclusive. One cannot exclude the possibility that either test outperforms the other.

Remark 2.2.3. Depending on (s_1, s_0) and $P(t, r)$, "apparent" measures $(\tilde{\theta}_1, \tilde{\theta}_0)$ need not be contained in the identified set for (θ_1, θ_0). In that sense, $(\tilde{\theta}_1, \tilde{\theta}_0)$ may be over- or understating (θ_1, θ_0). A relevant empirical example is found in Section 2.5.

Remark 2.2.4. The FDA Statistical Guidance defines a *reference standard* for a condition as: *"The best available method for establishing the presence or absence of the target condition. ... established by opinion and practice within the medical, laboratory, and regulatory community."* The guidance does not require a reference standard to be perfect, as it rarely is. When used as a reference test, the estimates may be reported as pertaining to sensitivity and specificity, even though the estimands are "apparent" measures when it is imperfect. This practice can be misleading. Tests other than the reference standard may

be used as reference tests. However, then the estimates should be reported as "apparent". If one wishes to learn true test performance, they are typically not of interest.

The guidance does not require or suggest any corrections that would allow researchers to form adequate estimates of the true operating characteristics in either case. The method in this paper proposes a solution by forming the smallest possible bounds on true performance measures under standard assumptions. Furthermore, the guidance emphasizes that in a conventional study one cannot determine whether t or r has better performance. Remark 2.2.2 clarifies that bounds allow for comparison of performance between the two tests in certain cases.

Remark 2.2.5. Bounds in (2.7) could be formed from the marginals $P(t = j)$ and $P_{s_1,s_0}(y = j)$ as $\theta_j \in \Big[max\Big(0, P(t = j) + P_{s_1,s_0}(y = 1 - j)\Big), min\Big(P(t = j), P_{s_1,s_0}(y = j)\Big)\Big]_{\frac{1}{P_{s_1,s_0}(y=j)}}$. The literature on data combination suggests that these are not sharp, as outlined by Ridder and Moffitt (2007). Lemma B.4.2 in Appendix B.4 shows that they are at least as wide as those in Proposition 2.2.1.

Subsection 2.2.3. Misclassification Assumptions

Points in the identified set $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ derived in the previous section correspond to different non-observable probability distributions $P_{s_1,s_0}(t, r, y)$ that are consistent with the identified distribution $P(t, r)$ and (s_1, s_0). Until this point no additional restrictions on the dependence structure between t, r and y were imposed. Literature on gold standard bias suggests that t and r may frequently be dependent conditional on y in ways that would further restrict the set of distributions $P_{s_1,s_0}(t, r, y)$ consistent with the data, resulting in

more informative identified sets for (θ_1, θ_0). It is thus important to incorporate assumptions on the dependence structure into the framework.

A particular kind of restrictions that researchers may be willing to consider concern the error probabilities of t conditional on r making a misclassification error for a specific value of y. Researchers may scrutinize the credibility of such assumptions based on physical properties of the two tests. Valenstein (1990) informally discusses one such restriction. The author analyzes the magnitude of the difference $\theta_j - \tilde{\theta}_j$ for $j = 0, 1$ by means of a numerical example when the two tests have classification errors that are referred to as "highly correlated". The meaning of highly correlated errors is not formally defined, and in the numerical example the assumption is imposed as $P(t \neq y | r \neq y, y) = P(t = 1 - y | r = 1 - y, y) = 1$ for all y. I formalize this assumption and derive the resulting sharp identified set for (θ_1, θ_0). Given that its plausibility may vary across health statuses, I allow it to hold only for a particular value of y.

Definition 2.2.1. (Tendency to wrongly agree) An index test has a tendency to wrongly agree with the reference test for disease status $\bar{y}$ given (s_1, s_0) if $P_{s_1,s_0}(t = 1 - \bar{y} | r = 1 - \bar{y}, y = \bar{y}) \geq P_{s_1,s_0}(t = \bar{y} | r = 1 - \bar{y}, y = \bar{y})$.

If an index test exhibits a tendency to wrongly agree with the reference test for $\bar{y}$, conditional on the reference test making a classification error, the index test is more likely to misdiagnose the patient than to diagnose them correctly. Valenstein (1990) explains that the tendency may arise if the two tests have common properties, such as the type of sample used, e.g. the same swab type.

Proposition 2.2.2. *Let θ_j^L be as in (2.7). When the index and reference tests have a tendency to wrongly agree only for $y = j$, the sharp bounds on θ_j given (s_1, s_0) are $\bar{\mathcal{H}}_{\theta_j}(s_1, s_0) = [\theta_j^L, \bar{\theta}_j^U]$, where:*

$$\begin{aligned}\bar{\theta}_j^U = &\Bigg[min\Big(P(t=j, r=1-j), \frac{P_{s_1,s_0}(r=1-j, y=j)}{2}\Big) \\ &+ min\Big(P(t=j, r=j), P_{s_1,s_0}(r=j, y=j)\Big)\Bigg] \frac{1}{P_{s_1,s_0}(y=j)}.\end{aligned} \tag{2.8}$$

If the index and reference tests have a tendency to wrongly agree for $y = 0$ and $y = 1$, the sharp bounds on θ_j for $j = 0, 1$ given (s_1, s_0) are $\bar{\bar{\mathcal{H}}}_{\theta_j}(s_1, s_0) = [\theta_j^L, \bar{\bar{\theta}}_j^U]$, where:

$$\begin{aligned}\bar{\bar{\theta}}_j^U = &\Bigg[min\Big(P(t=j, r=1-j), \frac{P_{s_1,s_0}(r=1-j, y=j)}{2}\Big) \\ &+ min\Big(P(t=j, r=j) - \frac{P_{s_1,s_0}(r=j, y=1-j)}{2}, P_{s_1,s_0}(r=j, y=j)\Big)\Bigg] \\ &\times \frac{1}{P_{s_1,s_0}(y=j)}.\end{aligned} \tag{2.9}$$

Sharp joint identified sets $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0)$ and $\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1, s_0)$ for (θ_1, θ_0) given (s_1, s_0) follow from (2.6), $\bar{\mathcal{H}}_{\theta_j}(s_1, s_0)$, and $\bar{\bar{\mathcal{H}}}_{\theta_j}(s_1, s_0)$.

Proposition 2.2.2 provides sharp identified sets for (θ_1, θ_0) when the researcher maintains that the tests have a tendency to wrongly agree for only one or both health statuses.[11] Both sets given (s_1, s_0) are again line segments in $[0, 1]^2$. The bounds $[\theta_j^L, \bar{\theta}_j^U]$, and $[\theta_j^L, \bar{\bar{\theta}}_j^U]$

11. One can also define the tendency to correctly disagree for disease status $\bar{y}$ as $P_{s_1,s_0}(t = 1 - \bar{y} | r = 1 - \bar{y}, y = \bar{y}) \leq P_{s_1,s_0}(t = \bar{y} | r = 1 - \bar{y}, y = \bar{y})$. Identified sets that follow can easily be derived symmetrically. Thibodeau (1981) emphasizes that tests are generally not expected to exhibit negative dependence. However, the formulation may be beneficial in applications described in Section 2.6.

imply that the sets may be reduced in size only from above compared to $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$. This can be seen in Example 2.2.2 below.

It is important to note that the assumption may or may not have identifying power for a given $P(t, r)$ and (s_1, s_0). This is evident in the empirical application. Remark 2.5.3 notes that the assumption effectively halves the size of the estimated identified set in one population, but has no effect in the remaining two. Remark 2.2.6 characterizes sufficient and necessary conditions for the assumption to have identifying power, and provides an easily verifiable necessary condition.

Remark 2.2.6. Lemma B.4.4 in Appendix B.4 shows that the tendency to wrongly agree for $y = j$ has identifying power if and only if $P(t = j, r = 1-j) > \frac{P_{s_1,s_0}(r=1-j,y=j)}{2} > 0$. If $s_j = 1$, the assumption cannot have identifying power.

For the purpose of interpreting this result, suppose that the tendency to wrongly agree holds for $y = j$. If $s_j = 1$, $\{r = 1 - j, y = j\}$ is a probability zero event and properties of t on it are inconsequential. Having $s_j < 1$ is thus necessary for reducing the size of the identified set. Assumption 2.2.3 and $s_j < 1$ imply $P_{s_1,s_0}(r = 1 - j, y = j) > 0$. For the assumption to have identifying power we then only need $P(t = j, r = 1 - j)$ to be "large enough". The definition of "large enough" is contingent upon $P(r = 1)$ and (s_1, s_0). In Example 2.2.2 below, the threshold is $P(t = j, r = 1 - j) > 2.5\%$ for $j \in \{0, 1\}$.

Example 2.2.2. Consider the study as in Example 2.2.1. If the tests have a tendency to wrongly agree for $y = 1$, $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0)$ is a line segment with end points $(0.8, 0.8)$ and $(0.95, 0.95)$. If they have a tendency to wrongly agree for any y, $\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1, s_0)$ is a line segment with end points $(0.8, 0.8)$ and $(0.9, 0.9)$.

Remark 2.2.7. The identified set $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ was derived by finding all distributions $P_{s_1,s_0}(t, r, y)$ that are consistent with the data given (s_1, s_0). It thus represents a domain of consensus for the values of (θ_1, θ_0) under additional assumptions restricting the set of $P_{s_1,s_0}(t, r, y)$ that are considered to be feasible. In other words, any sharp identified set obtained under further assumptions on the statistical dependence of t, r, and y will be a subset of $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$.

Given that SARS-CoV-2 RT-PCR and rapid antigen swab tests rely on the same type of sample usually taken from the same location (e.g. nasopharynx, nares or oropharynx), it may be plausible to maintain that the two tests have a tendency to wrongly agree. We will use the assumption in the empirical application in Section 2.5. More examples can be found in the literature. Hadgu (1999) observes that the same assumption is credible for the ligase chain reaction (LCR) and culture tests for *Chlamydia trachomatis* by the same reasoning. Valenstein (1990) indicates that when determining the performance of direct immunoassay swab tests for *Group A streptococci* using a culture as a reference, the tendency to wrongly agree may hold for $y = 1$ due to inadequately obtained samples leading to false negatives. The same is suggested for $y = 0$. Patients who are ill with viral pharyngitis, but incidentally carry the bacteria elsewhere, may appear falsely positive on both tests. Vacek (1985) argues that tine and Mantoux tuberculin tests may have a tendency to wrongly agree for any y as both rely on the antibody reaction to tuberculin.

Subsection 2.2.4. Imperfect Knowledge of Reference Test Characteristics

For simplicity of exposition, previously derived identified sets for (θ_1, θ_0) were presented under the premise that (s_1, s_0) are known exactly. That assumption might be implausible

depending on the setting. Researchers may instead prefer to maintain that they do not possess exact, but rather approximate knowledge of (s_1, s_0). I thus relax Assumption 2.2.2 by supposing that we only have knowledge of a set $\mathcal{S}$ that contains true sensitivity and specificity of the reference test.

Assumption 2A. Sensitivity and specificity of the reference test are contained in a known compact set $\mathcal{S} \subset [0,1]^2$. All values $(s_1, s_0) \in \mathcal{S}$ satisfy $s_1 > 1 - s_0$.

Assumption 2A is a weaker form of Assumption 2.2.2, since it is implied by it. Similarly, jointly with Assumption 2.2.3, Assumption 2A implies that $\forall (s_1, s_0) \in \mathcal{S} : P(r = 1) \in (1-s_0, s_1)$. If the condition fails, at least one of the two assumptions is refuted. Compactness of $\mathcal{S}$ is not relevant for identification, but it is utilized in the inference procedure constructed in Section 2.4.

For an element $(s_1, s_0) \in \mathcal{S}$, let the identified set $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ for (θ_1, θ_0) be found using Proposition 2.2.1 or Proposition 2.2.2, depending on which of the discussed assumptions the researcher is willing to maintain. Denote by $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ the corresponding identified set for (θ_1, θ_0) when (s_1, s_0) is known to be in $\mathcal{S}$. All values (θ_1, θ_0) that are found in at least one set G within a collection of sets $G \in \{\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : (s_1, s_0) \in \mathcal{S}\}$ then constitute $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$. In other words, the set $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ contains all values of (θ_1, θ_0) that are consistent with the observed data and at least one $(s_1, s_0) \in \mathcal{S}$. We can formally define:

$$\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S}) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0). \tag{2.10}$$

Corollary 2.2.1. *Suppose that Assumption 2A holds. Let* $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ *be a sharp identified set for* (θ_1, θ_0) *given a value* (s_1, s_0) *as defined in Proposition 2.2.1, or Proposition 2.2.2. Then* $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ *in* (2.10) *is a sharp identified set for* (θ_1, θ_0) *if* $(s_1, s_0) \in \mathcal{S}$.

Any set $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ contains only the values of (θ_1, θ_0) that are consistent with the observed data and (s_1, s_0). The union of sets $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ over all possible $(s_1, s_0) \in \mathcal{S}$ then only contains the values of (θ_1, θ_0) that are consistent with the observed data and at least one $(s_1, s_0) \in \mathcal{S}$. Hence, the identified set $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ is the smallest possible under the maintained assumptions.

The set $\mathcal{S}$ may take different forms. Expected ones include finite sets, line segments or rectangles. In general, within $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ test performance measures θ_1 and θ_0 will no longer necessarily be linearly dependent. The set $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ may not be a line segment in $[0, 1]^2$, but rather a union of line segments of positive and bounded slopes. Hence, it will not be rectangular. It is still possible to demonstrate that r is more precise than t. As in Remark 2.2.2, it is feasible for all $(s_1, s_0) \in \mathcal{S}$ and $(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ to have $(s_1, s_0) > (\theta_1, \theta_0)$ component-wise.

Section 2.3. Bounding Prevalence in Screened Populations

Sensitivity and specificity in the performance study population are often extrapolated to other populations and used to identify different parameters of interest. Notable examples are prevalence in a population undergoing screening and predictive values. In this section, I show how the specific structure of the identified set for (θ_1, θ_0) helps reduce the width of bounds on prevalence when test t is used for screening. Bounds on predictive values are discussed in Appendix B.1.

Population disease prevalence is both a research- and policy-relevant parameter. Suppose that we are interested in learning the true prevalence in a certain population that is being screened. Identification of prevalence based on results of an imperfect screening test is a standard epidemiological problem. Assume that each individual is tested exactly once using only the test t. Maintain that r is not used for screening. Depending on the setting, t may be preferred over r for the purpose due to resource constraints, turnaround time or invasiveness, as explained in Remark 2.2.1. A prominent recent example was the use of antigen testing in university and institutional settings to monitor prevalence during the COVID-19 pandemic.

To make the distinction between the screened and performance study populations explicit, let $Q(t, y)$ denote the probability distribution which generates the data in the screened population. Unlike in the performance study, r is not available, making the use of $Q(t, r, y)$ superfluous. The data alone identify only $Q(t = 1)$. As before, y is not observed, and we are interested in learning $Q(y = 1)$. Let $(\tau_1, \tau_0) = (Q(t = 1|y = 1), Q(t = 0|y = 0))$ be the sensitivity and specificity of t among the screened individuals. We can then write the following identities for the two populations:

$$\begin{aligned} Q(t = 1) &= \tau_1 Q(y = 1) + (1 - \tau_0) Q(y = 0) \\ P(t = 1) &= \theta_1 P_{s_1,s_0}(y = 1) + (1 - \theta_0) P_{s_1,s_0}(y = 0). \end{aligned} \tag{2.11}$$

Remark 2.3.1. It is important to emphasize that $Q(y = 1)$ may differ from $P_{s_1,s_0}(y = 1)$ in the performance study population. In the performance study, one can point identify or bound $P_{s_1,s_0}(y = 1)$ using $P(r = 1)$ and knowledge of (s_1, s_0) or $\mathcal{S}$ as shown in (2.3). In

the context of this section, this is not possible since r is not used for screening, making $Q(r=1)$ unidentified.

Exact knowledge of (τ_1, τ_0) identifies the prevalence (Gart and Buck (1966), Diggle (2011)). In the population of interest it directly follows from (2.11):

$$Q(y=1) = \frac{Q(t=1) + \tau_0 - 1}{\tau_1 + \tau_0 - 1}. \tag{2.12}$$

Walter and Irwig (1988), and Greenland (1996) explain that knowledge of (τ_1, τ_0) is commonly extrapolated from test performance studies. (see also Gastwirth (1987)) That is, researchers maintain the extrapolation assumption: $(\tau_1, \tau_0) = (\theta_1, \theta_0)$, where (θ_1, θ_0) are assumed to be identified in a performance study. I follow this practice and generalize (2.12) to the case when $(\tau_1, \tau_0) = (\theta_1, \theta_0)$, but (θ_1, θ_0) are partially identified. Denote by $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ the identified set for (θ_1, θ_0) obtained in a performance study. $\mathcal{S}$ can be a singleton, as when (s_1, s_0) are assumed to be known, in which case I write $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$.

Assumption 2.3.1. *(Test Performance Extrapolation)* $(\tau_1, \tau_0) = (\theta_1, \theta_0)$.

The assumption maintains that test performance is identical in the performance study and the screened populations. It does not require (θ_1, θ_0) to be known exactly, and it implies that $(\tau_1, \tau_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$. While Walter and Irwig (1988) state that sensitivity and specificity may readily extrapolate to other populations in many cases, one should be aware that credibility of Assumption 2.3.1 critically depends on the details of the empirical setting. One notable potential threat to its validity is variability of test performance across subpopulations, otherwise known as *spectrum effects*. For example, it is known that test sensitivity may vary across subpopulations with different illness severity.

When spectrum effects exist and the two populations differ in terms of relevant subpopulation proportions, Willis (2008a) argues that $(\theta_1, \theta_0) = (\tau_1, \tau_0)$ may be implausible. To see this, observe that in this case (θ_1, θ_0) and (τ_1, τ_0) are weighted averages of sensitivity and specificity across relevant subpopulations, but with different weights. However, if test performance is known for all relevant subpopulations, one could maintain Assumption 2.3.1 and identify prevalence at the subpopulation level using arguments that follow. (see also Mulherin and Miller (2002a))

Proposition 2.3.1. *Suppose that Assumption 2.3.1 holds and the population is screened only using t. Let $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ be a known sharp identified set from Proposition 2.2.1 or Proposition 2.2.2. Denote by θ_j^L and θ_j^U the smallest and largest values of θ_j in $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$. The sharp bounds on prevalence $Q(y=1)$ are:*

$$Q(y=1) \in \Pi_{s_1,s_0} := \left[\min \left\{ \frac{Q(t=1)+\theta_0^L-1}{\theta_1^L+\theta_0^L-1}, \frac{Q(t=1)+\theta_0^U-1}{\theta_1^U+\theta_0^U-1} \right\}, \right. \\ \left. \max \left\{ \frac{Q(t=1)+\theta_0^L-1}{\theta_1^L+\theta_0^L-1}, \frac{Q(t=1)+\theta_0^U-1}{\theta_1^U+\theta_0^U-1} \right\} \right] \cap [0,1] \tag{2.13}$$

when $\forall(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : \ \theta_1 \neq 1 - \theta_0$, and $Q(y=1) \in [0,1]$ otherwise.

Remark 2.3.2. The proof of Proposition 2.3.1 remains valid if we replace Assumption 2.3.1 with a weaker condition $(\tau_1, \tau_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$. The benefits of doing so are primarily technical, as it difficult to think of settings in which the weaker condition is plausible and Assumption 2.3.1 is not.

Proposition 2.3.1 extends the identity (2.12) to the case when (θ_1, θ_0) extrapolate to $Q(t, y)$. Note that it maintains that (s_1, s_0) are known exactly in the performance study.

Corollary 2.3.1 generalizes the results to the case when (s_1, s_0) are known to lie in $\mathcal{S}$. The resulting bounds on the screened population prevalence $Q(y = 1)$ are sharp in the absence of additional data, namely results of other tests such as r.

Whenever there exist $(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : \theta_1 + \theta_0 = 1$, prevalence is unidentified as t may not be informative of y. Such tests are not useful for screening purposes. When t is informative of y, that is $\forall(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : \ \theta_1 \neq 1 - \theta_0$, the importance of the linear structure of $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ for bounding $Q(y = 1)$ becomes apparent. Identifying power of the structure can be substantial as highlighted by Remark 2.3.3. For certain $Q(t, y)$, resulting bounds may even point identify $Q(y = 1)$ using data only on t, despite (θ_1, θ_0) being partially identified.

It is important to highlight that Assumption 2.3.1 is refutable. If t is informative of y, then it is possible that $\Pi_{s_1,s_0} = \emptyset$. By (2.12), that happens if all assumed values (τ_1, τ_0) consistent with the assumption $(\tau_1, \tau_0) = (\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ result in $Q(y = 1) \notin [0, 1]$. This would contradict the definition of $Q(y = 1)$, implying that $(\tau_1, \tau_0) \neq (\theta_1, \theta_0)$.

Remark 2.3.3. Proof of Proposition 2.3.1 reveals that when (s_1, s_0) are known, $Q(y = 1)$ is point identified if $Q(t = 1) = P(t = 1)$. One can intuitively see this from the fact that $Q(t = 1) = P(t = 1)$, $(\tau_1, \tau_0) = (\theta_1, \theta_0)$ and (2.11) jointly imply that $P_{s_1,s_0}(y = 1) = Q(y = 1)$. Since $P_{s_1,s_0}(y = 1)$ is point-identified when (s_1, s_0) are known, then $Q(y = 1)$ is too.

Let $\mathcal{G}_{\theta_j}(s_1, s_0) = \{\theta_j : (\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)\}$ denote the individual bounds on θ_j for $j = 0, 1$. The sets $\mathcal{G}_{\theta_1}(s_1, s_0)$ and $\mathcal{G}_{\theta_0}(s_1, s_0)$ are also referred to as projection bounds on θ_1 and θ_0.

Remark 2.3.4. Let $\forall(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : \ \theta_1 \neq 1 - \theta_0$ so that t is informative for screening. If we were to disregard the linear structure of the sharp identified set by supposing that it is a rectangle $\mathcal{G}_{\theta_1}(s_1, s_0) \times \mathcal{G}_{\theta_0}(s_1, s_0)$, then the bounds on the prevalence would be:

$$Q(y=1) \in \bar{\Pi}_{s_1,s_0} := \left[\frac{Q(t=1)+\theta_0^L-1}{\theta_1^U+\theta_0^L-1}, \frac{Q(t=1)+\theta_0^U-1}{\theta_1^L+\theta_0^U-1}\right] \cap [0,1]. \tag{2.14}$$

It is direct that $\Pi_{s_1,s_0} \subset \bar{\Pi}_{s_1,s_0}$ whenever $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ is not a singleton, so that $\theta_j^U > \theta_j^L$ for $j \in \{0, 1\}$. Disregarding the linear structure of the identified set for (θ_1, θ_0) yields strictly wider bounds on prevalence. For any $Q(t = 1)$, $\bar{\Pi}_{s_1,s_0}$ is an infinite set. If $P(t = 1) = Q(t = 1)$, Π_{s_1,s_0} is a singleton.

Bounds in (2.14) would follow from methods that do not establish the linear structure of $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$, such as Thibodeau (1981) and Emerson et al. (2018). Remark 2.3.4 shows that using such methods to bound test performance will yield wider bounds on prevalence in a population screened by an informative t. Moreover, depending on $P(t, r)$ and $Q(t)$ the difference in width can be extreme, since Π_{s_1,s_0} can be a singleton, while $\bar{\Pi}_{s_1,s_0}$ is always an infinite set. Section 2.5 illustrates how relying on rectangular identified sets affects the width of prevalence bounds using empirical examples. It compares prevalence bound widths implied by estimated identified sets for (θ_1, θ_0) constructed using Thibodeau (1981), Emerson et al. (2018), and the method described here, for hypothetical screened populations with different $Q(t = 1)$.

Corollary 2.3.1. *Suppose that Assumption 2.3.1 holds and the population is screened only using t. Let $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S}) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$, where $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ are known*

sharp identified sets from Proposition 2.2.1 or Proposition 2.2.2. The sharp bounds on prevalence $Q(y=1)$ *are:*

$$Q(y=1) \in \Pi_{\mathcal{S}} := \bigcup_{(s_1,s_0)\in\mathcal{S}} \Pi_{s_1,s_0} \tag{2.15}$$

when $\forall(\theta_1,\theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S}): \ \ \theta_1 \neq 1-\theta_0$, *and* $Q(y=1) \in [0,1]$ *otherwise.*

Corollary 2.3.1 generalizes Proposition 2.3.1 to the case when (s_1, s_0) are not known exactly. If the shape of $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ was disregarded by assuming that the identified set was a rectangle, bounds $\bar{\Pi}_{\mathcal{S}}$ analogous to the ones in (2.14) can still be formed, and it would hold that $\Pi_{\mathcal{S}} \subset \bar{\Pi}_{\mathcal{S}}$.

Throughout this section we have assumed that screening is performed once in the population. If screening is done repeatedly, a time series of prevalence bounds can be constructed. When there is selection into testing, bounds on prevalence by Stoye (2022) may be used, for which the bounds on (θ_1, θ_0) derived in Section 2.2 are natural inputs.

Section 2.4. Estimation and Inference

Identified sets in Section 2.2 can be found when $P(t,r)$ is known. In practice, researchers must use sample data to estimate the identified set and conduct inference. This section demonstrates consistent estimation of the identified set and construction of confidence sets for the points in the identified set that are uniformly consistent in level over a large family of permissible distributions.

Let $W_i = (t_i, r_i) \in \{0,1\}^2$ for $i = 1, \ldots, n$ constitute the observed data of n i.i.d observations from the distribution $P(t,r) \in \mathbf{P}$, where $\mathbf{P}$ is a family of categorical distributions with 4 categories. Let $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ denote an arbitrary identified set for (θ_1, θ_0) given

(s_1, s_0) from any of the propositions above, and $\mathcal{G}_{\theta_j}(s_1, s_0)$ the corresponding identified set for θ_j with $j = 0, 1$. Replacing population parameters with their consistent estimators in closed form expressions for $\mathcal{G}_{\theta_j}(s_1, s_0)$ and $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ yields the consistent plug-in estimator of the identified sets (Manski and Pepper (2000b), Tamer (2010)).

Let $\mathbb{1}\{\cdot\}$ denote the indicator function. Suppose first that (s_1, s_0) are known. $\hat{P}(t = j, r = k) = \frac{\sum_{i=1}^{n} \mathbb{1}\{t_i=j, r_i=k\}}{n}$ are consistent estimators of $P(t = j, r = k)$ for all $(j, k) \in \{0, 1\}^2$. Combining $\hat{P}(t = j, r = k)$ with the knowledge of (s_1, s_0) yields $\hat{P}_{s_1,s_0}(r = k, y = l)$ for every $k, l \in \{0, 1\}^2$. The plug-in estimator $\hat{\mathcal{G}}_{\theta_j}(s_1, s_0)$ for the identified set of a single parameter θ_j follows immediately by inputting $\hat{P}(t = j, r = k)$ and $\hat{P}_{s_1,s_0}(r = k, y = l)$ into the bounds in Proposition 2.2.1, or Proposition 2.2.2. Consistent estimator $\hat{\mathcal{G}}_{(\theta_1,\theta_0)}(s_1, s_0)$ of the joint identified set for (θ_1, θ_0) follows from (2.6).

In the case when (s_1, s_0) are only known to be bounded by some compact set $\mathcal{S}$, one can obtain the consistent estimator $\hat{\mathcal{G}}_{(\theta_1,\theta_0)}(\mathcal{S}) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \hat{\mathcal{G}}_{(\theta_1,\theta_0)}(s_1, s_0)$. This is done by finding a union of $\hat{\mathcal{G}}_{(\theta_1,\theta_0)}(s_1, s_0)$ over a fine grid of (s_1, s_0) covering $\mathcal{S}$. The procedure requires two nested grid-search algorithms, and the level of coarseness of the two grids can impact computation time.

FDA Statistical Guidance on Reporting Results Evaluating Diagnostic Tests requires all diagnostic performance studies to report confidence intervals for θ_1 and θ_0. I show how one can use the method for inference based on moment inequalities from Romano, Shaikh, and Wolf (2014) to form confidence sets that cover the true parameters with at least some pre-specified probability $1 - \alpha$ and that are uniformly consistent over a large family of permissible distributions $\mathbf{P}$.

Let C_n be the confidence set of interest and let $\mathbf{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \Big(\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0) \times \{(s_1,s_0)\}\Big)$ be an identification region for $\theta = (\theta_1,\theta_0,s_1,s_0)$ that depends on $P \in \mathbf{P}$ through $\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)$, and where $\mathcal{S}$ can be a singleton.[12] Note that θ includes reference test performance (s_1,s_0). This is done to facilitate convenient definition of moment inequalities that represent the identified set of interest, regardless of whether (s_1,s_0) are known exactly or not. The confidence set C_n should satisfy:

$$\liminf_{n\longrightarrow\infty} \inf_{P\in\mathbf{P}} \inf_{\theta\in\mathbf{\Theta}(P)} P(\theta \in C_n) \geq (1-\alpha). \tag{2.16}$$

Canay and Shaikh (2017) provide an overview of the recent advances in inference based on moment inequalities that are focused on finding C_n in partially identified models. They underline the importance of uniform consistency of C_n in level in these settings. If it fails, it may be possible to construct a distribution of the data $P(t,r)$ such that for any sample size finite-sample coverage probability of some points in the identified set is arbitrarily low. In that sense, inference based on confidence intervals that are consistent only pointwise may be severely misleading in finite samples. To exploit existing inference methods based on moment inequalities to construct C_n, the identified set $\mathbf{\Theta}(P)$ must be equivalent to some set $\tilde{\mathbf{\Theta}}(P)$:

(2.17)

$$\tilde{\mathbf{\Theta}}(P) = \{\theta \in [0,1]^2 \times \mathcal{S} : E_P\big(m_j(W_i,\theta)\big) \leq 0 \text{ for } j \in J_1 \ , E_P\big(m_j(W_i,\theta)\big) = 0 \text{ for } j \in J_2\}$$

12. More precisely, we are interested in C_n for the points in $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S}) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0) = \{(\theta_1,\theta_0) : \theta \in \mathbf{\Theta}(P)\}$. When $P(t,r)$ is known, whether one defines the identified set as $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$ or $\mathbf{\Theta}(P)$ is inconsequential.

where $m_j(W_i, \theta)$ for $j \in J_1 \cup J_2$ are the components of a random function $m : \{0,1\}^2 \times [0,1]^2 \times \mathcal{S} \to \mathbb{R}^k$ such that $|J_1| + |J_2| = k$. Construction of the confidence set for points in the identified set $\tilde{\mathbf{\Theta}}(P)$ is done by imposing a fine grid over the parameter space $[0,1]^2 \times \mathcal{S}$ for θ and performing test inversion.

Identified sets derived in the previous section are representable by (2.17). Focus on the bounds for θ_1 in Proposition 2.2.2 when the tests have the tendency to wrongly agree for $y = 1$ for intuition. Observe that there are four values that are all lower bounds on θ_1 given (s_1, s_0). Similarly there are four values that are all upper bounds. One of the lower bounds is trivial: $\theta_1 \geq 0$. $\theta_1 \leq \left(\frac{P_{s_1,s_0}(r=0,y=1)}{2} + P_{s_1,s_0}(r=1,y=1) \right) \frac{1}{P_{s_1,s_0}(y=1)} = \frac{1+s_1}{2}$ is one upper bound. There are no parameters pertaining to the population distribution in the bound. This is a restriction on the parameter space, under which $\theta \in \bigcup_{(s_1,s_0)\in\mathcal{S}} [0, \frac{1+s_1}{2}] \times [0,1] \times \{(s_1, s_0)\}$. With the appropriate parameter space, there are six relevant values for the bounds on θ_1 that depend on parameters of the population distribution, three for the upper and three for the lower bound. Hence, we can represent the bounds on θ_1 using six moment inequalities.

Proposition 2.2.2 implies that we only need to include one additional moment equality to represent the joint identification region $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0)$ for (θ_1, θ_0). Then the moment function $\bar{m}^1(W_i, \theta)$ representing the identified set $\mathbf{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \left(\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0) \times \{(s_1, s_0)\} \right)$ will have $k = 7$, where $J_1 = \{1, \ldots, 6\}$ and $J_2 = \{7\}$.

Proposition 2.4.1. *Assume that the index and reference tests have a tendency to wrongly agree only for $y = 1$. Let the moment function $\bar{m}^1$ be:*

$$\bar{m}^1(W_i,\theta) = \begin{pmatrix} \bar{m}^1_1(W_i,\theta) \\ \bar{m}^1_2(W_i,\theta) \\ \bar{m}^1_3(W_i,\theta) \\ \bar{m}^1_4(W_i,\theta) \\ \bar{m}^1_5(W_i,\theta) \\ \bar{m}^1_6(W_i,\theta) \\ \bar{m}^1_7(W_i,\theta) \end{pmatrix} = \begin{pmatrix} (-\theta_1 + s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + (t_i - 1)r_i \\ (-\theta_1 + 1 - s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + (r_i - 1)(1 - t_i) \\ (-\theta_1 + 1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + (t_i - 1) \\ \theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - t_i \\ (\theta_1 - s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - t_i(1 - r_i) \\ \left(\theta_1 + \frac{-1 + s_1}{2}\right)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - t_i r_i \\ (\theta_0 - 1)(1 - \frac{r_i - 1 + s_0}{s_1 - 1 + s_0}) - \theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i \end{pmatrix}. \tag{2.18}$$

Moment inequalities and equalities defined by $\bar{m}^1$ for $J_1 = \{1,\ldots,6\}$ and $J_2 = \{7\}$ represent the joint identification region $\boldsymbol{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \left(\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1,s_0) \times \{(s_1,s_0)\}\right)$ for $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1,s_0)$ defined in Proposition 2.2.2 for $y = 1$. For each $\theta \in \bigcup_{(s_1,s_0)\in\mathcal{S}} [0, \frac{1+s_1}{2}] \times [0,1] \times \{(s_1,s_0)\}$ such that $E_P\big(\bar{m}^1_j(W_i,\theta)\big) \leq 0$ for $j = 1,\ldots,6$ and $E_P\big(\bar{m}^1_7(W_i,\theta)\big) = 0$, it must be that $\theta \in \boldsymbol{\Theta}(P)$. Conversely, if $\theta \in \boldsymbol{\Theta}(P)$, then $E_P\big(\bar{m}^1_j(W_i,\theta)\big) \leq 0$ for $j = 1,\ldots,6$ and $E_P\big(\bar{m}^1_7(W_i,\theta)\big) = 0$.

Similarly, it is possible to define moment inequality functions that represent remaining identified sets in Propositions 2.2.1 and 2.2.2. They are found in equations (B.5), (B.7), and (B.8) in Appendix B.2.

Romano, Shaikh, and Wolf (2014), Theorem 3.1 provides sufficient conditions for uniform consistency of confidence sets over a large family of distributions. Assumption 2.4.1 defines a family $\mathbf{P}$ to which the conclusions of Theorem 3.1 apply. This is demonstrated by Theorem 2.4.1 below.

Assumption 2.4.1. *There exists a number* $\varepsilon > 0$ *such that* $P(t = j, r = k) \geq \varepsilon$ *for all* $(j, k) \in \{0, 1\}^2$ *and any* $P(t, r) \in \mathbf{P}$.

The assumption restricts $\mathbf{P}$ to distributions $P(t, r)$ such that all outcomes $(t, r) \in \{0, 1\}^2$ have probability that is bounded away from zero. It serves a technical purpose, ensuring that the uniform integrability condition required by Romano, Shaikh, and Wolf (2014), Theorem 3.1 holds. The assumption is easily interpretable and it appears reasonable in the analyzed data, as discussed in Section 2.5.

Theorem 2.4.1. *Suppose that Assumptions 2.2.1, 2A, 2.2.3, and 2.4.1 hold. Then for any component* $m_j(W_i, \theta)$ *in* (2.18), (B.5), (B.7), *and* (B.8):

(1) $Var_P(m_j(W_i, \theta)) > 0$ *and for all* $P \in \mathbf{P}$ *and* $\theta \in [0, 1]^2 \times \mathcal{S}$;

(2) $\limsup_{\lambda \longrightarrow \infty} \sup_{P \in \mathbf{P}} \sup_{\theta \in \mathbf{\Theta}(P)} E_P \left[\left(\frac{m_j(W_i, \theta) - \mu_j(\theta, P)}{\sigma_j(\theta, P)} \right)^2 \mathbb{1} \left\{ \left| \frac{m_j(W_i, \theta) - \mu_j(\theta, P)}{\sigma_j(\theta, P)} \right| > \lambda \right\} \right] = 0$;

where $\mu_j(\theta, P) = E_P(m_j(W_i, \theta))$ *and* $\sigma_j(\theta, P) = Var_P(m_j(W_i, \theta))$.

Theorem 2.4.1 enables us to use the method from Romano, Shaikh, and Wolf (2014) to construct confidence sets C_n for points (θ_1, θ_0) in the identified sets defined by Proposition 2.2.1 and Proposition 2.2.2 that satisfy (2.16) when the relevant family of population distributions conforms to Assumption 2.4.1.

Remark 2.4.1. In certain cases researchers may estimate (s_1, s_0) based on an independent sample of size m, rather than assume them to be known. For example, this happens if one has access to a sample from an independent validation study identifying performance of r. It is possible to account for statistical imprecision of both samples using Šidák's correction for independent tests (Lehmann and Romano (2022), Chapter 9.1.2).

Let $\alpha_S = 1 - (1-\alpha)^{\frac{1}{2}}$. One can construct an asymptotic confidence set for (s_1, s_0) at the $1-\alpha_S$ confidence level, and treat it as $\mathcal{S}$ in the inference procedure. Then, the confidence set C_n at the significance level α_S ensures at least $1-\alpha$ asymptotic coverage of (θ_1, θ_0) as $n, m \to \infty$. For example, if $\alpha = 5\%$, then $\alpha_S = 2.53\%$ which modestly improves upon the Bonferroni correction.

Section 2.5. Application - Abbott BinaxNOW COVID-19 Antigen Test

In this section, I apply the developed method to existing study data to provide confidence and estimated identified sets for (θ_1, θ_0) of the rapid antigen COVID-19 test with the currently highest market share in the United States - *Abbott BinaxNOW COVID-19 Ag2 CARD* test.

Template for Developers of Antigen Tests required by the FDA for EUA mandates that the reference for all COVID-19 antigen test studies must be an approved RT-PCR test.[13] However, Arevalo-Rodriguez et al. (2020), Kucirka et al. (2020), Dramé et al. (2020), Hernández-Huerta et al. (2020), and Kanji et al. (2021) explain that these tests are imperfectly sensitive. Using them as a reference yields "apparent" and generally not true sensitivity and specificity. Interpreting the results as measures of true performance may be severely misleading. Fitzpatrick et al. (2021) emphasize that the false negative rate of the *BinaxNOW* test may be substantially understated by the reported "apparent" analog due to imperfect reference tests. They highlight that "apparent" measures can unjustifiably lead the users to believe that the test must have high sensitivity, even when this is not true.

13. Link: https://www.fda.gov/media/137907/download (Last accessed: 12/25/2022)

To verify the claim, I revisit the test performance study data from the submitted EUA documentation[14], as well data from an independent study by Shah et al. (2021). I compare the results with the corresponding "apparent" estimates from the original documentation and the instructions for use pamphlet. I also use the data to compare the developed method with existing bounds by Thibodeau (1981) and Emerson et al. (2018), henceforth referred to as comparable methods.

By established notation, t is the antigen test, r is the RT-PCR test and y determines whether the person truly has COVID-19. To construct the confidence sets, I implement the test from Romano, Shaikh, and Wolf (2014) denoted by ϕ_n^{RSW2} in Bai, Santos, and Shaikh (2021). The test relies on the maximum statistic $T_n = \max\left\{\max_{1\leq j\leq k} \frac{\sqrt{n}\bar{m}_j}{S_j}, 0\right\}$, where $\bar{m}_j = \frac{1}{n}\sum_{i=1}^n m_j(W_i,\theta)$ and $S_j^2 = \frac{1}{n}\sum_{i=1}^n (m_j(W_i,\theta) - \bar{m}_j)^2$ for a value θ and components of the appropriate moment function $m_j(W_i,\theta)$ $j = 1,\ldots,k$. The testing procedure has two steps: 1) Construction of confidence regions for the moments; 2) Formation of a critical value incorporating information on which moment inequalities are "negative". I perform test inversion over a fine grid of 10^5 points for the relevant parameter space for (θ_1,θ_0), and additionally over 10 points over $\mathcal{S}$, where applicable. Following the original paper, I use 500 bootstrap samples to find the critical values and set $\beta = \alpha/10$. The results do not change significantly with alternative values $\beta = \alpha/5$ and $\beta = \alpha/20$.

Subsection 2.5.1. Identification Assumptions

Use of comparable and bounds developed in the paper requires a credible set of values $(s_1, s_0) \in \mathcal{S}$ for the reference RT-PCR test. I maintain that $s_0 = 1$ following Kucirka

14. Link: https://www.fda.gov/media/141570/download (Last accessed: 12/25/2022)

et al. (2020) who do the same, citing perfect analytical specificity.[15] The same assumption has been used in other existing work, such as Manski (2020), Manski and Molinari (2021) Kanji et al. (2021), Ziegler (2021), and Stoye (2022).

In the absence of a perfect gold standard, a conventional diagnostic test performance study cannot identify sensitivity s_1 of the RT-PCR tests. Some studies rely on a different set of assumptions to identify the parameter of interest. Kanji et al. (2021) provide a discordant result analysis of the RT-PCR test used for frontline testing of symptomatic individuals. The authors define discordant results as initially negative RT-PCR findings followed by a positive test result within the incubation period. The negative samples were retested by three alternative RT-PCR assays targeting different genes. If at least two alternative assays yielded positive results, the initial result was considered to be a false negative finding. Assuming perfect specificities of each assay, and perfect sensitivity of the combined testing procedure, they estimate the sensitivity of the used RT-PCR test at 90.3%. Perfect specificity is maintained based on perfect analytical specificity. Arevalo-Rodriguez et al. (2020) use data from published studies to estimate sensitivity, defining false negatives as patients who were symptomatic and negative, but subsequently positive on an RT-PCR test within the incubation period. It is implicitly maintained that all initial results must have been false negatives. Three estimates are based on data from the United States. Sensitivity of 90% is the only estimate which is not considered to be at high risk of bias according to the QUADAS-2 tool (Whiting et al. (2011)). Following the two references, I assume that $s_1 = 0.9$. One should note that assumed (s_1, s_0) are a

15. Specificity on contrived laboratory samples containing other pathogens, but not SARS-CoV-2.

critical identifying assumption which directly affect the obtained estimates. Appendix B.3 thus discusses robustness of findings to different assumed values of s_1.

Since the antigen and RT-PCR test rely on the same type of sample, I maintain that they have a tendency to wrongly agree for $y = 1$. Since it is assumed that $s_0 = 1$, the tendency to wrongly agree for $y = 0$ has no identifying power, as explained by Remark 2.2.6. It is thus not maintained.

Subsection 2.5.2. Data and Results

EUA documentation and the instructions for use outline the initial performance study. The estimates were obtained on a sample of 460 participants tested within 7 days of symptom onset. Shah et al. (2021) perform the same analysis on an independent sample of 2110 individuals enrolled at a community testing site. This includes 1188 symptomatic individuals, of which 929 were tested within 7 days of symptom onset. I omit symptomatic individuals tested more than 7 days after initial symptoms for comparability with the EUA study. I separately analyze the performance on 877 asymptomatic participants to provide plausible estimates of performance in the absence of symptoms. The data are summarized in Table 2.1. In all three samples, estimates of joint probabilities $\hat{P}(t = j, r = k)$ for $(r, k) \in \{0, 1\}^2$ are bounded away from zero. That the distributions which have generated the data lie in a family of distributions satisfying Assumption 2.4.1 is reasonable.

Panels (a), (b), and (c) of Figure 2.1 represent the estimated "apparent" operating characteristics and joint identified sets for (θ_1, θ_0), as well as corresponding 95% confidence sets. The joint confidence set for apparent measures is the projection Clopper-Pearson exact confidence set. The areas of the two confidence sets are similar. Table 2.2 shows the

Table 2.1. Study Data.

Data	N	(t_i, r_i)			
		$(1,1)$	$(0,1)$	$(1,0)$	$(0,0)$
EUA Symptomatic	460	99	18	5	338
Shah et al. (2021) Symptomatic	929	199	44	2	684
Shah et al. (2021) Asymptomatic	877	33	15	5	824

Note: Number of outcomes (t_i, r_i) in analyzed studies.

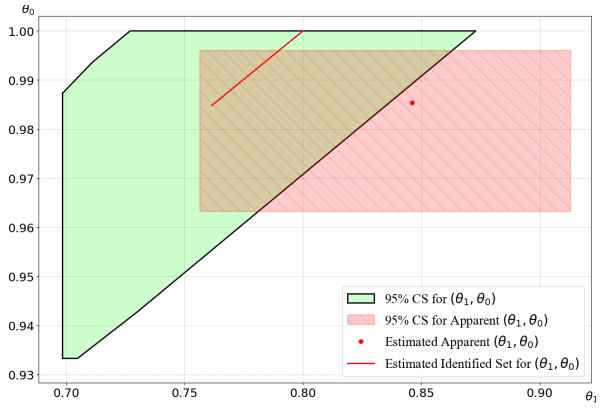
estimates of projected individual bounds on sensitivity and specificity. The bounds are revealing.

Table 2.2. Estimates and Estimated Projection Bounds.

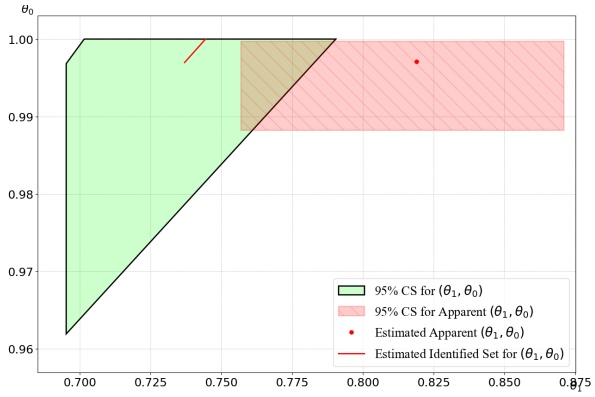
		Data		
		EUA Sx	Shah et al. Sx	Shah et al. ASx
θ_1 Estimates	Apparent	0.846	0.819	0.688
	Projection	$[0.761, 0.800]$	$[0.737, 0.744]$	$[0.619, 0.669]$
	Emerson et al. (2018)	$[0.761, 0.800]$	$[0.737, 0.744]$	$[0.619, 0.712]$
	Thibodeau (1981)	$[0.761, 0.846]$	$[0.737, 0.819]$	$[0.619, 0.688]$
θ_0 Estimates	Apparent	0.985	0.997	0.994
	Projection	$[0.985, 1.000]$	$[0.997, 1.000]$	$[0.994, 0.997]$
	Emerson et al. (2018)	$[0.985, 1.000]$	$[0.997, 1.000]$	$[0.994, 1.000]$
	Thibodeau (1981)	$[0.985, 1.000]$	$[0.997, 1.000]$	$[0.994, 0.998]$

Note: Estimates of "apparent" performance measures and projections of estimated identified sets for (θ_1, θ_0) shown in Figure 2.1. Sx denotes the symptomatic, and ASx the asymptomatic individuals.

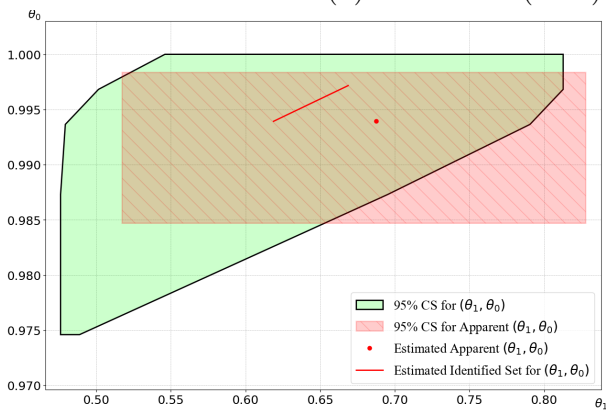
The original EUA was granted based on interim results of the study in which the test exhibited estimated "apparent" sensitivity and specificity of $(91.7\%, 100\%)$. Subsequent results of the full study yielded "apparent" operating characteristics estimates of $(84.6\%, 98.5\%)$. Public statements and media releases erroneously cite all of the estimates



(a) EUA: Results



(b) Shah et al. (2021) Symptomatic: Results

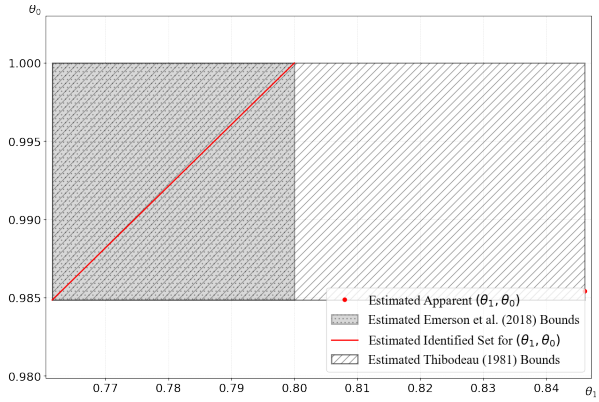


(c) Shah et al. (2021) Asymptomatic: Results

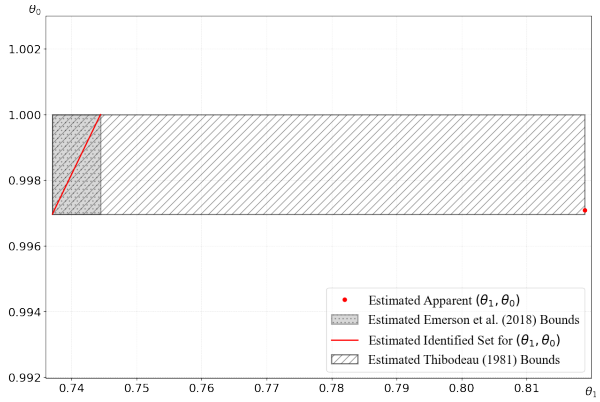
Figure 2.1. Estimates, and 95% confidence sets for "apparent" measures and points in the identified set for (θ_1, θ_0).

as estimates of true performance.[16] Both the interim and final estimates are reported on the instructions-for-use pamphlet accompanying the test. First two rows of Table 2.2 show that both estimates of "apparent" sensitivity lie strictly above the estimated projected upper bound for true sensitivity in all samples. Hence, sensitivity may be overstated by the "apparent" analog, as Fitzpatrick et al. (2021) suggest. The fifth and sixth rows

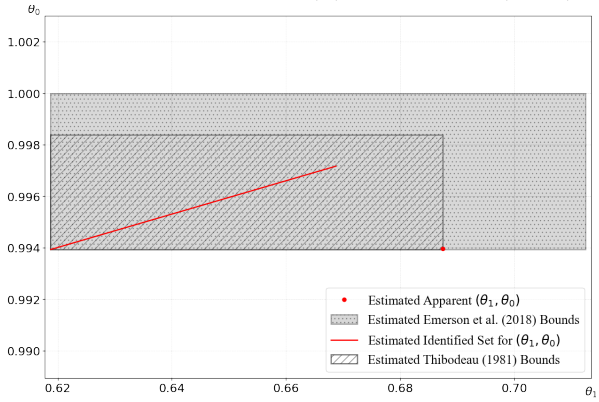
16. For example: https://www.bloomberg.com/press-releases/2020-12-16/abbott-s-binaxnow-covid-19-rapid-test-receives-fda-emergency-use-authorization-for-first-virtually-guided-at-home-rapid-test-u.



(a) EUA: Methods Comparison



(b) Shah et al. (2021) Sx: Methods Comparison

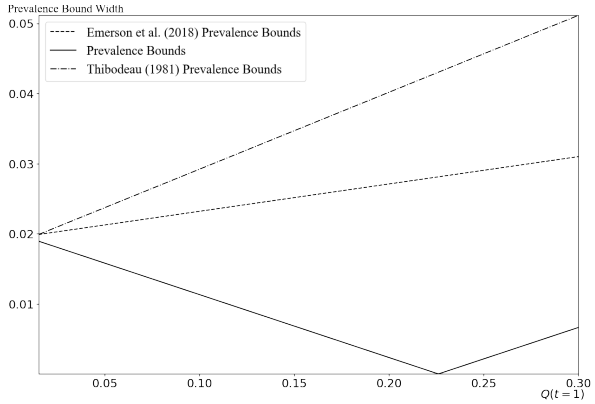


(c) Shah et al. (2021) ASx: Methods Comparison

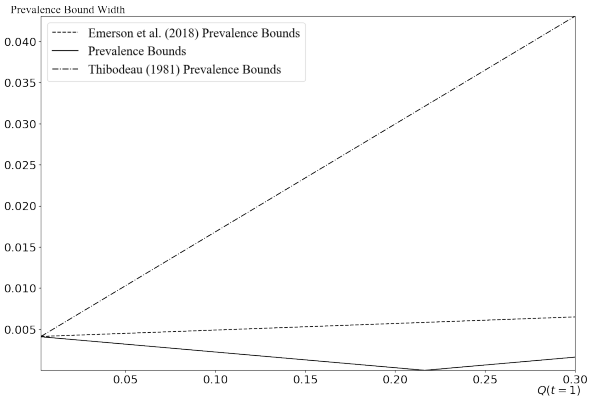
Figure 2.2. Comparison of estimated identified sets with estimates by comparable methods.

demonstrate that final estimates of "apparent" specificity are at the estimated projected lower bounds for true specificity. True specificity may be understated.

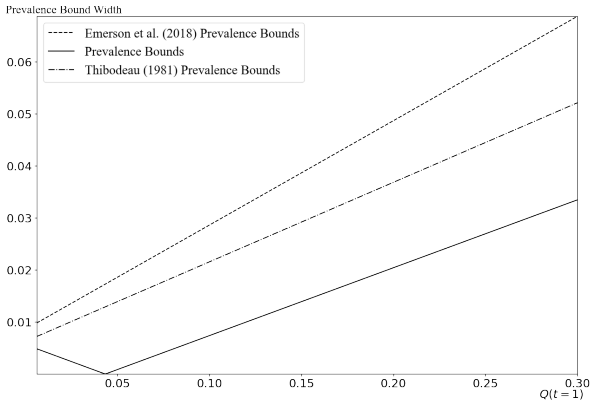
In Figure 2.1: (a), the estimate of "apparent" measures is outside the confidence set for (θ_1, θ_0). At the 5% significance level the hypothesis $H_0 : (\theta_1, \theta_0) = (84.6\%, 98.5\%)$ would be rejected. In other words, under the assumptions, the true sensitivity and specificity are not jointly equal to currently often cited "apparent" values $(84.6\%, 98.5\%)$ at the ubiquitous



(a) EUA: Prevalence Bounds Width



(b) Shah et al. (2021) Sx: Prevalence Bounds Width



(c) Shah et al. (2021) ASx: Prevalence Bounds Width

Figure 2.3. Comparison of prevalence bounds widths implied by the estimated identified sets.

level of significance. The argument for the same value holds in all other samples, as well as for the interim "apparent" estimates $(91.7\%, 100\%)$.

Estimated bounds on false negative rate for symptomatic individuals within 7 days of symptom onset are $[20\%, 23.9\%]$ in the final EUA study data, which is between 1.3 and 1.55 times larger than the corresponding "apparent" estimate of 15.4%. Comparison with the often-cited interim estimate of 8.3% reveals that the estimated true false negative

rate is between 2.41 and 2.88 times larger than the "apparent" analog. Data from Shah et al. (2021) yield estimated bounds of $[25.6\%, 26.3\%]$ for symptomatic and $[33.1\%, 38.1\%]$ for asymptomatic individuals. These estimates suggest that the true false negative rates may be up to 3.17 and 4.59 times higher than "apparent" interim analogs for symptomatic and asymptomatic individuals, respectively. Appendix Appendix B.3 shows that assuming lower s_1 further exacerbates the difference between true and "apparent" false negative rates. It also notes that assuming any $0.9 < s_1 < 1$ lessens the difference. However, the apparent measure is never contained by the set. Hence for any feasible s_1, apparent sensitivity overestimates true sensitivity.

Remark 2.5.1. Estimated average number of infected symptomatic people who are missed by the antigen test is up to 3.17 times higher than the test users may be led to believe by reported "apparent" estimates.

Hadgu (1999) highlights that the errors in measurement of 2.9 percentage points for sensitivity are significant. The differences I find in this paper between the estimates of "apparent" and true sensitivity are substantially larger under plausible assumptions. The differences vary between 4.6 and 8.5 percentage points using the final EUA study data. Results from Shah et al. (2021) exacerbate the discrepancies when compared to the final EUA study "apparent" sensitivity to as much 10.9 percentage points in the symptomatic population and 22.7 percentage points in the asymptomatic population. Even though the estimates of specificity remain close to the estimates of "apparent" specificity, the findings for sensitivity warrant further attention.

Remark 2.5.2. FDA has granted EUA to tests demonstrating at least 80% estimated sensitivity. The results show that, depending on interpretation and assumed (s_1, s_0), the test may not satisfy the requirement.

Panels (a), (b), and (c) of Figure 2.2 show estimates of the identified set for (θ_1, θ_0) and compare them with estimates obtained using comparable methods. Results are represented graphically in order preserve the specific linear structure of the identified set that is lost through projection. The sharp identified set provides a substantial reduction in size in all three samples over the comparable methods, and can be very informative. Estimates of the identified set do not contain the estimates of "apparent" measures in any of the samples. Owing to the lack of sharpness, bounds estimated using other methods do not necessarily exclude the "apparent" measures. Table 2.2 shows that projected bounds on θ_1 and θ_0 can also be proper subsets of those produced by comparable methods.

Remark 2.5.3. Projected bounds on θ_j from Emerson et al. (2018) are equivalent to projected bounds from Proposition 2.2.1 without imposing the tendency to wrongly agree for any y. Rows of Table 2.2 marked by "*Projection*" and "Emerson et al. (2018)" thus correspond to projection bounds with and without assuming the tendency to wrongly agree for $y = 1$, respectively. As mentioned by Remark 2.2.6, the assumption may have identifying power depending on $P(t, r)$ and (s_1, s_0). For data in the first two columns, estimates suggest that the assumption has no identifying power. However, among Shah et al. (2021) ASx individuals, it effectively halves the size of the estimated identified set.

Panels (a), (b), and (c) of Figure 2.3 depict the width of prevalence bounds implied by estimates from the three methods when extrapolated to populations screened by the

antigen test. The solid line represents bound width given estimates of the identified set for (θ_1, θ_0) and (2.13) for various hypothetical values of $Q(t = 1)$. As previously highlighted, for $Q(t = 1) = P(t = 1)$ prevalence becomes point-identified, despite (θ_1, θ_0) being only partially identified. The remaining lines refer to widths of bounds in (2.14), following from estimates obtained by comparable methods which yield rectangular bounds on (θ_1, θ_0). The resulting sharp bounds on prevalence are always proper subsets of bounds found via the other two methods. Benefits stemming from the particular shape of the sharp joint identified set are immediate. Even when the projected bounds are not strictly narrower compared to other methods, the identified set can yield substantially narrower bounds on derived parameters, as shown by Figure 2.3:(a), (b).

Section 2.6. Applications Beyond Diagnostic Test Performance

Derived results have applications that extend beyond diagnostic test performance studies. This section offers three illustrative examples, highlighting further utility of the bounds on (θ_1, θ_0) and $Q(y = 1)$. It also interprets the tendency to wrongly agree in the relevant contexts, and contrasts it with the exclusion restrictions $E[t|r, y] = E[t|r]$ and $E[t|r, y] = E[t|y]$ from Cross and Manski (2002).[17]

In abstraction, suppose $P(t, r)$ is identified and $P(r|y)$ can be identified or credibly bounded, for $(t, r, y) \in \{0, 1\}^3$. A salient case is the one in which a validation dataset identifying $P(r|y)$ exists, but it cannot be matched with the dataset identifying $P(t, r)$. This may happen due to legal or privacy concerns, lack of adequate identifiers, or because the two datasets are independent. Sharp bounds on $P(t, y)$ and its features follow from

17. Note that $E[t|r, y] = E[t|y]$ is equivalent to $t \perp\!\!\!\perp r|y$, which is frequently considered implausible in the context of diagnostic tests, as discussed by Section 2.2.1.

the results. If (θ_1, θ_0) are bounded and they extrapolate to other populations where data only on t is available, one can also sharply bound $Q(y = 1)$.

Example 2.6.1. (Surveys and Validation Data) Suppose y is measured using a survey response r, and t is a binary outcome of interest for surveyed individuals. This identifies $P(t, r)$. It is well known that survey responses are susceptible to misclassification errors. This has been discussed in the contexts where y is participation in government welfare programs, disability, or employment status (Poterba and Summers (1986), Kreider and Pepper (2007)). For example, let y be true participation in the Food Stamp program, r self-reported participation, and let t denote whether a person has completed higher education. Bollinger and David (1997) and Meyer, Mittag, and Goerge (2022) identify $P(r|y)$ using administrative validation data for the American Community Survey, the Current Population Survey and the Survey of Income and Program Participation. One can then bound functions of $E[t|y]$ in these surveys.

Researchers may find the tendency to wrongly agree for $y = j$ credible, maintaining it to further tighten the bounds. For $y = 1$ in the example above, it would mean that people who falsely report no participation are more likely not to have a higher education degree than to have one. The restriction $E[t|r, y] = E[t|r]$ holds if rates of higher education depend on program participation only through the survey response. That is, for people who provide a response r, higher education rates must not change with true program participation. Conversely, $E[t|r, y] = E[t|y]$ holds if for individuals who have true participation y, higher education rates must be the same among those who gave correct and incorrect answers. Hence, one may consider the tendency to wrongly agree to be more plausible in this context.

Results in Section 2.3 could be useful if one wishes to learn $Q(y = 1)$ in a different population to which $E[t|y]$ extrapolates, and for which only t is available. In the context of the example, if access to a dataset containing local rates of higher education is available, then local Food Stamp program participation rates can be sharply bounded using knowledge of $E[t|y]$ from the survey.

Example 2.6.2. (Protected Classes and Privacy) Let t be an outcome of interest and y a protected class. Administrative data often do not contain y, but its proxy r may be available, identifying $P(t, r)$. Some commonly used proxies are constructed based on datasets in which both r and y are observed, so their performance may be known. For a t of interest, we can then sharply bound various parameters of $P(t, y)$. For example, y may be a latent binary indicator for a certain race, and its proxy r may be constructed using the Bayesian Improved Surname Geocoding (BISG) method. Performance of BISG has been validated in several datasets, potentially providing information on $P(r|y)$ (Elliott et al. (2008), Imai and Khanna (2016)). Elzayn et al. (2023) consider a similar setting where t is a tax audit flag, and seek to identify racial tax audit disparity $E[t|y = 1] - E[t|y = 0] = \theta_1 + \theta_0 - 1$ in a dataset where only r is available.[18] This parameter can be bounded using $P(r|y)$. One could also use results from Section 2.3 to sharply bound racial composition $Q(y = 1)$ in a population with the same $E[t|y]$ if a dataset containing only tax audit rates is available.

The tendency to wrongly agree for $y = 0$ would mean that people misclassified as being of race $r = 1$ are more likely to be audited than not to be audited. This would be plausible if characteristics of individuals with $y = 0$ that lead to racial misclassification also make audit the more likely outcome. The assumption $E[t|r, y] = E[t|r]$ maintains

18. The parameter $\theta_1 + \theta_0 - 1$ is known as the Youden's J statistic in the medical literature.

that tax audit rates for people classified as r would not vary with their true race. This would hold if audit decisions depend on race only through information summarized by r. Conversely, $E[t|r, y] = E[t|y]$ would hold if for people of race y audit rates do not vary with their classified race r. This would be the case if audit decisions depend on information summarized by r only through true race. If any of these assumptions are plausible, researchers may use them to obtain tighter bounds.

The same arguments apply to measurement of racial disparities in health care, where t can be, for example, medication nonadherence. Weissman and Hasnain-Wynia (2011) explain that race is often missing from medical claims data and studies validating performance of BISG in such datasets are available (Adjaye-Gbewonyo et al. (2014)).

Example 2.6.3. (Binary Classifiers) Let t be a binary classifier whose performance is determined using an imperfect binary classifier or label r as a reference. The discussion herein readily applies to any such setting. In general, r may be imperfect when determined by labelers or algorithms (Cannings, Fan, and Samworth (2020)). For example, labels r are often obtained through services like the Amazon Mechanical Turk. Mislabeling may happen due to human error, inattentive labelers, or malicious mislabelling activity. If researchers are be able to determine misclassification rates (s_1, s_0), then (θ_1, θ_0) may be bounded. Foody (2010) notes that (θ_1, θ_0) are often of interest in the context of remote sensing applications, such as satellite imaging. However, reference data r are commonly imperfect. Carlotto (2009) explains that in some cases, it may be possible to learn (s_1, s_0) by observing the ground truth y in validation studies. When the performance of r cannot be validated, one can use the bounds to perform sensitivity analyses and determine how much apparent and real performance may differ for various possible values of (s_1, s_0).

Binary diagnostic tests are specific examples of binary classifiers, and the interpretation of the tendency to wrongly agree remains unchanged for general t and r. Its plausibility can be argued based on the properties of the classifiers. Foody (2023) notes that the exclusion restriction $E[t|r, y] = E[t|y]$ may also commonly be implausible in the context of binary classifiers, especially when t and r are based on the same phenomenon or process. The restriction $E[t|r, y] = E[t|r]$ asserts that $t \perp\!\!\!\perp y|r$, or that t cannot provide additional information about y over r, which may be unappealing depending on the context.

Results from Section 2.3 also apply directly. When using the classifier t to determine the prevalence of y in a screening study, one can use them to obtain sharp bounds on $Q(y = 1)$ if the performance of t extrapolates to the screened population.

Section 2.7. Concluding Remarks

This paper derives the smallest possible identified set for sensitivity and specificity of a diagnostic test of interest in standard settings, when the reference test is imperfect. It formalizes an existing assumption on dependence between the reference and the test of interest, and shows how it can further reduce the size of the identified set. Finally, it develops an appropriate uniform inference procedure for the points in the identified set, enabling construction of confidence sets. The study also indicates applicability of the method beyond the context of diagnostic test performance studies.

The framework is proposed as a solution to a ubiquitous problem in diagnostic test performance studies, and it can be directly applied to existing study data to bound true test performance. Doing so demonstrates that a widely used COVID-19 antigen test tends to produce significantly more false negative results than what the currently cited figures

suggest. Since other rapid COVID-19 antigen tests may exhibit similar tendencies, these findings warrant further investigation.

CHAPTER 3

Diagnostic Tests as Dilations

Section 3.1. Introduction

Diagnostic tests serve to provide information about a patient's health status. Diagnostic and treatment decisions are thus frequently made using a *post-test* probability—the probability that a patient has a suspected illness conditional on an observed test result—which follows from Bayesian updating (McNeil, Keeler, and Adelstein 1975, Watson, Whiting, and Brush 2020a). The clinician first forms a *pre-test* probability that the patient has the illness, relying on heuristics and expert knowledge. They then observe a test result and use Bayes' theorem to determine the corresponding *post-test* probability based on the test's false positive and false negative rates.[1]

The post-test probability is a central parameter in clinical decision-making. Its use rests on the presumption that conditioning on the test result indeed provides additional information about the patient's health status. However, Good (1974) demonstrates that conditioning on *any* event from a partition can, in fact, *only increase* uncertainty. In this context, it may be that the set of possible post-test probability values following *any* test result strictly includes the pre-test probability. Intuitively, conditioning on any test result then only *dilates* the pre-test probability; the post-test probability *only* introduces

1. In practice, updating is commonly done via online and graphical tools, Fagan's nomograms (Fagan 1975). Examples of online calculators are Link 1 and Link 2. See Section C.2 for nomograms.

uncertainty about the patient's health status relative to the pre-test probability. Seidenfeld and Wasserman (1993) call this phenomenon *dilation*.[2]

Dilation has generated considerable attention in the literature. However, it has traditionally been viewed as an abstract theoretical concept. We show that diagnostic tests may induce dilation in real-world contexts under existing clinical decision-making practices.

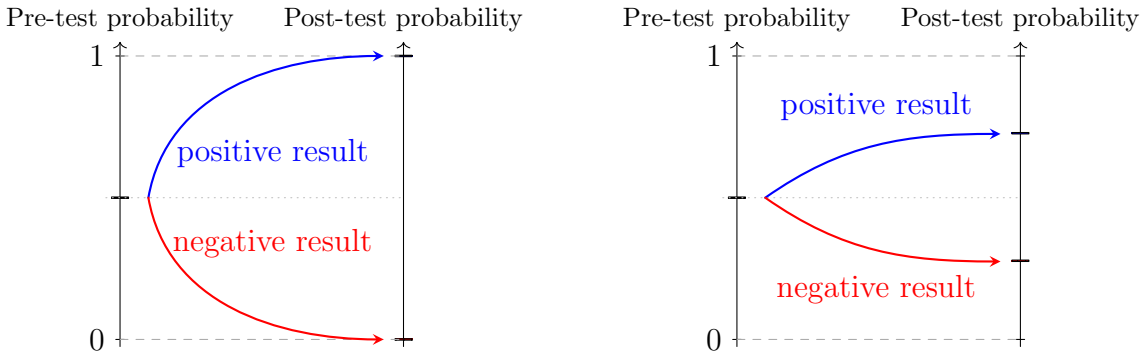
This paper has three main contributions. First, we show that dilation may arise in ubiquitous clinical settings—when clinicians update the pre-test probability using a diagnostic test whose misclassification rates are evaluated against an imperfect reference. Even though dilation-inducing tests are uninformative of the patient's health status, conventional performance measures may not reflect this, leading policy-makers and researchers to adopt or recommend them. Yet, regulators and clinicians alike might reasonably wish to avoid such tests since virtually any test carries (unmodeled) testing cost.[3] This motivates our second contribution. We enable decision-makers to identify dilation-inducing diagnostic tests by equivalently characterizing when dilation occurs in this context, and providing a statistical testing procedure that is uniformly consistent in level for large family of relevant data-generating processes (DGPs). Finally, we study computed tomography (CT) chest scans for detecting COVID-19 infection that were recommended as a primary detection tool in epidemic areas based on conventional performance measures. We find that they

2. They briefly mention a related example: "*To emphasize the counterintuitive nature of dilation, imagine that a physician tells you that you have probability* $\frac{1}{2}$ *of having a fatal disease. He then informs you that he will carry out a blood test tomorrow. Regardless of the outcome of the test, if he conditions on the new evidence, he will then have a lower probability of 0 and an upper probability of 1 that you have the disease. Should you allow the test to be performed?*" Levi (1977) and Kyburg (1977) debate whether one may retain the more informative unconditional probability after conditioning.

3. Moreover, Grunwald and Halpern (2004) argue that ignoring costless dilation-inducing information may be *minimax*-optimal in certain settings; Shishkin and Ortoleva (2023) explain that ambiguity-averse decision makers may be strictly worse off after updating based on any result of such a test.

induced dilation and were thus only increasing uncertainty about the patient's health status.

To further illustrate dilation, consider the updating from a given pre-test to a post-test probability in the context of clinical decision-making. If the test is perfect, the post-test probability would be 0 in the case of a negative result or 1 in the case of a positive result. This is depicted in Figure 3.1a. In practice, the test is almost always imperfect. If the misclassification rates are point identified, the post-test probabilities may not be 0 or 1, but may still be informative, as in Figure 3.1b. This can be seen since the pre-test probability is shifted upwards (downwards) due to a positive (negative) test result. Importantly, note that the post-test probability is a unique value in either case.



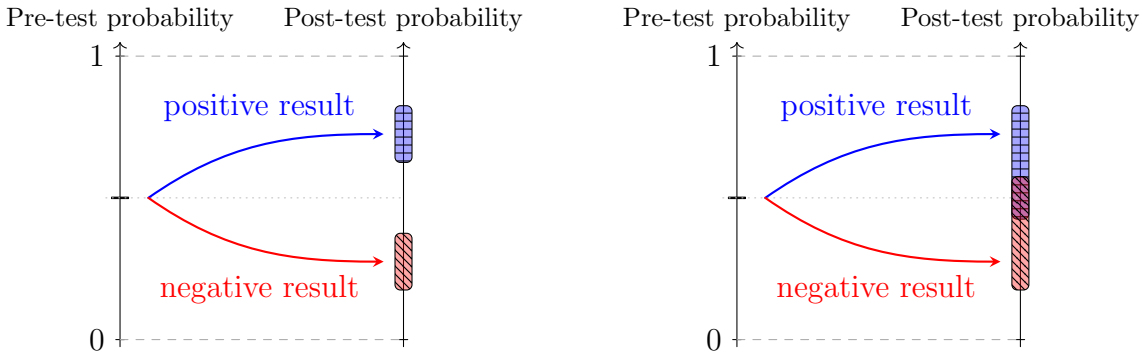
(a) Perfect information (b) Imperfect, but unambiguous information

Figure 3.1. Unambiguous information

Whenever the pre-test probability or the false positive/negative rates can only be bounded, the resulting post-test probability is no longer unique for any test result. Then the clinicians commonly determine the corresponding bounds on the post-test probability via full Bayesian updating (see, e.g., Baron 1994; Bianchi, Alexander, and Cash 2009; Srinivasan, Westover, and Bianchi 2012; and Manski 2021). If the misclassification rates

of the diagnostic test are uncertain, we say that the test provides *ambiguous* information. However, even when the test provides ambiguous information, it can be informative of the health status if the post-test probability for at least one test result is shifted with respect to the pre-test probability. This is depicted in Figure 3.2a.

In certain cases, the test providing ambiguous information may not be informative: the pre-test probability can be included in the set of post-test probabilities, *regardless of the test result* and *dilation* occurs. Figure 3.2b exemplifies such a situation.



(a) Ambiguous information (b) Dilation

Figure 3.2. Ambiguous information

Partially identified misclassification rates are necessary for dilation to occur. The rates are generally evaluated with respect to a reference test since the true health status is unobservable. Reference tests are almost always imperfect, typically resulting in partially identified misclassification rates under tenable assumptions in practice (Thibodeau 1981; Emerson et al. 2018).

However, regulators and researchers usually rely on point-identified "apparent" misclassification rates—misclassification rates with respect to the reference—to approve or recommend tests. It has been well documented that "apparent" and true misclassification

rates can differ significantly (Deneef 1987; Boyko, Alderman, and Baron 1988; Valenstein 1990). Nevertheless, test evaluation still centers on "apparent" rates. We speak to this practice by demonstrating that tests may appear precise with respect to the reference, while simultaneously inducing dilation.

This paper enables policy-makers and researchers to detect whether a test is dilation-inducing. We first equivalently characterize when a test induces dilation in the context of clinical decision-making. Intuitively, dilation occurs if and only if the set of true positive and true negative rates consistent with the data and assumptions, *the identified set*, contains at least two values—one that would update the pre-test probability and one that would not affect it. By Bayes' theorem, this occurs if and only if the sum of the true positive and true negative rates can be both equal and not equal to one.

When the partially identified rates occur due to imperfect reference, as it is commonly the case, we provide a simple equivalent characterization for the test to be dilation-inducing, relying on results from Obradović (2024). In a large class of DGPs that are practically relevant, the identified set for true positive and true negative rates is a line segment with positive and finite slope under standard assumptions. The set intersects the antidiagonal of the unit square if and only if the sum of the true positive and true negative rate can be both equal and not equal to one. Thus, the test is dilation-inducing if and only if the identified set intersects the antidiagonal of the unit square.

This characterization states that the test is dilation-inducing if and only if the true positive and true negative rates can sum to one, which is a standard subvector inference problem. We rely on this to propose a statistical test for dilation. We use a representation of the identified set for the misclassification rates via moment inequalities and leverage

the results from Bugni, Canay, and Shi (2017). The resulting statistical test for dilation is uniformly consistent in level over a broad class of permissible distributions.

We apply the developed method to revisit the use of computed tomography (CT) chest scans that were recommended for COVID-19 detection in epidemic areas by highly-cited work in Ai et al. (2020), based on "apparent" performance. We find that the test has indeed induced dilation, despite being considered well-suited for the task. This example indicates that following established practice of relying only "apparent" performance may lead to recommendation or approval of tests that are completely uninformative of the patient's health status and that *only* introduce ambiguity. Moreover, it shows a real-world example of dilation, which has previously been considered a theoretical phenomenon.

We therefore offer a direct policy recommendation to regulators and researchers. We advise verifying whether new diagnostic tests under evaluation are dilation-inducing whenever imperfect reference tests are used, which is pervasive in practice. Our proposed hypothesis test may be used for this purpose, supplementing currently reported results in test performance studies. The results herein rely on standard assumptions in the literature and are readily compatible with prevalent study designs.

In the literature on gold standard bias, it is commonly accepted to assume perfect knowledge of the reference test misclassification rate (see, e.g., Gart and Buck 1966; Staquet et al. 1981; and Zhou, McClish, and Obuchowski 2009). This may be untenable in certain empirical settings. We thus also discuss how the assumption can be weakened by assuming that the rates are only approximately known. We also provide tools for sensitivity analysis when the performance may not be known at all. This may be particularly appealing in other empirical settings not pertaining to diagnostic tests and clinical decision making.

Our results may be directly applied in any context where: 1) a decision-maker updates their beliefs via full Bayesian updating based on information from a binary classifier; 2) misclassification rates of the binary classifier are identified by means of a reference classifier or a label imperfectly measuring the ground truth. For example, one may find them applicable for remote sensing classification (Carlotto 2009 and Foody 2010), or loan risk prediction (Abakarim, Lahby, and Attioui 2018).

The remainder of the paper is structured as follows. Section 3.1.1 reviews the related literature. Section 3.2 introduces the setting, defines dilation and provides identification results. Section 3.3 formalizes the statistical test for the occurrence of dilation. Section 3.4 discusses the use of CT chest scans for COVID-19 detection. Section 3.5 explores extensions that relax the knowledge assumptions about the reference test. Section 3.6 discusses policy implications and concludes. All formal proofs are relegated to the appendix. Section C.4 shows an illustrative application in the context of loan risk prediction.

Subsection 3.1.1. Related Literature

Diagnostic test performance literature has introduced misclassification rates as a parameter of interest (Yerushalmy 1947; Binney, Hyde, and Bossuyt 2021). The concept of *gold standard bias*—the discrepancy between actual and observed misclassification rates as measured against a reference test—is well-known. Early research by Gart and Buck (1966), Staquet et al. (1981), and Zhou, McClish, and Obuchowski (2009) established that when the reference and the test of interest are statistically independent given the patient's true health status, one can point-identify relevant misclassification rates, provided the reference test's performance is known exactly. However, the assumption of conditional independence

is often untenable, particularly when the two tests share physiological bases, as noted by Valenstein (1990), Hui and Zhou (1998), and Emerson et al. (2018). Subsequent studies explored how the disease prevalence and correlation of classification errors of the two tests affect the gold standard bias (Deneef 1987; Boyko, Alderman, and Baron 1988; Valenstein 1990). The main focus of this work is the qualitative direction of the bias. Yet, the practical application of the findings is limited since they depend on latent parameters. Notable exceptions include Thibodeau (1981), and Emerson et al. (2018). Our analyses directly build on the sharp joint identified set for the true misclassification rates from Obradović (2024).

As we illustrate here, this lack of point identification induces ambiguity in the interpretation of the test's result: the test's post-test probabilities are *imprecise*—that is, the probability assignment is not necessarily a unique number.[4] In the realm of imprecise probabilities, the concept of dilation was initially demonstrated by Good (1974). Subsequently, Walley (1991), Seidenfeld and Wasserman (1993) and Herron, Seidenfeld, and Wasserman (1997) systematically analyzed dilation. Our identification results are based on the definitions in the latter two papers.

A substantial corpus of theoretical literature on dilation, which is too extensive to comprehensively summarize here, followed this work. The interested reader may find summaries of recent contributions in Bradley (2019, in particular, Section 3.1) and Gong and Meng (2021), along with the references therein.[5] Implicitly in many of these approaches,

4. Imprecise probability can be seen as a natural extension of usual probability theory and has a long history in the foundations thereof and decision theory. Bradley (2019) provides an overview.

5. Theoretical applications of dilation to economic questions include mechanism design (Bose and Renou 2014), game theory (Beißner and Khan 2019), persuasion (Beauchêne, Li, and Li 2019; Cheng 2023; Pahlke 2024), ordering ambiguous information structures (Wang 2024), Aumann's (1976) agreement theorem (Zhang, Liu, and Seidenfeld 2018), and combining forecasts (Levy and Razin 2022).

and shared by ours, is a sort of *full Bayesian updating* (Pacheco Pires 2002). Alternative updating procedures for ambiguous information have been recently investigated by Gul and Pesendorfer (2021), Dominiak, Kovach, and Tserenjigmid (2022) and Lin and Payró (2024). Moving beyond theoretical work, there has been a recent surge of interest in studying ambiguous information and dilation in experimental economics, exemplified by Hayashi and Wada (2010), Kellner, Le Quement, and Riener (2022), Shishkin and Ortoleva (2023), Epstein and Halevy (2024), Kops and Pasichnichenko (2023), and Liang (2024). Manski (2018) mentions the possibility of dilation in concrete questions about personalized patient care. In contrast to prior research, to the best of our knowledge, this paper is the first to study the identification of dilation in a real-world context and propose a statistical test for its occurrence. However, we do not take a stance on the decision-making processes individuals employ when encountering dilation. Thus, our stance on non-informative tests is (decision) model free.

The test is based on the method for subvector inference in moment inequality models proposed by Bugni, Canay, and Shi (2017). Thus, we also contribute to the recent developments exploring issues of partial identification in medical and epidemiological settings such as Bhattacharya, Shaikh, and Vytlacil (2012), Manski and Molinari (2021), Toulis (2021), Manski (2021), Stoye (2022), Sacks et al. (2022), and Obradović (2024).

Section 3.2. Dilation Identification

We evaluate whether a binary diagnostic test $t \in \{0, 1\}$ induces dilation in the context of clinical decision-making. As we will show, this is closely associated with the misclassification rates of t, also called performance measures.

Misclassification rates are evaluated and reported in test performance studies. The rates are frequently partially identified in practice due the inability to perfectly measure the true health status in the studies. Section 3.2.1 expounds on the details in practically relevant settings. Section 3.2.2 discusses how clinicians use the rates to assess the probability of a patient having the underlying condition upon observing a test result. In Section 3.2.3 we show that utilizing a test whose performance measures are partially identified can induce dilation. Moreover, Section 3.2.4 demonstrates that dilation may happen even if the test seemingly performs well based on conventionally used measures. We thus develop equivalent characterizations that allow practitioners to detect when dilation occurs. Section 3.3 provides a statistical test for dilation based on the characterizations.

Subsection 3.2.1. Test Performance Identification

Let $t = 1$ denote a positive, and $t = 0$ a negative result. Similarly, denote by $y = 1$ the existence of the underlying condition we are classifying, and by $y = 0$ the absence of it. Following standard nomenclature, we refer to t as the *index test* and to y as the *true health status*. Identification of test misclassification rates requires knowledge of y, which is most often unobservable. Otherwise, testing would be superfluous. For this reason, the true health status is commonly measured by a *reference test* r. Let $r = 1$ and $r = 0$ denote positive and negative reference test results, respectively. Each individual in the performance study population is thus characterized by a triple $(t, r, y) \in \{0, 1\}$. Let P denote the joint distribution of the triple.[6]

6. In accordance with the medical literature, we focus on binary tests and health statuses. FDA Statistical Guidance on Reporting Results Evaluating Diagnostic Tests recognizes only binary reference tests and health statuses. Many tests, such as RT-PCR tests, are reduced to binary tests by thresholding.

Test performance is predominantly quantified in the form of *sensitivity* and *specificity*, also referred to as performance measures.

$$\text{Sensitivity: } \theta_1 := P(t = 1|y = 1) \tag{3.1}$$

$$\text{Specificity: } \theta_0 := P(t = 0|y = 0) \tag{3.2}$$

The test r is usually the best currently available test for y. In spite of this, in practice r is almost always imperfect and $P(r = y) < 1$. Consequently, it is critical to allow r to be an imperfect test. Define reference test sensitivity $s_1 := P(r = 1|y = 1)$ and specificity $s_0 := P(r = 0|y = 0)$. For conciseness, let $\theta := (\theta_0, \theta_1)$ and $s := (s_0, s_1)$. It should be noted that r itself may not be adequate for clinical decision-making due to invasiveness, cost or slow turnover. For example, r used for Alzheimer's disease is a post-mortem pathological report (Suemoto and Leite (2023)).

Data in test performance studies are collected by randomly sampling from a population of interest, and testing each observation with both the reference and index tests. The observed outcome for each participant is $(t, r) \in \{0, 1\}^2$. Sampling identifies the joint probability distribution $P(t, r)$. We henceforth denote probability distributions that are derived from observable distributions given s by P_s. All directly observable distributions, such as $P(t, r)$, do not have the subscript.

When $s_1 < 1$ or $s_0 < 1$, so that the reference test is imperfect, $P(t, r)$ will not point identify θ without further assumptions. This fact is well documented in the literature on gold standard bias (Zhou, McClish, and Obuchowski 2009). Moreover, it is known that

While the results of this paper can be extended to cases in which ranges of tests are finite sets, we limit the analysis to the binary setting to conform to current research norms.

incorrectly assuming $s_1 = s_0 = 1$ will not identify the true θ that is needed for clinical decision-making.

Early work accounts for imperfect r and point identifies θ by assuming exact knowledge of s and conditional independence of t and r, i.e. $t \perp\!\!\!\perp r|y$ (Buck and Gart (1966), Staquet et al. (1981)). However, multiple authors, e.g. Vacek (1985), Valenstein (1990), or Hui and Zhou (1998), have argued that conditional independence is implausible in practice. A salient case is when t and r are physiologically related, such as when they rely on the same type of sample (e.g. nasal swab or capillary blood) or measure the same quantities (e.g. antibody reaction to tuberculin).

Thibodeau (1981), and Emerson et al. (2018) thus allow $t \not\perp\!\!\!\perp r|y$ when s is known, partially identifying θ. In other words, there exists a set of values of θ that are consistent with s and observed data $P(t, r)$, called the identified set. Obradović (2024) derives the sharp identified set under standard assumptions. This is the smallest set that contains all values of θ that are consistent with $P(t, r)$ and s, denoted by $\Theta_P(s)$. We first restate the assumptions sufficient to characterize $\Theta_P(s)$ and build upon these results.

Assumption 3.2.1. *(Reference Performance) Sensitivity and specificity of the reference test are known and satisfy* $s_1 + s_0 > 1$.

Knowledge of s is a non-trivial assumption, but it is commonly assumed in literature concerned with gold standard bias correction, such as Gart and Buck (1966), Thibodeau (1981), Staquet et al. (1981), and Emerson et al. (2018). Moreover, the current norm of

assuming that the reference test is perfect imposes a stronger assumption—$s = (1, 1)$—which implies Assumption 3.2.1. Examples in the literature that assume perfect references are too numerous to cite.

Ideally, a performance study with a perfect reference that would identify s may exist. Alternatively, the assumption may be imposed based on knowledge of the physical characteristics of r or its analytical performance – misclassification rates measured based on contrived samples. For further discussion and examples, see Obradović (2024, Section 2.1). Regardless, Assumption 3.2.1 may be restrictive for some applications. However, the commonly maintained assumption $s = (1, 1)$ has been disputed for a plethora of reference tests. This fact indicates that at least a set $\mathcal{S}$ of more credible values s exists for a variety of tests used as r. An extension leveraging this fact is found in Section 3.5. There we will additionally provide an alternative formulation which characterizes the set of values for s such that t is induces dilation. One can then consider whether it is plausible for performance of the reference test to lie in this set or use the results as a sensitivity analysis.

Assumption 3.2.1 further maintains that $s_1 + s_0 > 1$. This implies that $s_1 + s_0 \neq 1$, which excludes the possibility that $r \perp\!\!\!\perp y$. In other words, we require the reference to perform better than a simple coin toss. Otherwise, r provides no information on y, and it cannot be used as a reasonable reference. This is also a minimal requirement for r to be called a test, *c.f.* Rogan and Gladen (1978). Note also that $s_1 > 1 - s_0$ is merely a normalization and therefore without loss of generality. To see that it without loss, consider the alternative case $s_1 < 1 - s_0$. Then it would be possible to redefine $r^* = 1 - r$, so that $s_1^* = 1 - s_1$ and $s_0^* = 1 - s_0$ and therefore also $s_1^* > 1 - s_0^*$.

Assumption 3.2.2. *(Bounded Prevalence) The reference test yield* $P(r=1)$ *satisfies* $1-s_0 < P(r=1) < s_1$*, where* $s=(s_0,s_1)$ *satisfies Assumption 3.2.1.*

Assumptions 3.2.1 and 3.2.2 jointly point identify the study population prevalence via $P_s(y=1) = \frac{P(r=1)+s_0-1}{s_1+s_0-1}$. If $P_s(y=1) \notin [0,1]$ at least one of the two assumptions is refuted. We call Assumption 3.2.2 *Bounded Prevalence* because it is then equivalent to assuming the study population prevalence satisfies $P_s(y=1) \in (0,1)$. This assumption is implicit in any test performance study identifying sensitivity or specificity. Without it, the performance measures are not properly defined.

Under Assumptions 3.2.1 and 3.2.2, and given the data distribution $P(t,r)$, Obradović (2024, Proposition 1) defines $\Theta_P(s)$ as:

$$\Theta_P(s) := \left\{ (\theta_0,\theta_1) \in [0,1]^2 \;\middle|\; \begin{array}{l} \theta_1 \in [\theta_1^L(s), \theta_1^U(s)] \text{ and} \\ \theta_0 = \dfrac{\theta_1 P_s(y=1) - P(t=1)}{P_s(y=0)} + 1 \end{array} \right\}, \tag{3.3}$$

where:

$$\theta_1^L(s) := \frac{1}{P_s(y=1)} \Big[\max\{0, P(t=1,r=0) - s_0 P_s(y=0)\} + \max\{0, P(t=1,r=1) - (1-s_0) P_s(y=0)\} \Big], \tag{3.4}$$

and:

$$\theta_1^U(s) := \frac{1}{P_s(y=1)} \Big[\min\{P(t=1,r=0), (1-s_1) P_s(y=1)\} + \min\{P(t=1,r=1), s_1 P_s(y=1)\} \Big]. \tag{3.5}$$

The result indicates that θ may commonly be partially identified. This depends on the data-generating process (DGP) and the maintained s. As we will show in Section 3.2.3, this will indeed be the case for a large and practically relevant class of DGPs whenever r is imperfect.

Subsection 3.2.2. Clinical Test Use and Post-Test Probabilities

For clinical decision-making it is essential to determine the probability of a patient having the underlying condition upon observing a test result—the *post-test probabilities*. Sensitivity and specificity only measure probabilities of obtaining a particular test result given a specific health status. However, it is standard practice to extrapolate sensitivity and specificity from test performance studies to find post-test probabilities for members of relevant clinical populations (see e.g. Altman and Bland (1994)).[7]

To that end, let $Q(t, y)$ denote a clinical population distribution of interest such that the test has the same sensitivity and specificity as in the performance study population. Formally, $Q(t, y)$ is a distribution such that $Q(t = j|y = j) = \theta_j$ for $j = 0, 1$. The parameters of interest to the clinican are the *positive post-test probability* (PPP) $Q(y = 1|t = 1)$, and the *negative post-test probability* (NPP), $Q(y = 1|t = 0)$. Note that it is possible that $Q(y = 1) \neq P_s(y = 1)$. Suppose also $Q(y = 1) \in (0, 1)$ since using t is not warranted otherwise.

The diagnosis process is commonly formalized in the medical literature as follows (see e.g. McNeil, Keeler, and Adelstein (1975), Watson, Whiting, and Brush (2020a)). The

7. Mulherin and Miller (2002b) and Willis (2008b) discuss designs of performance studies intended to improve generalizability and provide guidance to physicians on how to assess whether performance study measures extrapolate to populations of interest.

clinician first examines a patient and assess the *pre-test probability* $\pi := Q(y = 1) \in (0, 1)$ that the patient has the underlying condition, prior to conducting the test. This is done based on local rates of illness, patients' symptoms and signs, differential diagnoses, and history of relevant exposure. They then use the assessed π and misclassification rates θ, obtained in a test performance study, to find PPP and NPP via Bayes' theorem:

$$v_1(\theta; \pi) := Q(y = 1|t = 1) = \frac{\theta_1 \pi}{\theta_1 \pi + (1 - \theta_0)(1 - \pi)} \text{ and}$$
$$v_0(\theta; \pi) := Q(y = 1|t = 0) = \frac{(1 - \theta_1)\pi}{\theta_0(1 - \pi) + (1 - \theta_1)\pi},$$

In practice, calculating PPP and NPP is typically done via an online or a graphical calculator, nomogram, where the clinician inputs values for π and θ to directly find the relevant post-test probability (Fagan (1975)). See Section C.2 for an example from Caraguel and Vanderstichel (2013). Decisions to undertake further testing or treatment are made based on the post-test probability. In this paper, we focus on informativeness of t in terms of post-test probabilities, and do not discuss the intricacies of the subsequent decision-making.

Remark 3.2.1. The following results do not depend on the clinician accurately assessing π, as they will hold uniformly for all $\pi \in (0, 1)$. However, the results will apply only to clinical populations where the parameter θ reflects the test's performance.

If π and θ are given by points, unique post-test probabilities follow from Bayes' theorem. Figure 3.3a illustrates this updating graphically when the test t is reasonable, i.e. $\theta_1 + \theta_0 > 1$. A positive test results in a post-test probability higher than π, as indicated

by the blue arrow. Conversely, a negative result yields a post-test probability lower than π, as indicated by the red arrow.

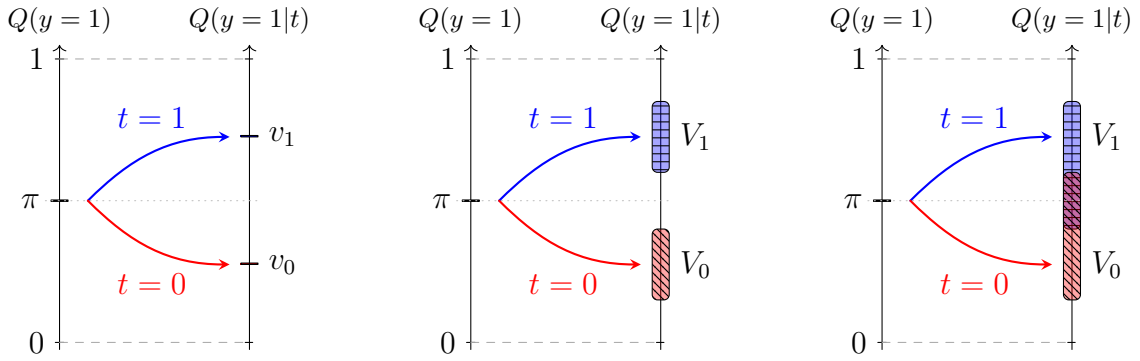
As noted in the preceding section, the parameters in Bayes' theorem may frequently only be bounded and not known exactly because of imperfect reference tests. Clinicians typically account for such ambiguity by finding the corresponding bounds on the post-test probability via full Bayesian updating (see e.g. Baron (1994), Bianchi, Alexander, and Cash (2009), Manski (2021)). Appendix C.2.1 presents an example from Srinivasan, Westover, and Bianchi (2012) showing how this is done via a nomogram. If θ is partially identified, post-test probabilities will also be partially identified. For a generic identified set Θ for θ, we denote the identified sets for PPP and NPP given a value for π as:

$$V_j(\Theta;\pi) := \Big\{ v_j(\theta;\pi) : \theta \in \Theta \Big\} \text{ for } j = 0, 1, \tag{3.6}$$

which are depicted in Figure 3.3b. The interpretation of the post-test probabilities is unchanged, but they are not known exactly. We thus say that the test provides *ambiguous information* when Θ contains more than one point. However, in the figure, the test is still informative. Upon observing $t = 1$, the lower bound on the post-test probability of being diseased lies above π. Conversely, upon observing $t = 0$, the upper bound is below π.

Subsection 3.2.3. Characterizing Dilation

If θ is partially identified, then the ambiguity in the post-test probabilities can be so severe that the test ceases to be informative. Depending on the identified set Θ, the pre-test probability can be strictly contained within the identified set for the post-test probability, regardless of the observed test result. This is the main idea behind the phenomenon known



(a) Unambiguous information (b) Ambiguous test (c) Dilation

Figure 3.3. Updating pre-test to post-test probabilities.

Dependence on π and Θ is suppressed. The left panel depicts point-identified θ with $\theta_1 + \theta_0 > 1$, in the middle panel, θ is partially identified but t is informative. The right panel presents dilation.

as *dilation*. Then test result *only* introduces ambiguity by dilating the pre-test probability and does not provide information about the patient's true health status. Dilation is depicted by Figure 3.3c.

Definition 3.2.1 (Seidenfeld and Wasserman 1993)**.** Let Θ be the identified set for θ. The index test induces *dilation for the pre-test probability* π if

$$\{\pi\} \subsetneq V_1(\Theta;\pi) \text{ and } \{\pi\} \subsetneq V_0(\Theta;\pi). \tag{3.7}$$

An index test is *dilation-inducing* if it induces dilation for any pre-test probability $\pi \in (0,1)$.

We say that t *induces dilation for the pre-test probability* π if π is strictly contained within the identified set for the post-test probability, *regardless of the test outcome*. We

refer to a test t as *dilation-inducing* if it induces dilation for any non-trivial pre-test probability π.

Remark 3.2.2. The original definition by Seidenfeld and Wasserman (1993) allows for ambiguous pre-test probability, where π is only contained within a known set. Extending Definition 3.2.1 and the analysis to allow for this is straightforward, as discussed by Section C.1.

The definition suggests that whether a test t is dilation-inducing critically depends on the identified sets for its performance measures θ. Our first main result characterizes necessary and sufficient conditions for this to occur for a generic connected identified set Θ.

Proposition 3.2.1. *Let Θ be a connected identified set for the performance measures of the index test. The index test is dilation-inducing if and only if there exist $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict. Equivalently, t is dilation-inducing if and only if there exist $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 = 1$ and $\theta'_0 + \theta'_1 \neq 1$.*

Proposition 3.2.1 characterizes whether t is dilation-inducing in the context of clinical decision-making in terms of the identified set for θ. For intuition, focus on the statement $\exists \theta, \theta' \in \Theta$ such that (1) $\theta_1 + \theta_0 = 1$ and (2) $\theta'_1 + \theta'_0 \neq 1$ (and hence $\theta \neq \theta'$). The first condition is equivalent to $t \perp\!\!\!\perp y$, so that t is possibly uninformative. The second condition means that the post-test probabilities is partially identified, i.e. t provides ambiguous information. Intuitively, $\theta_1 + \theta_0 = 1$ ensures that π is always in the identified sets for the post-test probabilities regardless of the test result; $\theta'_1 + \theta'_0 \neq 1$ ensures that there will be at

least one additional value in the identified sets for the post-test probabilities apart from π so that it is non-singleton, i.e. that it "expands" in comparison to π.

Remark 3.2.3. According to Seidenfeld and Wasserman (1993, Theorem 2.1), dilation implies that $t \perp\!\!\!\perp y$ is consistent with the data and assumptions. Proposition 3.2.1 nests this result.

Proposition 3.2.1 shows that partial identification of θ is necessary for the test t to be dilation-inducing. The characterization holds irrespective of the reason for the partial identification of test performance measures. However, Θ is frequently non-singleton in practice because imperfect reference tests are almost always used to identify θ in test performance studies. We thus now put together Proposition 3.2.1 and the results in Section 3.2.1, noting that the findings of the proposition may similarly be used for more general identified sets for θ.[8] By definition in (3.3), $\Theta_P(s)$ is either a point or a line segment, so a connected set, satisfying the conditions of the proposition. This yields the following corollary.

Corollary 3.2.1. *Suppose Assumptions 3.2.1 and 3.2.2 hold, and let $\Theta_P(s)$ denote the corresponding identified set for the performance measures of the index test. The index test is dilation-inducing if and only if there exist $\theta, \theta' \in \Theta_P(s)$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict. Equivalently, t is dilation-inducing if and only if there exist $\theta, \theta' \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$ and $\theta'_1 + \theta'_0 \neq 1$.*

8. Another reason why θ may be partially identified is missing data in the test performance study for either of the tests t or r. This typically occurs in specific study designs relying test result-based sampling, i.e. testing conditional on a test result. Such designs are less common and we do not consider them here.

Remark 3.2.4. Since $\Theta_P(s)$ is either a point or a line segment with positive and finite slope, it is immediate that t is dilation-inducing if and only if there exist $\theta, \theta' \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$ and $\theta \neq \theta'$. That is, it is sufficient and necessary for it to be possible that $\theta_1 + \theta_0 = 1$ and for θ to be partially identified.

The corollary characterizes when t is dilation-inducing in the real-world context of clinical decision making due to imperfect references. As we discuss in Section 3.2.4, this may happen even if t is deemed to have good performance judging based on conventionally-used performance measures. Since imperfect references are ubiquitous and dilation inducing tests are uninformative of the patient's health status, it is critical to enable verification of whether t is dilation-inducing.

Regulatory bodies approving diagnostic tests, such as the FDA, usually require methods of accounting for statistical imprecision. We thus now form a basis for a statistical test of whether t is dilation-inducing. By Remark 3.2.4, t is dilation-inducing if and only if 1) there exists $\theta \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$ and 2) θ is partially identified. The first condition is a subvector inference problem. Thus, if the second condition is fulfilled, the first is conducive to testing by existing inference methods. To this end, we show that θ is indeed partially identified under imperfect r for a broad class of DGPs that are practically relevant in test performance studies. Assumption 3.2.3 defines the class.

Assumption 3.2.3. *(Anything Goes) For any* $(j,k) \in \{0,1\}^2$, $P(t = j, r = k) > 0$.

This condition is mild and realistic in many practical settings. It fails only if a certain result for $r = k$ makes a particular outcome for $t = j$ impossible P-almost surely, or

vice-versa. Such dependence is testable and generally not expected in conventional test performance studies where the index and reference tests are used on each participant.

The assumption typically fails if one test is performed conditional upon a result of the other, i.e. under test result-based sampling. Such designs are not common, and the literature cautions against their use as it may introduce further identification issues (see e.g. Kohn (2022)). Nevertheless, we emphasize that Assumption 3.2.3 is not necessary to identify $\Theta_P(s)$ nor to characterize whether t is dilation-inducing. It is only used to further simplify the characterization which will be used to form a statistical test for dilation in Section 3.3.

Lemma 3.2.1. *Suppose $s \in [0,1]^2$ satisfies Assumption 3.2.1 and maintain Assumption 3.2.2. Then $\Theta_P(s)$—as defined in (3.3)—is non-empty. Furthermore, if Assumption 3.2.3 additionally holds, then $\Theta_P(s)$ is not a singleton set if and only if $s = (s_0, s_1) \neq (1,1)$.*

Lemma 3.2.1 shows that θ will be partially identified if and only if the reference test r is imperfect under a large class of practically-relevant data generating processes given by Assumption 3.2.3. By Remark 3.2.4, in this class and when r is imperfect, t is dilation-inducing if and only if $\exists \theta \in \Theta_P(s) : \theta_1 + \theta_0 = 1$. We formalize this argument in the following main identification result.

Theorem 3.2.1. *Maintain Assumptions 3.2.1, 3.2.2 and 3.2.3, and let $\Theta_P(s)$ be the resulting identified set as in (3.3). Then t is dilation-inducing if and only if (1) $s \neq (1,1)$ and (2) there exists $\theta \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$.*

The results in this section assume knowledge of s, in line with common practice in the literature on gold standard bias correction. However, this might be unsatisfactory for some applications. In Section 3.5, we extend the results leading up to and including Theorem 3.2.1 to cases where s is only known approximately or not at all.

Subsection 3.2.4. Test Approval and Numerical Examples

For approval, a new diagnostic test must often satisfy explicit or implicit minimum requirements for specificity and sensitivity. For example, rapid antigen tests for COVID-19 (ECDC 2021) or influenza (Green and StGeorge 2018) must surpass thresholds of 97% for specificity and 80% for sensitivity to be approved. During public health crises caused by infectious diseases, new tests may be recommended for detection purposes if they exhibit high sensitivity, even when the specificity may be low (Ai et al. 2020).

Tests are usually evaluated based on their "apparent" sensitivity and specificity $\tilde{\theta}_j := P(t = j|r = j)$ for $j = 0, 1$, or the performance of t *with respect to* r. If the reference test is perfect so $s = (1, 1)$, then "apparent" sensitivity and specificity are equal to the *true* sensitivity and specificity $\tilde{\theta}_j = \theta_j$. However, the reference test is almost always imperfect, and so the "apparent" performance measures may not represent *true* performance measures. This discrepancy is called gold standard bias.

Consequently, it is possible that "apparent" measures $\tilde{\theta}_j$ meet the performance requirements while the test is dilation-inducing and thus completely uninformative in the context of clinical decision making. We now show one such example. In Section 3.4 we find that a real-world test that was recommended based on high "apparent" sensitivity for COVID-19 detection induces dilation.

Consider the distribution $P(t, r)$ given by Table 3.1 and representing a DGP for a test performance study. This study corresponds to a test that would (exactly) meet the mentioned minimum threshold for approval of 97% specificity and 80% sensitivity, when measured by "apparent" performance $\tilde{\theta}_j$. That is, the test may be approved based on conventionally-used performance measures.

Table 3.1. A Potentially Worrisome Example of a Diagnostic Test.

$P(t \downarrow, r \rightarrow)$	$r = 0$	$r = 1$	$P(t)$
$t = 0$	38.8%	12.0%	50.8%
$t = 1$	1.2%	48.0%	49.2%
$P(r)$	40%	60%	

However, despite meeting the minimum requirements for approval based on "apparent" performance, depending on s, the test may be dilation-inducing. For example, even with perfect specificity ($s_0 = 1$), the index test could still be dilation-inducing if the reference sensitivity is $s_1 = 62\%$. The same would be true if the reference test had perfect sensitivity ($s_1 = 1$), with specificity of $s_0 = 51\%$.

Although these cases may seem extreme, as they require a seemingly poor reference test, they may still be relevant for specific applications. For instance, some forms of PCR tests for COVID-19 may fall into the latter category.[9] Moreover, Section 3.4 will show an example of a test recommended for use in the literature that may be dilation-inducing even when the reference test is very accurate with $s = (1, 0.9)$. In general, the s needed for the test to induce dilation depends on the distribution $P(t, r)$. We provide further numerical examples to clarify this and other key points from the preceding results.

9. Alcoba-Florez et al. (2020) estimate that the sensitivity for the tests they analyze was as low as 60.2%. Specificity for these tests is typically close to 100% as mentioned above.

3.2.4.1. Additional Numerical Examples. We provide three additional numerical examples for intuition. First, we analyze the extreme case where the index test is independent of the reference test, with the joint distribution given in Table 3.2. Second, we examine a case where the index test is weakly correlated with the reference test, as shown in Table 3.3. Finally, we explore a case where the index test is highly correlated with the reference test, with the corresponding joint distribution depicted in Table 3.4. In all cases, we set $s = (0.9, 0.9)$, indicating that the reference test is imperfect, but performs well. Assumption 3.2.1 under the normalization $s_1 > 1 - s_0$. Additionally, Assumption 3.2.3 is satisfied in all three cases.

Table 3.2. Joint Distribution of Independent Index and Reference Tests.

$P(t \downarrow, r \rightarrow)$	$r = 0$	$r = 1$	$P(t)$
$t = 0$	25%	25%	50%
$t = 1$	25%	25%	50%
$P(r)$	50%	50%	

Table 3.3. Joint Distribution of Two Tests with Weak Correlation.

$P(t \downarrow, r \rightarrow)$	$r = 0$	$r = 1$	$P(t)$
$t = 0$	30%	20%	50%
$t = 1$	20%	30%	50%
$P(r)$	50%	50%	

Table 3.4. Joint Distribution of Two Tests with Strong Correlation.

$P(t \downarrow, r \rightarrow)$	$r = 0$	$r = 1$	$P(t)$
$t = 0$	40%	10%	50%
$t = 1$	10%	40%	50%
$P(r)$	50%	50%	

The left panel of Figure 3.4 illustrates the resulting identified set for θ for the first DGP where the index test is independent of the reference test. Even though the performance of the reference test is known precisely, the set $\Theta_P(s)$ still contains multiple possible values for θ, represented by the dark, solid line in the figure. Equivalently, we lack point identification. Intuitively, this is because we do not know the exact correlation between the index test and reference tests *conditional on* the underlying health condition, despite knowing that the index and reference tests are independent. There is a set of correlation structures between t and r conditional on y that are consistent with $P(t, r)$. Hence, there will also be a set of possible correlation structures between t and y, and θ. For the other two DGPs, partial identification occurs for the same reason, and is depicted in the center and right panels of Figure 3.4.

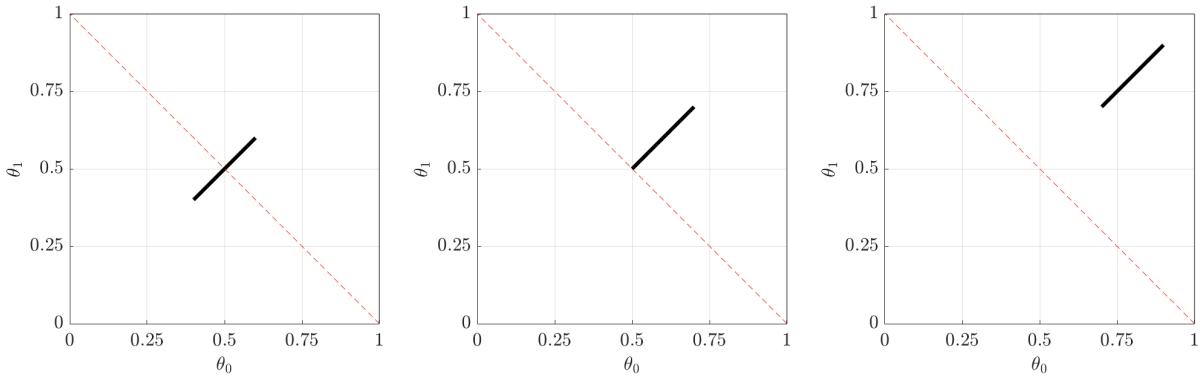


Figure 3.4. $\Theta_P(s)$ for independent tests, weakly correlated, and highly correlated test, respectively from left to right.

Perhaps unsurprisingly, the index test induces dilation in the first case.[10] This is clearly visible in Figure 3.4 by applying Theorem 3.2.1: the index test is dilation-inducing if and

10. It is worth noting, however, that the index test would *not* induce dilation if the reference test were perfect. Then $P(t, r) = P(t, y)$ so θ would be point identified and partial identification is necessary for dilation to occur.

only if the identified set intersects the antidiagonal, represented by the red, dashed line in the figure. This illustrates the simplification provided by Theorem 3.2.1. Now for the correlated cases in the center and right panel of Figure 3.4, we observe that the index test remains dilation-inducing in the case of weak correlation, but not when the correlation between the tests is high. Therefore, only in the high correlation case is the index test informative in this specific sense.

We can see whether t induces dilation for specific values of π from the post-test probabilities. In Figure 3.5, we focus on a specific pre-test probability, $\pi = 0.5$.[11] This figure is in line with the fact that t is dilation-inducing in the first two cases (independence and weak correlation), but not in the case of high correlation.

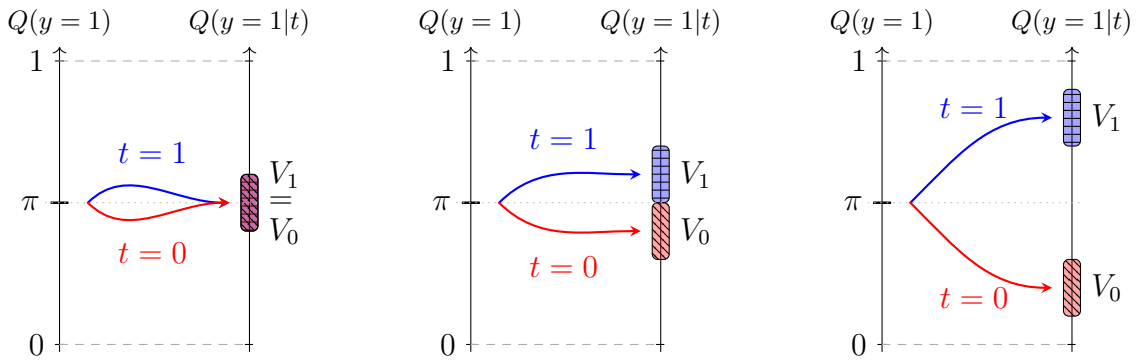


Figure 3.5. Sets of post-test probabilities for independent, weakly correlated, and strongly correlated tests.

Section 3.3. Statistical Inference

Theorem 3.2.1 equivalently characterizes when the index test t is dilation-inducing, but requires knowledge of the data distribution $P(t, r)$. In practice, this distribution is

11. Note, however, that the specific value of π does not affect the qualitative features of the figure, as dilation is defined uniformly over all $\pi \in (0, 1)$.

typically unknown, and only sample data from $P(t,r)$ are available. One can "estimate" whether t is dilation-inducing by analogy. Replacing population parameters with consistent estimators in closed-form expressions yields the consistent plug-in estimator $\hat{\Theta}_P(s)$ of the identified set $\Theta_P(s)$ (Manski and Pepper 2000b and Tamer 2010). Then, based on Theorem 3.2.1, t is "estimated" to be dilation-inducing if there exist $\theta \in \hat{\Theta}_P(s)$ such that $\theta_1 + \theta_0 = 1$.

To account for sampling variability, as typically required by regulatory bodies, we develop an inference procedure to test whether t is dilation-inducing based on Theorem 3.2.1. The procedure is uniformly consistent in level across a broad class of permissible distributions. Canay and Shaikh (2017) and Canay, Illanes, and Velez (2023, Remark 2.1) underline the importance of uniformity. If it fails, some distributions of the data $P(t,r)$ may result in coverage probability that is too low even arbitrarily large sample sizes. In that sense, inference based on confidence intervals that are consistent only pointwise may be severely misleading in finite samples.

Subsection 3.3.1. Baseline Assumptions

Let $W_i = (t_i, r_i) \in \{0,1\}^2$ for $i = 1, \ldots, n$ represent the observed data from n observations of the distribution $P(t,r)$. In our setting, the distribution of the observed data is a categorical distribution $P(t,r)$ for $(t,r) \in \{0,1\}^2$. We assume that the distribution of observed data P belongs to a baseline distribution space denoted by $\mathcal{P}$. Note that every $P \in \mathcal{P}$ can be represented as an element Δ^3—the three-simplex. Therefore, we represent $\mathcal{P}$ as a subset of a Euclidean space and endow $\mathcal{P}$ with the Euclidean topology, which in our case is the same as the (usual) weak topology on the space of probability distributions.

Following common practice, we will assume that we have access to a random sample.

Assumption 3.3.1. *(Random Sampling) The study sample is a sequence of i.i.d. random vectors $W_i = (t_i, r_i)$, where each W_i follows the distribution $P \in \mathcal{P}$.*

To provide an appropriately uniform statistical test, we strengthen some of the assumptions from Section 3.2.

Assumption 3.2.2'. *(Uniformly Bounded Prevalence) There exists $\varepsilon_r \in (0, \bar{\varepsilon})$ such that for every $P \in \mathcal{P}$, we have $P(r = 1) \in [1 - s_0 + \varepsilon_r, s_1 - \varepsilon_r]$, where $s = (s_0, s_1)$ satisfies Assumption 3.2.1.*

Together with Assumption 3.2.1, Assumption 3.2.2' is equivalent to a uniform bound on the implied prevalence $P_s(y = 1)$ over the admissible distributions $\mathcal{P}$. To see this, recall—from the discussion following Assumption 3.2.2—that specific s determines $P_s(y = 1)$ via $P(r = 1)$. Here, we have the corresponding statement that holds uniformly across all $P \in \mathcal{P}$. We also strengthen Assumption 3.2.3.

Assumption 3.2.3'. *There exists an $\varepsilon_d \in \left(0, \frac{1}{4}\right)$ such that for every $P \in \mathcal{P}$ and every $(j, k) \in \{0, 1\}^2$, $P(t = j, r = k) \geq \varepsilon_d$.*

Under these strengthened assumptions, the baseline distribution space $\mathcal{P}$ is compact, as formally established in the following lemma.

Lemma 3.3.1. *If Assumption 3.2.1, Assumption 3.2.2' and Assumption 3.2.3' hold, then $\mathcal{P}$ is compact.*

Subsection 3.3.2. The Proposed Test

Using Theorem 3.2.1, we can formulate the hypothesis as follows:

$$H_0 : \theta_0 + \theta_1 = 1 \quad \text{vs.} \quad H_1 : \theta_0 + \theta_1 \neq 1. \tag{3.8}$$

As mentioned previously, given that we have a partially identified model, it is crucial to propose a test that remains uniformly valid across all distributions in the set $\mathcal{P}$. To do so, we leverage on two recent results. First, following Obradović (2024) we characterize the identified set through moment (in)equalities. We then apply the minimum resampling test from Bugni, Canay, and Shi (2017) to this characterization and prove that uniformity holds.

We begin by recasting $\Theta_P(s)$ via moment (in)equalities through the introduction of an appropriate *moment function*, defined as follows:

$$m(W_i, \theta; s) := \begin{pmatrix} m_1(W_i, \theta; s) \\ m_2(W_i, \theta; s) \\ m_3(W_i, \theta; s) \\ m_4(W_i, \theta; s) \\ m_5(W_i, \theta; s) \\ m_6(W_i, \theta; s) \\ m_7(W_i, \theta; s) \end{pmatrix} := \begin{pmatrix} (\theta_1 - s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - (t_i - 1)r_i \\ (\theta_1 - 1 + s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - (r_i - 1)(1 - t_i) \\ (\theta_1 - 1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - (t_i - 1) \\ -\theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i \\ (-\theta_1 + s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i(1 - r_i) \\ (-\theta_1 + 1 - s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i r_i \\ (\theta_0 - 1)\left(1 - \frac{r_i - 1 + s_0}{s_1 - 1 + s_0}\right) - \theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i \end{pmatrix}, \tag{3.9}$$

where $W_i = (t_i, r_i)$. Intuitively, the first three functions represent the condition $\theta_1 \geq \theta_1^L(s)$ (*c.f.* (3.4)). Note that summing two max functions—each with two arguments—yields four cases. However, one of these cases, namely $\theta_1 \geq 0$, is already encompassed by the parameter

space for θ. Therefore, only three functions are both necessary and sufficient. Similarly, functions 4 through 6 represent $\theta_1 \leq \theta_1^U(s)$ (*c.f.* (3.5)). Thus, the first three functions constrain the lower bound, while the next three enforce the upper bound, ensuring the identified set respects both conditions of (3.3) on θ_1. Finally, the last function establishes the linear relationship between θ_0 and θ_1 (*c.f.* (3.3)). These functions collectively allow us to represent the identified set $\Theta_P(s)$ through moment (in)equalities, thus enabling the use of existing inference procedures.

Proposition 3.3.1 (Obradović 2024, Proposition 5)**.** *Suppose s satisfies Assumption 3.2.1 and P satisfies Assumption 3.2.2, then*

$$\Theta_P(s) = \left\{ (\theta_0, \theta_1) \in [0,1]^2 \;\middle|\; \begin{array}{l} (\forall j = 1, \ldots, 6)\ \mathbb{E}_P[m_j(\cdot, \theta, s)] \geq 0, \\ \\ \textit{and } \mathbb{E}_P[m_7(\cdot, \theta, s)] = 0. \end{array} \right\}.$$

Following Bugni, Canay, and Shi (2017), we define the test using this characterization. To define the test statistic T_n, we introduce the necessary additional notation. For $j = 1, \ldots, 7$, let

$$\bar{m}_{n,j}(\theta; s) := \frac{1}{n} \sum_{i=1}^{n} m_j(W_i, \theta; s), \text{ and}$$

$$\hat{\sigma}_{n,j}(\theta; s) := \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[m_j(W_i, \theta; s) - \bar{m}_{n,j}(\theta; s) \right]^2},$$

denote the sample mean and standard variance of the moment functions, respectively. Furthermore, let the so-called modified method of moments test statistic be:

$$Q_n(\theta; s) := \sum_{j=1}^{6} \left[\min \left\{ 0, \frac{\bar{m}_{n,j}(\theta; s)}{\hat{\sigma}_{n,j}(\theta; s)} \right\} \right]^2 + \left[\frac{\bar{m}_{n,7}(\theta; s)}{\hat{\sigma}_{n,7}(\theta; s)} \right]^2.$$

Then, we define the profiled test statistic as:

$$T_n := \min_{\theta \in \Theta_0} Q_n(\theta; s),$$

where $\Theta_0 = \{(\theta_0, \theta_1) \in [0,1]^2 \mid \theta_0 + \theta_1 = 1\}$ represents the antidiagonal of the unit square, that is the subset of the parameter space which satisfies the null hypothesis. To determine whether the test statistic is sufficiently large to reject the null hypothesis, we use the minimum resampling critical value $\hat{c}_n^{1-\alpha}$ for the significance level $\alpha \in (0,1)$, following Bugni, Canay, and Shi (2017, Section 2). $\hat{c}_n^{1-\alpha}$ is the $1-\alpha$ quantile of $T_n^{MR} := \min\{T_n^{DR}, T_n^{PR}\}$. To define T_n^{DR} and T_n^{PR}, we require additional notation. For $j \in \{1, \ldots, 7\}$ define the following stochastic process:

$$\nu_{n,j}(\theta; s) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{m_j(W_i, \theta; s) - \bar{m}_{n,j}(\theta; s)}{\hat{\sigma}_{n,j}(\theta; s)} \zeta_i,$$

where $\zeta_i \overset{i.i.d.}{\sim} N(0,1)$ for $i = 1, \ldots, n$ and independent of W_i. Next, for $j \in \{1, \ldots, 7\}$ define the slackness function:[12]

$$\ell_j(\theta; s) := \frac{\sqrt{n}}{\sqrt{\ln n}} \times \frac{\bar{m}_{n,j}(\theta; s)}{\hat{\sigma}_{n,j}(\theta; s)}$$

12. In our implementation, we use the tuning parameter $\kappa_n := \sqrt{\ln n}$ as suggested by Andrews and Soares (2010) and Bugni, Canay, and Shi (2017), but as explained in there, any $\kappa_n \to \infty$ with $\kappa_n / \sqrt{n} \to 0$ as $n \to \infty$ would work too.

and, for $j \in \{1, \ldots, 6\}$ set $\varphi_j(\theta; s) := \infty$ if and only if $\ell_j(\theta; s) > 1$ and zero otherwise. Then the statistics are:[13]

$$T_n^{DR} := \inf_{\theta \in \Theta_0 : Q_n(\theta;s) \leq T_n} \left\{ \sum_{j=1}^{6} \min\left\{0, \nu_{n,j}(\theta; s) + \varphi_j(\theta; s)\right\}^2 + \nu_{n,7}(\theta; s)^2 \right\}, \tag{3.10}$$

$$T_n^{PR} := \inf_{\theta \in \Theta_0} \left\{ \sum_{j=1}^{6} \min\left\{0, \nu_{n,j}(\theta; s) + \ell_j(\theta; s)\right\}^2 + \left[\nu_{n,7}(\theta; s) + \ell_7(\theta; s)\right]^2 \right\}.$$

With this notation, we formally establish that our proposed test controls size uniformly over all $P \in \mathcal{P}$, under the assumptions stated in Section 3.3.1.

Theorem 3.3.1. *Let Assumptions 3.2.1, 3.2.2', 3.2.3', and 3.3.1 hold. Then, for all* $\alpha \in \left(0, \frac{1}{2}\right)$,

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P} : \Theta_P(s) \cap \Theta_0 \neq \emptyset} P\left[T_n > \hat{c}_n^{1-\alpha}\right] \leq \alpha.$$

Theorem 3.3.1 demonstrates that the proposed test is uniformly consistent in level over the set of permissible distribution functions satisfying the assumptions, as the sample size n tends to infinity. The proof relies on an application of Theorem 4.1 from Bugni, Canay, and Shi (2017). To do so, we verify the conditions of their theorem, including the relevant polynomial minorant condition and the uniform Donsker and pre-Gaussian properties directly.

13. In our implementation, we take the infimum over $\left\{\theta \in \Theta_0 \mid Q_n(\theta; s) \leq T_n + 10^{-4}\right\}$ in (3.10) and use $\hat{c}_n^{1-\alpha} + 10^{-6}$ as the critical value. The introduction of these constants is explained by Bugni, Canay, and Shi (2017, Remark 4.1) and Bugni, Canay, and Shi (2017, Remark B.2), respectively. The value for the former follows from the implementation in Bugni, Canay, and Shi (2017) and for latter from Andrews and Shi (2013, p.625).

Subsection 3.3.3. Simluations

In this section, we analyze the finite sample behavior of our method through a simulation study. To provide a meaningful comparison, we evaluate the performance of our method alongside two alternative approaches.

First, utilizing the procedure from Romano, Shaikh, and Wolf (2014), we consider a projection-based test. Obradović (2024, Theorem 1) demonstrates the procedure yields a uniformly valid confidence set for θ. In turn, by projection, this yields a uniformly valid confidence set for $\theta_0 + \theta_1$, and therefore a test for the hypothesis $H_0 : \theta_1 + \theta_0 = 1$. Second, we develop and evaluate a simple projection test based on the results of Goodman (1965), which yield a joint confidence set for $P(t, r)$, leveraging the multinomial nature of the observed data. Since $P(t, r)$ are the only unknown parameters that determine θ, this confidence interval similarly yields a confidence interval for $\theta_1 + \theta_0$ by projection, and thus a test for dilation. We discuss the construction of this alternative test in detail in Section C.3.

Throughout the simulation study, we fix the reference test with performance measure $s = (0.9, 0.9)$ and use five data-generating processes: the three numerical examples used for intuition in Section 3.2.4 and two additional designs. Table 3.5 shows the first, involving a slight perturbation of the joint distribution in Table 3.3, which previously resulted in dilation. The second additional design, shown in Table 3.6, increases the correlation between the index and reference tests beyond the weak correlation case but remains less correlated than the highly correlated case in Table 3.4. Both designs serve to provide more insight into the power properties of the test.

Table 3.5. Perturbation of Table 3.3.

$P(t\downarrow, r\rightarrow)$	$r=0$	$r=1$	$P(t)$
$t=0$	31%	19%	50%
$t=1$	19%	31%	50%
$P(r)$	50%	50%	

Table 3.6. Joint Distribution of Two Tests with Intermediate Correlation.

$P(t\downarrow, r\rightarrow)$	$r=0$	$r=1$	$P(t)$
$t=0$	35%	15%	50%
$t=1$	15%	35%	50%
$P(r)$	50%	50%	

We use three different sample sizes: $n \in \{50, 100, 500\}$. These sample sizes are relatively small but are typical for the applications in mind. For each design, we perform $1,000$ Monte Carlo simulations and set the significance level at 5%.

Table 3.7. Simluation Tesults: Rejection Probabilites.

		Design 1	Design 2	Design 3	Design 4	Design 5
DGP		Table 3.2	Table 3.3	Table 3.5	Table 3.6	Table 3.4
H_0		true	true	false	false	false
	G	0.9%	0.6%	0.2%	0.7%	4.3%
$n=50$	RSW	0%	0%	0%	0.1%	4.3%
	BCS	0%	0.8%	3.9%	22%	76%
	G	0%	0%	0%	1%	30%
$n=100$	RSW	0%	0%	0%	0.1%	19%
	BCS	0%	1%	3.1%	43%	100%
	G	0%	0%	0%	50%	100%
$n=500$	RSW	0%	0%	0%	28%	100%
	BCS	0%	1.3%	9.2%	99%	100%

$1,000$ Monte Carlo iterations. G, RSW, and BCS denote the tests based on Goodman (1965), Romano, Shaikh, and Wolf (2014), and Bugni, Canay, and Shi (2017), respectively. The last one is our proposed test.

Table 3.7 presents the results from the simulation study, showing the rejection probabilities. All three tests are conservative, as is common in partially identified models: the tests reject the null hypothesis with a probability lower than the nominal 5% significance level when the null hypothesis is true (so dilation occurs). However, for the first two designs, where the null hypothesis is true, our test tends to reject it more frequently than the alternatives while maintaining significance below the nominal 5% level.

The proposed test also consistently shows substantially higher power across all sample sizes for the remaining three designs, where the null hypothesis is false. Focusing on Design 4 and the largest sample size, $n = 500$, our test rejects the null hypothesis almost 100% of the time, whereas the other tests have rejection probability of at most 50%. Notably, in the borderline case of Design 3—an especially difficult scenario with no dilation—our test still manages to reject the null hypothesis occasionally. In contrast, the other two tests never do. Thus, we conclude that even with relatively small sample sizes, our proposed test demonstrates reasonable power compared to the alternatives.

Section 3.4. Empirical Application: CT Chest Scans for Detection of COVID-19

Early in the COVID-19 pandemic, some hospitals used CT chest scans, interpreted by radiologists, as a method to detect COVID-19. This practice was endorsed by highly-cited work, such as Ai et al. (2020), based on the conventional use of "apparent" sensitivity and specificity $\tilde{\theta}_j = P(t = j|r = j)$, i.e. the performance with respect to the reference. We revisit these findings.

CT chest scans were evaluated against a PCR test, which served as the reference. PCR tests are known not to be perfectly sensitive (see e.g. Arevalo-Rodriguez et al. (2020)). In this context, the index test t is the CT chest scan, the reference test r is the PCR test, and the underlying health status y is whether the patient has COVID-19. We use the data from Ai et al. (2020) shown in Table 3.8, collected in a hospital in Wuhan, China, in early 2020. At that early stage of the pandemic, the authors (p. E32) concluded that "Chest CT may be considered as a primary tool for the current COVID-19 detection in epidemic areas." Following Kanji et al. (2021), we assume $s = (1, 0.9)$.

Table 3.8. Data from Ai et al. (2020).

	$r = 0$	$r = 1$	
$t = 0$	105	21	126
$t = 1$	308	580	888
	413	601	$n = 1014$

Figure 3.6 displays the plug-in estimate of the identified set for the performance measures of the CT chest scan, $\hat{\Theta}_P(s)$. Note that the estimated set intersects the antidiagonal, meaning that it contains a θ such that $\theta_1 + \theta_0 = 1$. This suggests that the test may be dilation-inducing. Additionally, on an example of a pre-test probability of $\pi = 1/3$, Figure 3.7 illustrates the ambiguity regarding the post-test probabilities following from the estimates. While a positive CT chest scan yields relatively little ambiguity—reflected by the size of the resulting set—there is significantly more ambiguity following a negative CT scan. The intersection of the two-sets is non-empty, suggesting that the test may be inducing dilation for $\pi = 0.3$.

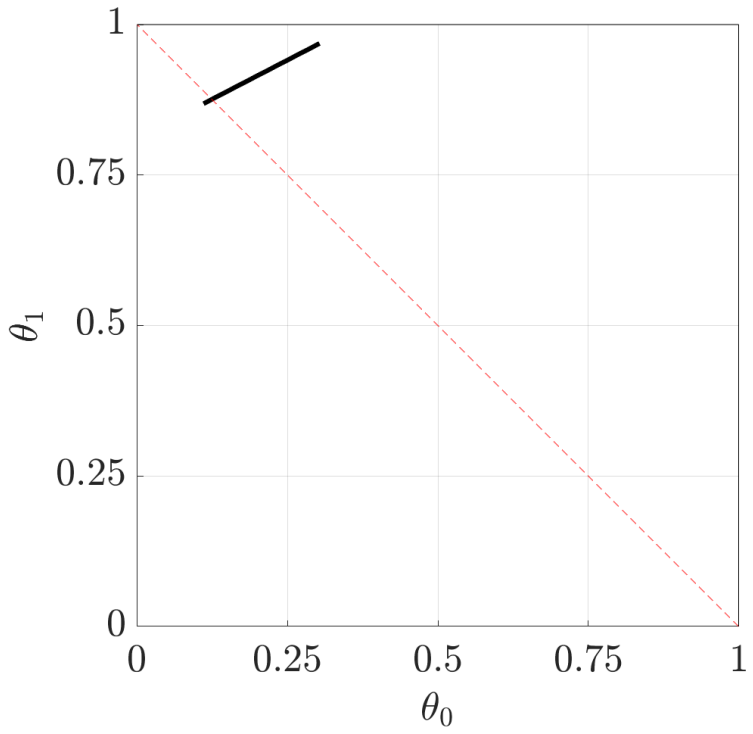


Figure 3.6. $\Theta_P(s)$ for the empirical distribution of Ai et al. (2020) assuming $s = (1, 0.9)$.

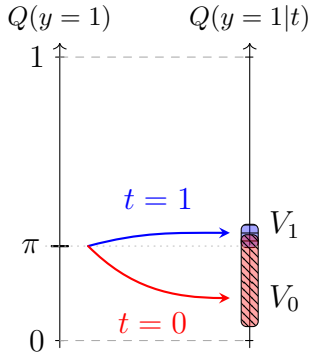


Figure 3.7. Sets of post-test probabilities from Ai et al. (2020) for $\pi = 1/3$.

Applying our proposed test from Section 3.3.2 at a nominal significance level of $\alpha = 5\%$, we obtain a test statistic of $T_n = 1.25 \times 10^{-18}$ and a critical value of $\hat{c}_n^{1-\alpha} = 1.11$. Therefore,

we cannot reject the null hypothesis that the CT chest scan is dilation-inducing. By varying the significance level, we find that the p-value for this null hypothesis is greater than 99%. This result strongly suggests that the CT chest scan is indeed dilation-inducing and completely uninformative of the health status, despite being recommended for use based on conventional performance measures. Moreover, to the extent of our knowledge, this finding represents the first concrete real-world instance of dilation.

Section 3.5. Extensions and Discussion

Subsection 3.5.1. Applicability Beyond Clinical Decision-making

While this paper focuses on the setting of clinical decision-making the results may directly be used in any context where: 1) a decision-maker updates their beliefs via full Bayesian updating based on information from a binary classifier t; 2) misclassification rates of the binary classifier are identified by means of a reference classifier or a label r imperfectly measuring the ground truth y.

The latter is particularly prominent in various empirical settings. In general, r may be imperfect when determined by labelers or algorithms (Cannings, Fan, and Samworth 2020). Labels r may frequently be assigned by an algorithm or human input via platforms like the Amazon Mechanical Turk. Mislabeling may happen due to algorithm imperfections, human error, inattentive labelers, or malicious mislabelling activity. For example, Carlotto (2009) explains that the ground truth is often imperfectly measured in remote sensing applications via a classifier r, and that it may be possible to learn (s_1, s_0) in validation studies. Foody (2010) notes post-test probabilities are formed using information from a

binary remote sensing classifier t by inputting π and θ into the Bayes rule, just as in the clinical decision-making case.

Similarly, an algorithm t may be used to classify loans as risky or not risky with the goal of informing the loan approval process (Abakarim, Lahby, and Attioui 2018). The algorithms seek to predict whether the borrower will default or not – y. However, the algorithm misclassification rates may be evaluated in a dataset where y is measured by an assessment of whether the borrower is "risky" – r. This assessment may be an imperfect measurement of y. Moreover, evaluations of such algorithms often report post-test probabilities obtained as outlined previously. As an illustration, in Section C.4 we revisit such an algorithm. In this case determining the true s may be challenging, which further motivates the following extensions.

Subsection 3.5.2. Uncertainty about the reference test's performance

Thus far, we have maintained Assumption 3.2.1, which assumes that the performance of the reference test, while allowed to be imperfect, is exactly known. This assumption may be restrictive. For example, in Section 3.4, to apply our framework, we assumed that the sensitivity and specificity of the reference test—specifically, the PCR test for COVID-19—were known exactly. Such assumptions may be difficult to justify in various settings.

In this section, we outline how our approach can be extended to account for uncertainty regarding the reference test's performance. Specifically, we aim to accommodate situations where the sensitivity and specificity of the reference test lie within a possibly non-singleton set $\mathcal{S}$. We then generalize Assumption 3.2.1 as follows.

Assumption 3.2.1S. *(Generalized Reference Performance) Sensitivity and specificity of the reference test are contained in a known (i.e. non-empty) and connected set* $\mathcal{S} \subset [0,1]^2$ *such that* $s_1 + s_0 > 1$ *holds for all* $s \in \mathcal{S}$.

Connectedness of $\mathcal{S}$ will be used to extend our characterization in Theorem 3.2.1. Assumption 3.2.1S accommodates sets such as singleton sets, line segments, rectangular Cartesian products of closed intervals, general convex polygons, or closed disks that do not contain points where $s_1 + s_0 = 1$.

Although knowledge of $\mathcal{S}$ is also a non-trivial assumption, it is clearly a relaxation of Assumption 3.2.1. The fact that the common assumption $s = (1,1)$ has been disputed for a plethora of reference tests indicates that at least a set $\mathcal{S}$ of more credible values s exists for a variety of tests used as r. We maintain that $s_1 + s_0 \neq 1$ holds for all $s \in \mathcal{S}$, extending this condition from earlier. Given that $s_1 + s_0 \neq 1$ for any s and that $\mathcal{S}$ is connected, the entire set lies either fully above or fully below the antidiagonal of the unit rectangle. Thus, we impose the same normalization, $s_0 + s_1 > 1$, but now across all $s \in \mathcal{S}$.

Treating $\Theta_P(\cdot)$, as defined in (3.3), as a correspondence, we can readily extend the sharp identified set of performance measures for the index test. Denote it as $\Theta_P(\mathcal{S})$ and observe that it is given by the image of $\mathcal{S}$ under $\Theta_P(\cdot)$. Moreover, the connectedness of $\mathcal{S}$ carries over to $\Theta_P(\mathcal{S})$ under a suitable extension of Assumption 3.2.2.

Assumption 3.2.2S. *(Generalized Bounded Prevalence) The reference test yield* $P(r=1)$ *satisfies* $1 - s_0 < P(r=1) < s_1$ *for all* $s \in \mathcal{S}$*, where* $\mathcal{S}$ *satisfies Assumption 3.2.1S.*

Lemma 3.5.1. *If Assumptions 3.2.1S and 3.2.2S hold, then* $\Theta_P(\mathcal{S})$ *is a non-empty connected set.*

Lemma 3.5.1 ensures that Proposition 3.2.1 remains applicable, and we therefore get a direct generalization of Corollary 3.2.1.

Corollary 3.5.1. *Suppose Assumptions 3.2.1S and 3.2.2S hold, and let* $\Theta_P(\mathcal{S})$ *denote the corresponding identified set for the performance measures of the index test. The index test is dilation-inducing if and only if there exist* $\theta, \theta' \in \Theta_P(\mathcal{S})$ *such that* $\theta_0 + \theta_1 \leq 1$ *and* $\theta'_0 + \theta'_1 \geq 1$*, where at least one inequality is strict. Equivalently,* t *is dilation-inducing if and only if there exist* $\theta, \theta' \in \Theta_P(\mathcal{S})$ *such that* $\theta_1 + \theta_0 = 1$ *and* $\theta'_1 + \theta'_0 \neq 1$*.*

Furthermore, we also derive a similar implication of Assumption 3.2.3 as in Lemma 3.2.1. Namely θ will be partially identified whenever the reference test is not perfect under the large class of practically relevant DGPs given by the assumption.

Lemma 3.5.2. *Suppose* $\mathcal{S} \subset [0,1]^2$ *satisfies Assumption 3.2.1S and maintain Assumptions 3.2.2S and 3.2.3.* $\Theta_P(\mathcal{S})$ *is not a singleton set if and only if* $\mathcal{S} \neq \{(1,1)\}$*.*

Now, all these results together allow us to present the generalization of the main identification of dilation in Theorem 3.2.1, which can form a basis for statistical inference.

Theorem 3.5.1. *Maintain Assumptions 3.2.1S, 3.2.2S and 3.2.3, and let* $\Theta_P(\mathcal{S})$ *be the resulting identified set. Then* t *is dilation-inducing if and only if (1)* $\mathcal{S} \neq \{(1,1)\}$ *and (2) there exists* $\theta \in \Theta_P(\mathcal{S})$ *such that* $\theta_1 + \theta_0 = 1$*.*

By incorporating s as part of the parameter space, we furthermore can characterize the identified set by means of moment (in)equalities. In fact, a direct application of Obradović

(2024, Proposition 5) gives here too that, under Assumptions 3.2.1S and 3.2.2, we have

$$\Theta_P(\mathcal{S}) = \left\{ (\theta_0, \theta_1, s_0, s_1) \in [0,1]^2 \times \mathcal{S} \;\middle|\; \begin{array}{l} (\forall j = 1, \ldots, 6)\ \mathbb{E}_P[m_j(\cdot, \theta, s)] \geq 0, \\ \text{and } \mathbb{E}_P[m_7(\cdot, \theta, s)] = 0. \end{array} \right\}.$$

With this in hand, one could proceed similarly to Section 3.3.2 to develop a test for the null hypothesis that the index test is dilation-inducing.

Subsection 3.5.3. Lack of knowledge of the reference test's performance: the dilator set

The previous section outlined an extension for cases where the performance of the reference test is not exactly known, but still assumes some knowledge of the reference test's performance. However, in certain applications, such as the one discussed in Section C.4, even this assumption may be prohibitively restrictive. In these situations, the researcher might prefer not to make any assumptions about the reference test. Therefore, we now lay out how our approach can be extended to accommodate this lack of knowledge by introducing the concept of a *dilator set*.

We can intuitively ask what reference test performance s would make the index test dilation-inducing, given the data $P(t, r)$. By collecting all such performance measures for the reference test, we define what we call the dilator set. More formally, recall that $\Theta_P(\cdot)$ can be viewed as a correspondence, with the reference test's performance as input. This correspondence can be easily extended to the domain $S_{\geq} := \{(s_0, s_1) \in [0,1]^2 \mid s_0 + s_1 \geq 1\}$ with some care regarding special cases. First, when $s_0 + s_1 = 1$ or $P_s(y = 1) \in \{0, 1\}$, we

define $\Theta_P(s) = [0,1]^2$.[14] Second, for all values of s such that $P_s(y=1) \notin [0,1]$, we define $\Theta_P(s) = \emptyset$ as those values are incompatible with the observed data and assumptions, i.e. they are refuted. The dilator set, $\mathcal{D}_P$, is then the (lower) inverse of the correspondence $\Theta_P(\cdot)$, evaluated at the antidiagonal, denoted by Θ_0:

$$\mathcal{D}_P := \{s \in S_{\geq} \mid \Theta_P(s) \cap \Theta_0 \neq \emptyset\}.$$

When Assumptions 3.2.2 and 3.2.3 hold, $s \in \mathcal{D}_P$ if and only if the index test is dilation-inducing by Theorem 3.2.1.[15] In this case, it is easy to see that $\mathcal{D}_P$ is non-empty and closed and, furthermore, its name as *dilator set* is justified. Figure 3.8 shows an example of a dilator set using the example in Table 3.1.

Furthermore, exploiting Obradović (2024, Proposition 5) once more, the dilator set $\mathcal{D}$ can be reformulated in terms of moment inequalities, as formally established in Proposition 3.5.1 below. This allows construction of confidence sets for the dilator set using existing methods.

Proposition 3.5.1.

$$\mathcal{D}_P = \mathcal{D}_P^{\llcorner} \cup S_0,$$

14. Note that $s_0 + s_1 = 1$ means that the reference test is independent of the underlying health status, and therefore θ is unindentified. If $P_s(y=1) \in \{0,1\}$, then either sensitivity or specificity of the index test is not well-defined. In this case, we could also set $\Theta_P(s)$ to be a proper subset of $[0,1]^2$ with one parameter being point-identified and the remaining being bounded by $[0,1]$. This would not affect any of our discussion or results.

15. Strictly speaking, Theorem 3.2.1 does not apply if $s_0 + s_1 = 1$, but with the convention that $\Theta_P(s) = [0,1]^2$ the theorem naturally extends.

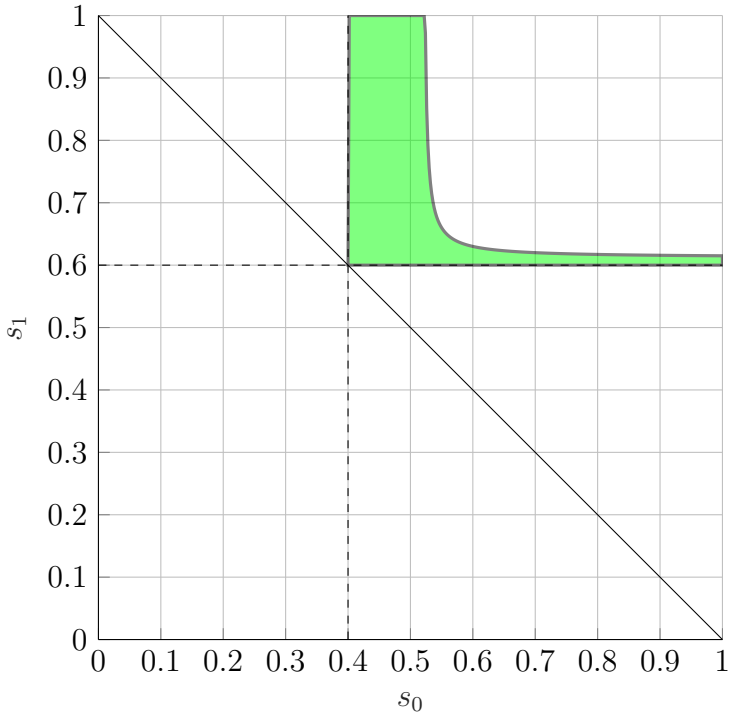


Figure 3.8. Dilator set $\mathcal{D}_P$ for the DGP in Table 3.1.

The dashed lines correspond to $s_0 = P(r = 0)$ and $s_1 = P(r = 1)$.

where

$$\mathcal{D}_P^{\llcorner} := \left\{ (s_0, s_1) \in [0,1]^2 \;\middle|\; \begin{array}{l} (1)\, \mathbb{E}_P\big[(\theta_1 - s_1)(r_i - 1 + s_0) - (s_1 - 1 + s_0)(t_i - 1)r_i\big] \geq 0, \\ (2)\, \mathbb{E}_P\big[(-\theta_1 + 1 - s_1)(r_i - 1 + s_0) + (s_1 - 1 + s_0)t_i r_i\big] \geq 0, \\ \quad \textit{where } \theta_1 = P(t = 1), \\ (3)\, \mathbb{E}_P\big[s_1 - r_i\big] \geq 0, \;\; \textit{and} \\ (4)\, \mathbb{E}_P\big[s_0 - 1 + r_i\big] \geq 0. \end{array} \right\}$$

and $S_0 := \{(s_0, s_1) \in [0,1]^2 \mid s_0 + s_1 = 1\}$.

The first two conditions are not moment functions, because $P(t = 1)$ is usually unknown. Nevertheless, these can be reformulated to include θ_1 as an additional parameter. Specifically, this means that

$$\left\{ (s_0, s_1, \theta_1) \in [0,1]^3 \;\middle|\; \begin{array}{l} (1)\, \mathbb{E}_P\big[(\theta_1 - s_1)(r_i - 1 + s_0) - (s_1 - 1 + s_0)(t_i - 1)r_i\big] \geq 0, \\ (2)\, \mathbb{E}_P\big[(-\theta_1 + 1 - s_1)(r_i - 1 + s_0) + (s_1 - 1 + s_0)t_i r_i\big] \geq 0, \\ (3)\, \mathbb{E}_P\big[s_1 - r_i\big] \geq 0, \\ (4)\, \mathbb{E}_P\big[s_0 - 1 + r_i\big] \geq 0, \text{ and} \\ (5)\, \mathbb{E}_P\big[\theta_1 - t_i\big] = 0. \end{array} \right\}$$

is isomorphic to $\mathcal{D}_P^{\perp\!\!\!\perp}$ and is defined using only moment (in)equalities.

Section 3.6. Conclusion and Policy Implications

In this paper we show that dilation is a real-world phenomenon which may be induced by tests recommended for use based on satisfactory "apparent" performance. Dilation may arise when imperfect references, which are ubiquitous in practice, are used to measure performance of novel diagnostic tests. Since dilation is an extreme form of non-informativeness, it may be particularly undesirable in the context of clinical decision making where clinicians may not wish to *only* increase ambiguity about their patient's health status. This further highlights the importance of considering true performance when evaluating novel tests, rather than following current norm of relying on "apparent" performance measures.

We conclude with a direct policy recommendation to regulators and researchers. We suggest verifying whether new diagnostic tests under evaluation are dilation-inducing

whenever an imperfect reference test is used. Doing so may reveal that seemingly well-performing tests may in fact be completely uninformative of the patient's true health status. To facilitate the verification, we provide an equivalent characterization of when a diagnostic test is dilation-inducing, develop an appropriately uniform hypothesis test, and outline methods for sensitivity analyses if the reference test performance may not be known.

Bibliography

Abakarim, Youness, Mohamed Lahby, and Abdelbaki Attioui. 2018. "Towards an efficient real-time approach to loan credit approval using deep learning." In *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC),* 306–313. IEEE.

Adjaye-Gbewonyo, Dzifa, Robert A Bednarczyk, Robert L Davis, and Saad B Omer. 2014. "Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study." *Health services research* 49 (1): 268–283.

Ai, Tao, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. 2020. "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases." *Radiology,* 200642.

Aizer, Anna, Nancy Early, Shari Eli, Guido Imbens, Keyoung Lee, Adriana Lleras-Muney, and Alexander Strand. 2024. "The Lifetime Impacts of the New Deal's Youth Employment Program." *The Quarterly Journal of Economics.*

Al-Khayyal, Faiz A. 1992. "Generalized bilinear programming: Part I. Models, applications and linear programming relaxation." *European Journal of Operational Research* 60 (3): 306–314.

Alcoba-Florez, Julia, Helena Gil-Campesino, Diego García-Martínez de Artola, Rafaela González-Montelongo, Agustín Valenzuela-Fernández, Laura Ciuffreda, and Carlos Flores. 2020. "Sensitivity of different RT-qPCR solutions for SARS-CoV-2 detection." *International Journal of Infectious Diseases* 99:190–192.

Aliprantis, C.D., and K.C. Border. 2006. *Infinite Dimensional Analysis: A Hitchhiker's Guide.* Springer Berlin Heidelberg. ISBN: 9783540295877. https://books.google.co.uk/books?id=4vyXtR3vUhoC.

Altman, Douglas G, and J Martin Bland. 1994. "Statistics Notes: Diagnostic tests 2: predictive values." *Bmj* 309 (6947): 102.

Andrews, Donald WK, and Panle Jia Barwick. 2012. "Inference for parameters defined by moment inequalities: A recommended moment selection procedure." *Econometrica* 80 (6): 2805–2826.

Andrews, Donald WK, and Xiaoxia Shi. 2013. "Inference based on conditional moment inequalities." *Econometrica* 81 (2): 609–666.

Andrews, Donald WK, and Gustavo Soares. 2010. "Inference for parameters defined by moment inequalities using generalized moment selection." *Econometrica* 78 (1): 119–157.

Arevalo-Rodriguez, Ingrid, Diana Buitrago-Garcia, Daniel Simancas-Racines, Paula Zambrano-Achig, Rosa Del Campo, Agustin Ciapponi, Omar Sued, Laura Martinez-Garcia, Anne W Rutjes, Nicola Low, et al. 2020. "False-negative results of initial RT-PCR assays for COVID-19: a systematic review." *PloS one* 15 (12): e0242958.

Artstein, Zvi. 1983. "Distributions of random sets and random selections." *Israel Journal of Mathematics* 46:313–324.

Athey, Susan, Raj Chetty, and Guido Imbens. 2020. "Combining experimental and observational data to estimate treatment effects on long term outcomes." *arXiv preprint arXiv:2006.09676.*

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2024. *Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index.* Technical report.

Attanasio, Orazio P, Costas Meghir, and Ana Santiago. 2012. "Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa." *The Review of Economic Studies* 79 (1): 37–66.

Aumann, Robert J. 1976. "Agreeing to Disagree." *The Annals of Statistics* 4 (6): 1236–1239. ISSN: 00905364, 21688966, accessed November 28, 2024. http://www.jstor.org/stable/2958591.

Bai, Yuehao, Andres Santos, and Azeem M Shaikh. 2021. "A Two-Step Method for Testing Many Moment Inequalities." *Journal of Business & Economic Statistics,* 1–11.

Baron, John A. 1994. "Uncertainty in bayes." *Medical Decision Making* 14 (1): 46–51.

Beauchêne, Dorian, Jian Li, and Ming Li. 2019. "Ambiguous persuasion." *Journal of Economic Theory* 179:312–365. ISSN: 0022-0531. https://doi.org/https://doi.org/

10.1016/j.jet.2018.10.008. https://www.sciencedirect.com/science/article/pii/S0022053118306653.

Beißner, Patrick, and M. Ali Khan. 2019. "On Hurwicz–Nash equilibria of non-Bayesian games under incomplete information." *Games and Economic Behavior* 115:470–490. ISSN: 0899-8256. https://doi.org/https://doi.org/10.1016/j.geb.2019.02.001. https://www.sciencedirect.com/science/article/pii/S0899825619300119.

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari. 2012. "Partial identification using random set theory." *Journal of Econometrics* 166 (1): 17–32.

Bhattacharya, Jay, Azeem M Shaikh, and Edward Vytlacil. 2012. "Treatment effect bounds: An application to Swan–Ganz catheterization." *Journal of Econometrics* 168 (2): 223–243.

Bianchi, Matt T, Brian M Alexander, and Sydney S Cash. 2009. "Incorporating uncertainty into medical decision making: an approach to unexpected test results." *Medical Decision Making* 29 (1): 116–124.

Binney, Nicholas, Christopher Hyde, and Patrick M Bossuyt. 2021. "On the Origin of Sensitivity and Specificity." *Annals of Internal Medicine* 174 (3): 401–407.

Bollinger, Christopher R, and Martin H David. 1997. "Modeling discrete choice with response error: Food Stamp participation." *Journal of the American Statistical Association* 92 (439): 827–835.

Bose, Subir, and Ludovic Renou. 2014. "Mechanism Design With Ambiguous Communication Devices." *Econometrica* 82 (5): 1853–1872. https://doi.org/https://doi.org/10.3982/ECTA10285. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA10285. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10285.

Boyko, Edward J, Beth W Alderman, and Anna E Baron. 1988. "Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease." *Journal of General Internal Medicine* 3 (5): 476–481.

Bradley, Seamus. 2019. "Imprecise Probabilities." In *The Stanford Encyclopedia of Philosophy,* Spring 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Buck, Alfred A, and John J Gart. 1966. "Comparison of a screening test and a reference test in epidemiologic studies: I. Indices of agreement and their relation to prevalence." *American Journal of Epidemiology* 83 (3): 586–592.

Budelier, Melissa M, and Randall J Bateman. 2020. "Biomarkers of Alzheimer disease." *The journal of applied laboratory medicine* 5 (1): 194–208.

Bugni, Federico A, Ivan A Canay, and Xiaoxia Shi. 2017. "Inference for subvectors and other functions of partially identified parameters in moment inequality models." *Quantitative Economics* 8 (1): 1–38.

———. 2015. "Specification tests for partially identified models defined by moment inequalities." *Journal of Econometrics* 185 (1): 259–282.

Canay, Ivan A, Gastón Illanes, and Amilcar Velez. 2023. "A User's guide for inference in models defined by moment inequalities." *Journal of Econometrics,* 105558.

Canay, Ivan A, and Azeem M Shaikh. 2017. "Practical and theoretical advances in inference for partially identified models." *Advances in Economics and Econometrics* 2:271–306.

Cannings, Timothy I, Yingying Fan, and Richard J Samworth. 2020. "Classification with imperfect training labels." *Biometrika* 107 (2): 311–330.

Caraguel, Charles GB, and Raphaël Vanderstichel. 2013. "The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation." *BMJ Evidence-Based Medicine* 18 (4): 125–128.

Carlotto, Mark J. 2009. "Effect of errors in ground truth on classification accuracy." *International Journal of Remote Sensing* 30 (18): 4831–4849.

Chen, Jiafeng, and David M Ritzwoller. 2023. "Semiparametric estimation of long-term treatment effects." *Journal of Econometrics* 237 (2): 105545.

Chen, Xuan, Carlos A Flores, and Alfonso Flores-Lagunes. 2018. "Going beyond LATE: bounding average treatment effects of Job Corps training." *Journal of Human Resources* 53 (4): 1050–1099.

Cheng, Xiaoyu. 2023. "Ambiguous Persuasion: An Ex-Ante Formulation." *Working paper,* arXiv: 2010.05376 [econ.TH]. https://arxiv.org/abs/2010.05376.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2019. "Inference on causal and structural parameters using many moment inequalities." *The Review of Economic Studies* 86 (5): 1867–1900.

Chernozhukov, Victor, Sokbae Lee, and Adam M Rosen. 2013. "Intersection bounds: estimation and inference." *Econometrica* 81 (2): 667–737.

Chesher, Andrew, and Adam M Rosen. 2017. "Generalized instrumental variable models." *Econometrica* 85 (3): 959–989.

———. 2020. "Generalized instrumental variable models, methods, and applications." In *Handbook of Econometrics,* 7:1–110. Elsevier.

Cross, Philip J, and Charles F Manski. 2002. "Regressions, short and long." *Econometrica* 70 (1): 357–368.

Currie, Janet, and Douglas Almond. 2011. "Human capital development before age five." In *Handbook of labor economics,* 4:1315–1486. Elsevier.

Deaton, Angus S. 2009. *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development.* Technical report. National bureau of economic research.

Deneef, Peter. 1987. "Evaluating rapid tests for streptococcal pharyngitis: the apparent accuracy of a diagnostic test when there are errors in the standard of comparison." *Medical Decision Making* 7 (2): 92–96.

Diggle, Peter J. 2011. "Estimating prevalence using an imperfect test." *Epidemiology Research International* 2011.

Dominiak, Adam, Matthew Kovach, and Gerelt Tserenjigmid. 2022. *Minimum distance belief updating with general information.* Technical report. Working paper.

Dramé, Moustapha, Maturin Tabue Teguo, Emeline Proye, Fanny Hequet, Maxime Hentzien, Lukshe Kanagaratnam, and Lidvine Godaert. 2020. "Should RT-PCR be considered a gold standard in the diagnosis of Covid-19?" *Journal of Medical Virology.*

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie Van Dijk. 2021. *Selection in surveys: Using randomized incentives to detect and account for nonresponse bias.* Technical report. National Bureau of Economic Research.

ECDC. 2021. "Options for the Use of Rapid Antigen Detection Tests for COVID-19 in the EU/EEA—First Update." *ECDC Technical Report.*

Elliott, Marc N, Allen Fremont, Peter A Morrison, Philip Pantoja, and Nicole Lurie. 2008. "A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity." *Health services research* 43 (5p1): 1722–1736.

Elzayn, Hadi, Evelyn Smith, Thomas Hertz, Arun Ramesh, Jacob Goldin, Daniel E Ho, and Robin Fisher. 2023. *Measuring and mitigating racial disparities in tax audits.* Stanford Institute for Economic Policy Research (SIEPR).

Emerson, Sarah C, Sushrut S Waikar, Claudio Fuentes, Joseph V Bonventre, and Rebecca A Betensky. 2018. "Biomarker validation with an imperfect reference: Issues and bounds." *Statistical methods in medical research* 27 (10): 2933–2945.

Epstein, Larry G, and Yoram Halevy. 2024. "Hard-to-interpret signals." *Journal of the European Economic Association* 22 (1): 393–427.

Fagan, Terrence J. 1975. "Nomogram for Bayes's theorem." *The New England journal of medicine* 293 (5): 257–257.

Fitzpatrick, Meagan C, Abhishek Pandey, Chad R Wells, Pratha Sah, and Alison P Galvani. 2021. "Buyer beware: inflated claims of sensitivity for rapid COVID-19 tests." *The Lancet* 397 (10268): 24–25.

Foody, Giles M. 2010. "Assessing the accuracy of land cover change with imperfect ground reference data." *Remote Sensing of Environment* 114 (10): 2271–2285.

———. 2023. "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient." *Plos one* 18 (10): e0291908.

Galichon, Alfred. 2018. *Optimal transport methods in economics.* Princeton University Press.

Galichon, Alfred, and Marc Henry. 2011. "Set identification in models with multiple equilibria." *The Review of Economic Studies* 78 (4): 1264–1298.

García, Jorge Luis, James J Heckman, Duncan Ermini Leaf, and María José Prados. 2020. "Quantifying the life-cycle benefits of an influential early-childhood program." *Journal of Political Economy* 128 (7): 2502–2541.

Gart, John J, and Alfred A Buck. 1966. "Comparison of a Screening Test and a Reference Test in Epidemiologic Studies: A Probabilistic Model for the Comparison of Diagnostic Tests." *American Journal of Epidemiology* 83 (3): 593–602.

Gastwirth, Joseph L. 1987. "The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data." *Statistical Science* 2 (3): 213–222.

Ghassami, AmirEmad, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. 2022. "Combining experimental and observational data for identification and estimation of long-term causal effects." *arXiv preprint arXiv:2201.10743*.

Gong, Ruobin, and Xiao-Li Meng. 2021. "Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox." *Statistical Science* 36 (2): 169–190.

Good, Irving John. 1974. "A little learning can be dangerous." *The British Journal for the Philosophy of Science* 25 (4): 340–342.

Goodman, Leo A. 1965. "On simultaneous confidence intervals for multinomial proportions." *Technometrics* 7 (2): 247–254.

Green, Daniel A, and Kirsten StGeorge. 2018. "Rapid antigen tests for influenza: rationale and significance of the FDA reclassification." *Journal of Clinical Microbiology* 56 (10): 10–1128.

Greenland, Sander. 1996. "Basic methods for sensitivity analysis of biases." *International journal of epidemiology* 25 (6): 1107–1116.

Groemping, Ulrike. 2019. "South german credit data: Correcting a widely used data set." *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep* 4:2019.

Grunwald, Peter D, and Joseph Y Halpern. 2004. "When ignorance is bliss." In *Proceedings of the Twentieth Conference on Uncertainty in AI,* 226–234.

Gul, Faruk, and Wolfgang Pesendorfer. 2021. "Evaluating ambiguous random variables from Choquet to maxmin expected utility." *Journal of Economic Theory* 192:105129. ISSN: 0022-0531. https://doi.org/https://doi.org/10.1016/j.jet.2020.105129. https://www.sciencedirect.com/science/article/pii/S0022053120301228.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. "Top challenges from the first practical online controlled experiments summit." *ACM SIGKDD Explorations Newsletter* 21 (1): 20–35.

Hadgu, Alula. 1999. "Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity." *Journal of clinical epidemiology* 52 (12): 1231–1237.

Hayashi, Takashi, and Ryoko Wada. 2010. "Choice with imprecise information: an experimental approach." *Theory and Decision* 69 (3): 355–373. https://doi.org/10.1007/s11238-008-9119-x. https://ideas.repec.org/a/kap/theord/v69y2010i3p355-373.html.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes." *American Economic Review* 103 (6): 2052–2086.

Heckman, James J, and Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156 (1): 27–37.

Heckman, James J, and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences* 96 (8): 4730–4734.

Hernández-Huerta, María Teresa, Laura Pérez-Campos Mayoral, Luis Manuel Sánchez Navarro, Gabriel Mayoral-Andrade, Eduardo Pérez-Campos Mayoral, Edgar Zenteno, and Eduardo Pérez-Campos. 2020. "Should RT-PCR be considered a gold standard in the diagnosis of Covid-19?" *Journal of Medical Virology.*

Herron, Timothy, Teddy Seidenfeld, and Larry Wasserman. 1997. "Divisive conditioning: further results on dilation." *Philosophy of Science* 64 (3): 411–444.

Hiriart-Urruty, J-B. 1985. "Images of connected sets by semicontinuous multifunctions." *Journal of Mathematical Analysis and Applications* 111 (2): 407–422.

Hofmann, Hans. 1994. *Statlog (German Credit Data).* UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Hoynes, Hilary W, and Diane Whitmore Schanzenbach. 2018. *Safety net investments in children.* Technical report. National Bureau of Economic Research.

Hu, Wenjie, Xiaohua Zhou, and Peng Wu. 2022. "Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data." *arXiv preprint arXiv:2208.10163.*

Hui, Siu L, and Xiao H Zhou. 1998. "Evaluation of diagnostic tests without gold standards." *Statistical methods in medical research* 7 (4): 354–370.

Hui, Sui L, and Steven D Walter. 1980. "Estimating the error rates of diagnostic tests." *Biometrics,* 167–171.

Imai, Kosuke, and Kabir Khanna. 2016. "Improving ecological inference by predicting individual ethnicity from voter registration records." *Political Analysis* 24 (2): 263–272.

Imbens, Guido, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2024. "Long-term causal inference under persistent confounding via data combination." *arXiv preprint arXiv:2202.07234.*

Imbens, Guido W. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic literature* 48 (2): 399–423.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475.

Joe, Harry. 1997. *Multivariate models and multivariate dependence concepts.* CRC press.

Kaido, Hiroaki, Francesca Molinari, and Jörg Stoye. 2019. "Confidence intervals for projections of partially identified parameters." *Econometrica* 87 (4): 1397–1432.

Kamat, Vishal. 2024. "Identifying the effects of a program offer with an application to head start." *Journal of Econometrics* 240 (1): 105679.

Kanji, Jamil N, Nathan Zelyas, Clayton MacDonald, Kanti Pabbaraju, Muhammad Naeem Khan, Abhaya Prasad, Jia Hu, Mathew Diggle, Byron M Berenger, and Graham Tipples. 2021. "False negative rate of COVID-19 PCR testing: a discordant testing analysis." *Virology journal* 18 (1): 1–6.

Kellner, Christian, Mark T Le Quement, and Gerhard Riener. 2022. "Reacting to ambiguous messages: An experimental analysis." *Games and Economic Behavior* 136:360–378.

Kline, Patrick, and Christopher R Walters. 2016. "Evaluating public programs with close substitutes: The case of Head Start." *The Quarterly Journal of Economics* 131 (4): 1795–1848.

Kohn, Michael A. 2022. "Studies of diagnostic test accuracy: partial verification bias and test result-based sampling." *Journal of Clinical Epidemiology* 145:179–182.

Kops, Christopher, and Illia Pasichnichenko. 2023. "Testing negative value of information and ambiguity aversion." *Journal of Economic Theory,* 105730.

Kreider, Brent, and John V Pepper. 2007. "Disability and employment: Reevaluating the evidence in light of reporting errors." *Journal of the American Statistical Association* 102 (478): 432–441.

Kucirka, Lauren M, Stephen A Lauer, Oliver Laeyendecker, Denali Boon, and Justin Lessler. 2020. "Variation in false-negative rate of reverse transcriptase polymerase

chain reaction–based SARS-CoV-2 tests by time since exposure." *Annals of Internal Medicine.*

Kyburg, Henry E. 1977. "Randomness and the right reference class." *The Journal of Philosophy* 74 (9): 501–521.

Lehmann, Erich L, and Joseph P Romano. 2022. *Testing Statistical Hypotheses.* Springer Science & Business Media.

Levi, Isaac. 1977. "Direct inference." *The journal of philosophy* 74 (1): 5–29.

Levy, Gilat, and Ronny Razin. 2022. "Combining forecasts in the presence of ambiguity over correlation structures." Symposium Issue on Ambiguity, Robustness, and Model Uncertainty, *Journal of Economic Theory* 199:105075. ISSN: 0022-0531. https://doi.org/https://doi.org/10.1016/j.jet.2020.105075. https://www.sciencedirect.com/science/article/pii/S0022053118303144.

Li, Shoumei, and Yukio Ogura. 1998. "Convergence of set valued sub-and supermartingales in the Kuratowski-Mosco sense." *Annals of probability,* 1384–1402.

Liang, Yucheng. 2024. "Learning from unknown information sources." *Management Science.*

Lin, Yi-Hsuan, and Fernando Payró. 2024. *Updating Under Imprecise Information.* Technical report. Working paper.

Luo, Ye, and Hai Wang. 2018. "Identifying and computing the exact core-determining class." *Available at SSRN 3154285.*

Manski, Charles F. 2020. "Bounding the accuracy of diagnostic tests, with application to COVID-19 antibody tests." *Epidemiology* 32 (2): 162–167.

———. 2021. "Bounding the accuracy of diagnostic tests, with application to COVID-19 antibody tests." *Epidemiology* 32 (2): 162–167.

———. 2018. "Credible ecological inference for medical decisions with personalized risk assessment." *Quantitative Economics* 9 (2): 541–569.

———. 2007. *Identification for Prediction and Decision.* Harvard University Press.

———. 1997. "Monotone treatment response." *Econometrica: Journal of the Econometric Society,* 1311–1334.

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

———. 2003. *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, Charles F, and Francesca Molinari. 2021. "Estimating the COVID-19 infection rate: Anatomy of an inference problem." *Journal of Econometrics* 220 (1): 181–192.

Manski, Charles F, and John V Pepper. 2000a. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4): 997–1010.

Manski, Charles F., and John V. Pepper. 2000b. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4): 997–1010. https://doi.org/https://doi.org/10.1111/1468-0262.t01-1-00144a. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.t01-1-00144a. https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.t01-1-00144a.

Manski, Charles F, and John V Pepper. 2009. "More on monotone instrumental variables." *The Econometrics Journal* 12 (suppl_1): S200–S216.

Masten, Matthew A, and Alexandre Poirier. 2023. "Choosing exogeneity assumptions in potential outcome models." *The Econometrics Journal* 26 (3): 327–349.

Mathews, William C, Edward R Cachay, Joseph Caperna, Amy Sitapati, Bard Cosman, and Ian Abramson. 2010. "Estimating the accuracy of anal cytology in the presence of an imperfect reference standard." *PloS one* 5 (8): e12284.

Matos, R, TF Novaes, Mariana Minatel Braga, WL Siqueira, DA Duarte, and Fausto Medeiros Mendes. 2011. "Clinical performance of two fluorescence-based methods in detecting occlusal caries lesions in primary teeth." *Caries Research* 45 (3): 294–302.

McNeil, Barbara J, Emmett Keeler, and S James Adelstein. 1975. "Primer on certain elements of medical decision making." *New England Journal of Medicine* 293 (5): 211–215.

Meyer, Bruce D, Nikolas Mittag, and Robert M Goerge. 2022. "Errors in survey reporting and imputation and their effects on estimates of food stamp program participation." *Journal of Human Resources* 57 (5): 1605–1644.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica* 86 (5): 1589–1619.

Molchanov, Ilya. 2017. *Theory of Random Sets.* 2nd ed. Vol. 87. Probability Theory and Stochastic Modelling. Springer.

Molchanov, Ilya, and Francesca Molinari. 2014. "Applications of random set theory in econometrics." *Annu. Rev. Econ.* 6 (1): 229–251.

———. 2018. *Random Sets in Econometrics.* Vol. 60. Cambridge University Press.

Molinari, Francesca. 2008. "Partial identification of probability distributions with misclassified data." *Journal of Econometrics* 144 (1): 81–117.

Moon, Sarah. 2024. "Partial Identification of Individual-Level Parameters Using Aggregate Data in a Nonparametric Binary Outcome Model." *arXiv preprint arXiv:2403.07236.*

Mulherin, Stephanie A, and William C Miller. 2002a. "Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation." *Annals of internal medicine* 137 (7): 598–602.

———. 2002b. "Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation." *Annals of internal medicine* 137 (7): 598–602.

Munkres, James R. 2000. *Topology.* Pearson Education.

Obradović, Filip. 2024. "Measuring diagnostic test performance using imperfect reference tests: A partial identification approach." *Journal of Econometrics* 244 (1): 105842. ISSN: 0304-4076. https://doi.org/https://doi.org/10.1016/j.jeconom.2024.105842. https://www.sciencedirect.com/science/article/pii/S0304407624001878.

Pacheco Pires, Cesaltina. 2002. "A rule for updating ambiguous beliefs." *Theory and Decision* 53:137–152.

Pahlke, Marieke. 2024. "Dynamic Consistency and Ambiguous Communication." *Working paper.*

Park, Yechan, and Yuya Sasaki. 2024a. "A Bracketing Relationship for Long-Term Policy Evaluation with Combined Experimental and Observational Data." *arXiv preprint arXiv:2401.12050.*

———. 2024b. "The Informativeness of Combined Experimental and Observational Data under Dynamic Selection." *arXiv preprint arXiv:2403.16177.*

Patwardhan, Meenal B, Douglas C McCrory, David B Matchar, Gregory P Samsa, and Olivier T Rutschmann. 2004. "Alzheimer disease: operating characteristics of PET—a meta-analysis." *Radiology* 231 (1): 73–80.

Pollock, Nira R, Jesica R Jacobs, Kristine Tran, Amber E Cranston, Sita Smith, Claire Y O'Kane, Tyler J Roady, Anne Moran, Alison Scarry, Melissa Carroll, et al. 2021. "Performance and implementation evaluation of the Abbott BinaxNOW rapid antigen test in a high-throughput drive-through community testing site in Massachusetts." *Journal of Clinical Microbiology* 59 (5): e00083–21.

Ponomarev, Kirill. 2024. "Selecting Inequalities for Sharp Identification in Models with Set-Valued Predictions," http://kponomarev.github.io/files_on_website/sharp%20inequalities.pdf.

Poterba, James M, and Lawrence H Summers. 1986. "Reporting errors and labor market dynamics." *Econometrica: Journal of the Econometric Society,* 1319–1338.

Prentice, Ross L. 1989. "Surrogate endpoints in clinical trials: definition and operational criteria." *Statistics in medicine* 8 (4): 431–440.

Ridder, Geert, and Robert Moffitt. 2007. "The econometrics of data combination." *Handbook of econometrics* 6:5469–5547.

Rockafellar, Ralph Tyrell. 1970. *Convex Analysis.* Princeton: Princeton University Press. ISBN: 9781400873173. https://doi.org/doi:10.1515/9781400873173. https://doi.org/10.1515/9781400873173.

Rogan, Walter J, and Beth Gladen. 1978. "Estimating prevalence from the results of a screening test." *American journal of epidemiology* 107 (1): 71–76.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf. 2014. "A practical two-step method for testing moment inequalities." *Econometrica* 82 (5): 1979–2002.

Russell, Thomas M. 2021. "Sharp bounds on functionals of the joint distribution in the analysis of treatment effects." *Journal of Business & Economic Statistics* 39 (2): 532–546.

Sacks, Daniel W, Nir Menachemi, Peter Embi, and Coady Wing. 2022. "What can we learn about SARS-CoV-2 prevalence from testing and hospital data?" *Review of Economics and Statistics,* 1–36.

Schaefer, Helmut H., and M. P. Wolff. 1999. *Topological Vector Spaces.* Springer.

Seidenfeld, Teddy, and Larry Wasserman. 1993. "Dilation for sets of probabilities." *The Annals of Statistics* 21 (3): 1139–1154.

Shah, Melisa M, Phillip P Salvatore, Laura Ford, Emiko Kamitani, Melissa J Whaley, Kaitlin Mitchell, Dustin W Currie, Clint N Morgan, Hannah E Segaloff, Shirley Lecher, et al. 2021. "Performance of Repeat BinaxNOW Severe Acute Respiratory Syndrome Coronavirus 2 Antigen Testing in a Community Setting, Wisconsin, November 2020–December 2020." *Clinical Infectious Diseases* 73 (Supplement_1): S54–S57.

Shea, Joshua. 2022. "Testing for racial bias in police traffic searches." *University of Illinois, Champaign Urbana, USA.*

Shi, Xiaoxia, and Matthew Shum. 2015. "Simple two-stage inference for a class of partially identified models." *Econometric Theory* 31 (3): 493–520.

Shishkin, Denis, and Pietro Ortoleva. 2023. "Ambiguous information and dilation: An experiment." *Journal of Economic Theory* 208:105610.

Siddiqui, Zishan K, Mihir Chaudhary, Matthew L Robinson, Anna B McCall, Ria Peralta, Rogette Esteve, Charles W Callahan, Yukari C Manabe, James D Campbell, J Kristie Johnson, et al. 2021. "Implementation and Accuracy of BinaxNOW Rapid Antigen COVID-19 Test in Asymptomatic and Symptomatic Populations in a High-Volume Self-Referred Testing Site." *Microbiology spectrum* 9 (3): e01008–21.

Srinivasan, Preethi, M Brandon Westover, and Matt T Bianchi. 2012. "Propagation of uncertainty in Bayesian diagnostic test interpretation." *Southern medical journal* 105 (9): 452.

Staquet, Maurice, Marcel Rozencweig, Young Jack Lee, and Franco M Muggia. 1981. "Methodology for the assessment of new dichotomous diagnostic tests." *Journal of chronic diseases* 34 (12): 599–610.

Stoye, Jörg. 2022. "Bounding infection prevalence by bounding selectivity and accuracy of tests: with application to early COVID-19." *The Econometrics Journal* 25 (1): 1–14.

Suemoto, Claudia K, and Renata EP Leite. 2023. "Autopsy studies are key to identifying dementia cause." *The Lancet Healthy Longevity* 4 (3): e94–e95.

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

Thibodeau, LA. 1981. "Evaluating diagnostic tests." *Biometrics,* 801–804.

Todd, Petra E, and Kenneth I Wolpin. 2006. "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility." *American economic review* 96 (5): 1384–1417.

———. 2023. "The best of both worlds: combining randomized controlled trials with structural modeling." *Journal of Economic Literature* 61 (1): 41–85.

Torgovitsky, Alexander. 2019. "Nonparametric inference on state dependence in unemployment." *Econometrica* 87 (5): 1475–1505.

Toulis, Panos. 2021. "Estimation of COVID-19 prevalence from serology tests: A partial identification approach." *Journal of Econometrics* 220 (1): 193–213.

Treves, François. 2016. *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics, Vol. 25.* Vol. 25. Elsevier.

Vaart, Aad W Van der. 2000. *Asymptotic statistics.* Vol. 3. Cambridge university press.

Vacek, Pamela M. 1985. "The effect of conditional dependence on the evaluation of diagnostic tests." *Biometrics,* 959–968.

Valenstein, Paul N. 1990. "Evaluating diagnostic tests with imperfect standards." *American Journal of Clinical Pathology* 93 (2): 252–258.

Van Der Vaart, Aad W, and Jon A Wellner. 1997. *Weak convergence and empirical processes: with applications to statistics.* Springer New York.

Van Goffrier, Graham, Lucas Maystre, and Ciarán Mark Gilligan-Lee. 2023. "Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding." In *Conference on Causal Learning and Reasoning,* 791–813. PMLR.

Vikström, Johan, Geert Ridder, and Martin Weidner. 2018. "Bounds on treatment effects on transitions." *Journal of Econometrics* 205 (2): 448–469.

Villani, Cédric, et al. 2009. *Optimal transport: old and new.* Vol. 338. Springer.

Walley, Peter. 1991. *Statistical reasoning with imprecise probabilities.* Vol. 42. Springer.

Walter, Steven D, and Les M Irwig. 1988. "Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review." *Journal of clinical epidemiology* 41 (9): 923–937.

Wang, Xiaonan, Fengjie Li, Jiang Tian, Qi Gao, and Huiping Zhu. 2023. "Bayesian estimation for the accuracy of three neuropsychological tests in detecting Alzheimer's disease and mild cognitive impairment: a retrospective analysis of the ADNI database." *European Journal of Medical Research* 28 (1): 1–9.

Wang, Zichang. 2024. "Informativeness orders over ambiguous experiments." *Journal of Economic Theory* 222:105937. ISSN: 0022-0531. https://doi.org/https://doi.org/10.1016/j.jet.2024.105937. https://www.sciencedirect.com/science/article/pii/S0022053124001431.

Watson, Jessica, Penny F Whiting, and John E Brush. 2020a. "Interpreting a covid-19 test result." *Bmj* 369.

———. 2020b. "Interpreting a COVID-19 Test Result." *Bmj* 369.

Weissman, Joel S, and Romana Hasnain-Wynia. 2011. "Advancing health care equity through improved data collection." *New England Journal of Medicine* 364 (24): 2276–2277.

Whiting, Penny F, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, Patrick MM Bossuyt, and QUADAS-2 Group*. 2011. "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies." *Annals of internal medicine* 155 (8): 529–536.

Willard, Stephen. 2004. *General topology.* Courier Corporation.

Willis, Brian H. 2008a. "Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies." *Family Practice* 25 (5): 390–396.

———. 2008b. "Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies." *Family Practice* 25 (5): 390–396.

Yerushalmy, Jacob. 1947. "Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques." *Public Health Reports (1896-1970),* 1432–1449.

Yildiz, Neşe. 2012. "Consistency of plug-in estimators of upper contour and level sets." *Econometric Theory* 28 (2): 309–327.

Zhang, Jiji, Hailin Liu, and Teddy Seidenfeld. 2018. "Agreeing to disagree and dilation." *International Journal of Approximate Reasoning* 101:150–162. ISSN: 0888-613X. https:

//doi.org/https://doi.org/10.1016/j.ijar.2018.07.006. https://www.sciencedirect.com/science/article/pii/S0888613X17305893.

Zhou, Xiao-Hua, Donna K McClish, and Nancy A Obuchowski. 2009. *Statistical methods in diagnostic medicine.* Vol. 569. John Wiley & Sons.

Ziegler, Gabriel. 2021. "Binary Classification Tests, Imperfect Standards, and Ambiguous Information," arXiv: 2012.11215 [econ.EM].

APPENDIX A

Appendix to Chapter 1

Section A.1. Extensions

Subsection A.1.1. Additional Results on the Roles of Data and Assumptions

This appendix collects complementary results for the discussion in Section 1.4.

A.1.1.1. Proposition 1.4.1 and Existing Bounds. Suppose first that no modeling assumptions are maintained.

Corollary A.1.1. *Suppose Assumptions RA and EV hold. If* $\mathcal{Y} = \mathbb{R}$*, the identified set for* τ *is* $\mathcal{H}(\tau) = \mathbb{R}$*. If* $\mathcal{Y} = [0, 1]$*:*

(A.1)

$$\mathcal{H}(\tau) = \left[E_O[YD] - E_O[Y(1-D)] - P_O(D=0), E_O[YD] - E_O[Y(1-D)] + P_O(D=1)\right].$$

In both cases, $0 \in \mathcal{H}(\tau)$ *and the sign of* τ *not identified.*

Corollary A.1.1 shows that if $\mathcal{Y}$ is unbounded and no modeling assumptions are imposed, then τ is unidentified. If the support is bounded, data combination reproduces bounds of Manski (1990), which utilize only the observational dataset. The bounds remain sharp even when the experimental dataset is added since it brings no identifying power, *on its own*.

Athey et al. (2024, Lemmas 1 and 2) provide bounds on long-term treatment effects in a different setting where D is unobserved in the observational data and experimental compliance is perfect.[1] Their bounds may be narrower than those in Corollary A.1.1, and do not maintain explicit modeling assumptions involving the potential outcomes. However, this does not contradict the result in Proposition 1.4.1. Namely, their bounds are derived under assumptions imposed on outcome variables: 1) $Y \perp\!\!\!\perp D|S, G = E$ (statistical surrogacy - Prentice (1989)); 2) $G \perp\!\!\!\perp Y|S$ (comparability). Appendix A.1.1.4 explains that these assumptions on outcomes imply underlying selection assumptions.

A.1.1.2. Assumption LUC and the Role of Experimental Data. Recall that $\mathcal{H}^O(\tau)$ the identified set for τ when only observational data are used and no modeling assumptions are imposed, and let $\mathcal{H}^{O/LUC}(\tau)$ denote the identified set under Assumption LUC. Finally, let $\mathcal{H}(\tau)$ be the identified set when combined data are used under Assumption LUC.

Proposition A.1.1. *Let Assumptions EV and LUC hold.*

i) Suppose the observed data distribution $P_O(Y, S, D)$ is such that $V_O[Y|S, D = d] > 0$ $P-$a.s. for some $d \in \{0, 1\}$ and that $\mathcal{Y}$ is a bounded set. Then $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.

ii) If the observed data distribution $P_O(Y, S, D)$ is such that $E_O[Y|S, D = d]$ is a trivial measurable function for all $d \in \{0, 1\}$, then τ is point-identified, and $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$.

A few observations are in order. First, the proposition shows that $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$ is possible. That is, LUC may have identifying power for τ for a large class of observable

1. More precisely, they bound $E_E[Y(1) - Y(0)]$. These bounds remain valid for τ when Assumption EV is imposed.

distributions $P_O(Y, S, D)$ even when experimental data are not used. A sufficient condition for this is that Y is bounded, and that S is not a perfect predictor of Y for at least some $D = d$.

Second, Athey, Chetty, and Imbens (2020) show that $\mathcal{H}(\tau)$ is a singleton under combined data and LUC. Since $\mathcal{H}^{O/LUC}(\tau)$ need not be a singleton, we usually have $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/LUC}$. Consequently, experimental data may *amplify* the identifying power of LUC.

Third, the proposition shows that $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$ is possible. That is, short-term experimental data are not necessary for point identification of τ under LUC. Thus, experimental data do not *necessarily* amplify the identifying power of LUC. This intuitively happens when the short-run outcomes S are not predictive of the mean long-term outcomes Y.[2] This condition is strong and may lack practical applicability. However, the result has important theoretical implications in clarifying the role of the experimental data.

A.1.1.3. An Example of Nested Misspecification. Section 1.4 explains that the amplifying role of experimental data has important implications when the modeling assumption fails. Then, adding experimental data may only produce identified sets for τ that are weakly farther away from the truth. Recalling the notation, $\tilde{\mathcal{H}}$ and $\tilde{\mathcal{H}}^{O/A}$ denote misspecified identified sets for τ using combined and just observational data. The following example shows that under a standard modeling assumption and a non-pathological data-generating process, $\tilde{\mathcal{H}}^{O/A}$ can be strictly closer to τ than $\tilde{\mathcal{H}}$. Moreover, $\tilde{\mathcal{H}}^{O/A}$ is informative of the sign of τ.

2. Observe that no restrictions on $\mathcal{Y}$ are required in this case.

Example A.1.1. Suppose $Y, S \in \{0,1\}$ and that the researcher maintains Assumption LUC. Let the DGP be given by:

$$E_O[Y|S=1,D=1] = 0.7 \qquad E_O[Y|S=0,D=1] = 0.4$$

$$E_O[Y|S=1,D=0] = 0.4 \qquad E_O[Y|S=0,D=0] = 0.2$$

$$E[Y(1)|S(1)=1] = 0.5 \qquad E[Y(0)|S(0)=1] = 0.3$$

$$E[Y(1)|S(1)=0] = 0.5 \qquad E[Y(0)|S(0)=0] = 0.3$$

$$P_O[S=1|D=1] = 0.6 \qquad P_O[S=1|D=0] = 0.4$$

$$P[S(1)=1] = 0.7 \qquad P[S(0)=1] = 0.3$$

$$P[D=1] = 0.5$$

Then $\tau = 0.2$, $\tilde{\mathcal{H}}^{O/A} = [0.15, 0.4]$ and $\tilde{\mathcal{H}} = \{0.35\}$.

A.1.1.4. More on Treatment Invariance and Surrogacy. By Lemma A.2.9 *ii)*, Assumption TI is implied by surrogacy when the experiment features perfect compliance. One may thus wish to intuitively interpret TI as stating that the treatment effect on the long-term outcome is fully mediated by the short-term outcome, an interpretation commonly used for the surrogacy assumption. However, surrogacy imposes *selection assumptions* when compliance is imperfect. Then it is immediate that by surrogacy $E_E[Y(1)|S(1)=s, D=1] = E_E[Y(0)|S(0)=s, D=0]$ for $s \in \mathcal{S}$. This is an a priori restriction on the selection mechanism of experimental individuals, because $Y(d)$ are never observed for $G = E$. On the other hand, TI is always a treatment response assumption.

Work relying on surrogacy for identification, such as Athey et al. (2024), commonly also maintains – $G \perp\!\!\!\perp Y|S$ (comparability). Comparability and surrogacy jointly imply a selection assumption even if compliance is perfect. Note that for any $s \in \mathcal{S}$ and $d \in \{0,1\}$:

$$\begin{aligned} E[Y(d)|S(d)=s] &= E_O[Y(d)|S(d), D=d]P_O(D=d|S(d)=s) \\ &\quad + E_O[Y(d)|S(d), D\neq d]P_O(D\neq d|S(d)=s) \\ E[Y(d)|S(d)=s] &= E_O[Y(1)|S(1)=s, D=1]P_O(D=1|S=s) \\ &\quad + E_O[Y(0)|S(0)=s, D=0]P_O(D=0|S=s) \end{aligned}$$

where the first identity is by the law of iterated expectations (LIE) and the second is by Lemma A.2.9 *vi)* and LIE. Therefore, for any s and d such that $P(D \neq d, S(d) = s) > 0$ by rearranging terms:

$$\begin{aligned} &E_O[Y(d)|S(d)=s, D\neq d] = \\ &= \frac{E_O[Y|S, D=d]\left(P_O(D=d|S=s) - P_O(D=d|S(d)=s)\right)}{P_O(D\neq d|S(d)=s)} \\ &+ \frac{E_O[Y|S, D\neq d]P_O(D\neq d|S=s)}{P_O(D\neq d|S(d)=s)} \end{aligned}$$

which relates $(Y(1), Y(0), S(1), S(0))$ and D in the observational data, and is hence a *selection assumption*.

Subsection A.1.2. Discretization of Short-term Outcomes

In this section, I clarify the implications of discretizing short-term outcomes. To this end, let a researcher pose a surjective discretization function $\lambda : \mathcal{S} \to \mathcal{S}^D := \{1, 2, \ldots, k\}$ for

some $k < \infty$, and define $S^D(d) = \lambda(S(d))$. Note that this subsumes the case in which $S(d)$ is finitely supported, since then $\lambda(s) = s$ for all $s \in \mathcal{S}$. I introduce λ to clarify the subtle differences in applications of results of Section 1.3.3 when $S(d)$ is finitely supported and discretized. Similarly define discretized temporal link functions $m_d^D : \mathcal{S}^D \to \mathcal{Y}$, given by $m_d^D = E[Y(d)|S^D(d)] = E[Y(d)|\lambda(S(d))]$, and let $m^D = (m_0^D, m_1^D)$. Pose the following analog of Assumption MA under the discretization.

Assumption MA:D. *Suppose $\mathcal{M}^A$ and $\mathcal{M}^D$ are known or identified sets, and that $m \in \mathcal{M}^A \subseteq \mathcal{M}$. Then λ is such that $m^D \in \mathcal{M}^D$.*

Assumptions MA and MA:D are closely related. The former maintains that the researcher imposes some modeling assumption that will restrict feasible m, as in Section 1.2.3. The latter strengthens this notion and assumes that additionally m^D satisfies known restrictions after discretization. Of course, if Assumption MA holds for a finitely supported $S(d)$, then Assumption MA:D trivially follows by taking λ to be an identity function up with necessary relabeling of $S(d)$ values, if any. The remark below explains that for some modeling assumptions and discretization functions, MA:D follows immediately from MA, but that it may be restrictive for others.

Remark A.1.1. Consider LIV which states that $E[Y(d)|S(d) = s]$ is in $\mathcal{M}^A$ which contains only non-decreasing temporal link functions. Then $E[Y(d)|S^D = s]$ must also be non-decreasing for any order-preserving λ, so Assumption MA:D holds for an appropriately chosen λ. However, LUC states that $m_d(s) = E_O[Y|S = s, D = d]$, which does not directly imply that $m_d^D(s) = E[Y|S^D = s, D = d]$. A similar remark can be made for treatment invariance.

If $S(d)$ is finitely supported, MA and MA:D are equivalent and Section 1.3.3 characterizes the identified set. If $S(d)$ is discretized and Assumption MA:D holds as a direct consequence of Assumption MA, such as under LIV, then results characterize the identified set $\mathcal{H}(\tau)$ that is sharp *under finitely-supported short-term outcomes*.[3] This is also the case if the researcher believes the modeling assumption holds under discretized data, i.e., is willing to maintain MA:D directly. Otherwise, the results in Section 1.3.3 should be viewed as providing an approximation of the identified set.

Section A.2. Proofs

This section contains the proofs of all results. It begins by summarizing notation. Appendix A.2.2 collects known supporting results. Appendix A.2.3 contains auxiliary results and their proofs. Appendix A.2.4 provides proofs to the main results.

Subsection A.2.1. Preliminaries and Notation

I denote laws of random elements using subscripts when the element needs to be specified (e.g. $P_{S(d)}$ is the law of $S(d)$). Laws conditional on an event $\mathcal{E}$ are denoted by $P_{\cdot|\mathcal{E}}$ (e.g. $P_{S(d)|\mathcal{E}}$ is the conditional law of $S(d)$). If the random element is clear, I write $P(\cdot|\mathcal{E}, G = g)$ as $P_g(\cdot|\mathcal{E})$ for $g \in \{O, E\}$. Whenever $P_E(\cdot|\mathcal{E}) = P_O(\cdot|\mathcal{E})$, I omit the subscript g. This is inherited by their features $E[\cdot|\mathcal{E}, G = g] = E_g[\cdot|\mathcal{E}]$ and $V[\cdot|\mathcal{E}, G = g] = V_g[\cdot|\mathcal{E}]$. Equality of distribution of two random elements or a random element and a law is denoted by $\stackrel{d}{=}$ (e.g. $Y \stackrel{d}{=} P_Y$ and $Y \stackrel{d}{=} Y'$). I denote random sets with boldface letters (e.g. $\mathbf{Y}$), their capacity functionals by boldface $\mathbf{T}$ (e.g. $\mathbf{T_Y}$) and containment functionals by boldface $\mathbf{C}$ (e.g.

3. Note that this set may be larger than the intractable identified set that would have been obtained using non-discretized data.

$\mathbf{C_Y}$). I use $(\mathbf{Y}, Z)$ to denote the random set $\mathbf{Y} \times \{Z\}$. $\mathbb{E}(\mathbf{Y}|X)$ is used for the conditional Aumann expectation of a random set $\mathbf{Y}$ given a sigma-algebra generated by a random variable X. If a distribution P_Y is selectionable from $\mathbf{Y}$ I write $P_Y \preccurlyeq \mathbf{Y}$. I use $\overset{d}{=}$ to denote that a random element has a law, or an equivalent distribution-determining functional (e.g. $Y \overset{d}{=} P_Y$ and $\mathbf{Y} \overset{d}{=} \mathbf{C_Y}$). A, B and K represent sets. $\mathcal{K}(A)$, $\mathcal{C}(A)$, $\mathcal{O}(A)$, $\mathcal{B}(A)$ are the families of all compact, closed, open and Borel subsets of the set A, respectively. $co(A)$ is the closed convex hull of the set A. The identified sets for a generic parameter θ is $\mathcal{H}(\theta)$. The set of distribution functions of random variables with support $\mathcal{Y}$ is $\mathcal{P}^{\mathcal{Y}}$. I assume throughout that $\mathcal{Y} \times \mathcal{S}$ is a locally compact, second countable Hausdorff space, more precisely $\mathbb{R}^{1+d}$ endowed with its natural topology, while any of its subspaces inherit their relative topologies.

In the proofs for simpler notation I will use the following random variable:

$$\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1). \tag{A.2}$$

I use LIE to refer to the "law of iterated expectations".

Subsection A.2.2. Known Supporting Results

A.2.2.1. Random Set Theory Preliminaries. I briefly introduce the necessary concepts, and refer the reader to Molchanov (2017) and Molchanov and Molinari (2018) for a textbook treatment of the topic. More concise overviews are available in Beresteanu, Molchanov, and Molinari (2012) and Molchanov and Molinari (2014).

Define $\mathbf{R}:\Omega\to\mathcal{C}(\mathbb{R}^d)$ to be a measurable correspondence recalling that $\mathcal{C}(\mathbb{R}^d)$ is the collection of all closed subsets of $\mathbb{R}^d$.[4] I refer to $\mathbf{R}$ as a *random (closed) set*. Define the *containment functional* $\mathbf{C}_\mathbf{R}:\mathcal{C}(\mathbb{R}^d)\to[0,1]$ of $\mathbf{R}$ as $\mathbf{C}_\mathbf{R}(B)=P(\mathbf{R}\subseteq B)$, and the *capacity functional* $\mathbf{T}_\mathbf{R}:\mathcal{K}(\mathbb{R}^d)\to[0,1]$ of $\mathbf{R}$ as $\mathbf{T}_\mathbf{R}(K)=P(\mathbf{R}\cap K\neq\emptyset)$, recalling that $\mathcal{K}(\mathbb{R}^d)$ is the collection of all compact subsets of $\mathbb{R}^d$. A *selection* of a random set $\mathbf{R}$ is a random vector R defined on the same probability space such that $P(R\in\mathbf{R})=1$. The set of all selections of $\mathbf{R}$ is denoted by $Sel(\mathbf{R})$. The set of all random vectors $R\in Sel(\mathbf{R})$ such that $E[||R||]<\infty$ is denoted by $Sel^1(\mathbf{R})$. Artstein's inequalities (Artstein (1983, Theorem 2.1), Beresteanu, Molchanov, and Molinari (2012, Theorem 2.1)) give an equivalent characterization of the set of distributions of all selections of a random set.

Lemma A.2.1. *(Artstein's Inequalities) A probability distribution μ on a locally compact second countable Hausdorff space $\mathfrak{X}$ is the distribution of a selection of a random closed set $\mathbf{R}$ on the same space if and only if:*

$$\forall B\in\mathfrak{F}_{cont}:\ \ \mu(B)\geq\mathbf{C}_\mathbf{R}(B)\Leftrightarrow\forall K\in\mathfrak{F}_{cap}:\ \ \mu(K)\leq\mathbf{T}_\mathbf{R}(K) \tag{A.3}$$

where $\mathfrak{F}_{cont}\in\{\mathcal{C}(\mathfrak{X}),\mathcal{O}(\mathfrak{X})\}$ and $\mathfrak{F}_{cap}\in\{\mathcal{C}(\mathfrak{X}),\mathcal{O}(\mathfrak{X}),\mathcal{K}(\mathfrak{X})\}$. If $\mathbf{R}$ is almost surely compact, then (A.3) is equivalent to:

$$\forall K\in\mathcal{K}(\mathfrak{X}):\ \ \mu(K)\geq\mathbf{C}_\mathbf{R}(K). \tag{A.4}$$

Proof. For proof see Molchanov and Molinari (2018, Theorem 2.13, Corollary 2.14). □

4. $\mathbf{R}$ is measurable if for every compact set $K\in\mathcal{K}(\mathbb{R}^d)$: $\{\omega\in\Omega:\mathbf{R}(\omega)\cap K\neq\emptyset\}\in\mathcal{F}$. The codomain of the map $\mathbf{R}$ is equipped by the $\sigma-$algebra generated by the families of sets $\{B\in\mathcal{C}(\mathbb{R}^d):B\cap K\neq\emptyset\}$ over $K\in\mathcal{K}(\mathbb{R}^d)$.

If (A.3) holds for a distribution function P_R, then I call P_R *selectionable* with respect to the distribution of $\mathbf{R}$, and write $P_R \preccurlyeq \mathbf{R}$. μ is selectionable if and only if there exists a random element $R' \overset{d}{=} P_R$ and a random set $\mathbf{R}' \overset{d}{=} \mathbf{R}$ defined on the same probability space such that $P(R' \in \mathbf{R}') = 1$. Family of all distributions that satisfy (A.3) are called the *core* of the capacity $\mathbf{T_R}$. A family of compact sets $\mathcal{K}_{CD} \subseteq \mathcal{K}(\mathfrak{X})$ is a *core-determining class* if $\forall K \in \mathcal{K}_{CD}: \ \mu(K) \leq \mathbf{T_R}(K)$ implies (A.3). A core-determining class may reduce the number of conditions that need to be verified to consider μ selectionable.

If $\mathbf{R}$ has at least one integrable selection, that is $Sel^1(\mathbf{R}) \neq \emptyset$, then $\mathbf{R}$ is an *integrable random set*. Whenever the random variable $||\mathbf{R}|| = sup\{||R|| : R \in Sel(\mathbf{R})\}$ is integrable $E[||\mathbf{R}||] < \infty$, then $\mathbf{R}$ is said to be *integrably bounded*.

Definition A.2.1. (Aumann Expectation) The Aumann expectation of an integrable random set $\mathbf{R}$ is defined as:

$$\mathbb{E}(\mathbf{R}) = cl\{E[R] : R \in Sel^1(\mathbf{R})\}. \tag{A.5}$$

If $\mathbf{R}$ is integrably bounded, then:

$$\mathbb{E}(\mathbf{R}) = \{E[R] : R \in Sel(\mathbf{R})\}. \tag{A.6}$$

Note that when $\mathbf{R}$ is a finite-dimensional and integrably bounded, $\mathbb{E}(\mathbf{R})$ is a closed set, and the closure operator is not used in the definition. (Molchanov (2017, Theorem 2.1.37))

The *support function* for a convex set $A \in \mathbb{R}^{d_A}$ is defined as $h_A(u) = \sup_{a \in A} a'u$ for $u \in \mathbb{R}^{d_A}$. The convex set A is uniquely determined by its its support function via

intersections of all half-spaces defined by h_A as:

$$A = \bigcap_{u \in \mathbb{R}^{d_A} : ||u|| = 1} \{a \in \mathbb{R}^{d_A} : a'u \leq h_A(u)\}. \tag{A.7}$$

If $\mathbf{R}$ is integrably bounded and if either the underlying probability space is non-atomic, or if $\mathbf{R}$ is almost surely convex, then $h_{\mathbb{E}(\mathbf{R})}(u) = E[h_{\mathbf{R}}(u)]$ for all $u \in \mathbb{R}^{d_{\mathbf{R}}}$. (Molchanov and Molinari (2018, Theorem 3.11))

Recalling that $(\Omega, \mathcal{F}, P)$ is the underlying probabilty space, let $\mathcal{F}_0 \subsetneq \mathcal{F}$ be some sub-σ-algebra.

Definition A.2.2. (Conditional Aumann Expectation) Let $\mathbf{R}$ be an integrable random set. For each sub-σ-algebra $\mathcal{F}_0 \subsetneq \mathcal{F}$, the conditional Aumann expectation of $\mathbf{R}$ given $\mathcal{F}_0$ is the $\mathcal{F}_0$-measurable random set $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ such that:

$$Sel^1(\mathbb{E}[\mathbf{R}|\mathcal{F}_0], \mathcal{F}_0) = cl\{E[R|\mathcal{F}_0] : R \in Sel^1(\mathbf{R})\} \tag{A.8}$$

where $Sel^1(\cdot, \mathcal{F}_0)$ denote the set of integrable selections measurable with respect to $\mathcal{F}_0$ and the closure is taken in L^1.

For any integrable random set $\mathbf{R}$, the conditional Aumann expectation $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ is integrable, unique and exists. If $\mathbf{R}$ is integrably bounded, so is $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ (Molchanov (2017, Theorem 2.1.71)). When $\mathcal{F}_0$ is countably generated, then $cl\{E[R|\mathcal{F}_0] : R \in Sel^1(\mathbf{R})\} = \{E[R|\mathcal{F}_0] : R \in Sel^1(\mathbf{R})\}$. (Molchanov (2017, pp. 271), Li and Ogura (1998, Theorem 1)) Recall that when $\mathcal{F}_0$ is a σ-algebra generated by a random vector, it is countably generated. Therefore, for any random vector W, $Sel^1(\mathbb{E}[\mathbf{R}|\sigma(W)], \sigma(W))$ is a closed set.

If for all $A \in \mathcal{F}$ with $P(A) > 0$ there exists $B \in \mathcal{F}$ with $B \subseteq A$ such that $0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)$ with positive probability, then the probability measure is said to have not atoms over $\mathcal{F}_0$. Then, $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$ is almost surely convex and $\mathbb{E}[\mathbf{R}|\mathcal{F}_0] = \mathbb{E}[co(\mathbf{R})|\mathcal{F}_0]$ a.s. (Molchanov (2017, Theorem 2.1.77)) Then, $h_{\mathbb{E}[\mathbf{R}|\mathcal{F}_0]}(u) = h_{\mathbb{E}[co(\mathbf{R})|\mathcal{F}_0]}(u) = E[h_{co(\mathbf{R})}(u)|\mathcal{F}_0]$ a.s. for all $u \in \mathbb{R}^{d_\mathbf{R}}$. (Molchanov (2017, Theorem 2.1.72))[5] Note that this will hold for any sub-σ-algebra $\mathcal{F}_0$ by Lemma A.2.2 when the probability space is non-atomic.

A.2.2.2. Other Known Results for Reference.

Theorem A.2.1. *Let E, F be metrizable and let G be any topological vector space. If E is a Baire space or if E is barreled and G is locally convex, then every separately equicontinuous family B of bilinear mappings of $E \times F$ into G is equicontinuous.*

Proof. See Schaefer and Wolff (1999, Theorem III.5.1). □

Corollary A.2.1. *Let E, F be metrizable and let G be any topological vector space. If E is a Baire space or if E is barreled and G is locally convex, then every separately continuous bilinear map of $E \times F$ into G is continuous (see also Treves (2016, pp. 425)).*

Proof. Direct from Theorem A.2.1. □

Subsection A.2.3. Auxiliary Lemmas

Lemma A.2.2. *Suppose the probability space $(\Omega, \mathcal{F}, P)$ is non-atomic and that $\mathcal{F}_0 \subseteq \mathcal{F}$ is a sub-σ-algebra. Then P is atomless over $(\Omega, \mathcal{F}_0)$. That is, for all $A \in \mathcal{F}$ with $P(A) > 0$ there exists $B \in \mathcal{F}$ with $B \subseteq A$ such that $0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)$ with positive probability.*

5. Theorem 2.1.72 states that $h_{\mathbb{E}[\mathbf{R}|\mathcal{F}_0]}(u) = E[h_\mathbf{R}(u)|\mathcal{F}_0]$ a.s. for all $u \in \mathbb{R}^{d_\mathbf{R}}$. If one wishes to use the support function to determine elements of $\mathbb{E}[\mathbf{R}|\mathcal{F}_0]$, the step $h_{\mathbb{E}[\mathbf{R}|\mathcal{F}_0]}(u) = h_{\mathbb{E}[co(\mathbf{R})|\mathcal{F}_0]}(u)$ by Theorem 2.1.77 is necessary.

Proof. Pick any $A \in \mathcal{F}$ with positive measure and fix any $B \in \mathcal{F}$ such that $B \subseteq A$ and $0 < P(B) < P(A)$. B exists since $(\Omega, \mathcal{F}, P)$ is non-atomic. Let $C = A \setminus B$ and observe that $A = B \cup C$ and $B \cap C = \emptyset$. Thus, $P(A) = P(B) + P(C)$, $P(C) > 0$ and $P(A|\mathcal{F}_0) = P(B|\mathcal{F}_0) + P(C|\mathcal{F}_0)$ a.s. I proceed by way of contradiction supposing that $P(B|\mathcal{F}_0) = P(A|\mathcal{F}_0)$ a.s. or $P(B|\mathcal{F}_0) = 0$ a.s. Consider $P(B|\mathcal{F}_0) = P(A|\mathcal{F}_0)$ a.s. first. Then, $P(C|\mathcal{F}_0) = 0$ which implies $P(C) = 0$, contradicting $P(C) > 0$. If $P(B|\mathcal{F}_0) = 0$ a.s., then $P(B) = 0$ a.s., contradicting $P(B) > 0$. Therefore, the set $\{\omega \in \Omega : 0 < P(B|\mathcal{F}_0)(\omega) < P(A|\mathcal{F}_0)(\omega)\}$ has positive probability, which concludes the proof. □

Lemma A.2.3. *Suppose that Assumption EV holds, and that exerimental data are unobserved. Then the identified set for the distribution function $P_{Y(d),S(d)}$ is:*

$$\mathcal{H}^O(P_{Y(d),S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \delta(B) \geq P_O((Y,S) \in B, D = d) \ \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})\} \tag{A.9}$$

Proof. The proof proceeds by extending arguments of Beresteanu, Molchanov, and Molinari (2012, Proposition 2.4). Define the random set for $d \in \{0, 1\}$:

$$(\mathbf{Y}^O(d), \mathbf{S}^O(d)) = \begin{cases} \{(Y,S)\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases}. \tag{A.10}$$

By definition, $(\mathbf{Y}^O(d), \mathbf{S}^O(d))$ summarizes all information on $(Y(d), S(d))$ *in the observational data*. Let $\tilde{I}$ be the set of triples random elements (E_1, E_2, E_3) such that $(E_1, E_2, E_3) \in \mathcal{Y} \times \mathcal{S} \times G$ and $(E_1, E_2) \perp\!\!\!\perp E_3$. Then all information in the data and

assumptions can be expressed as $(Y(d), S(d), G) \in Sel((\mathbf{Y}^O(d), \mathbf{S}^O(d), G)) \cap \tilde{I}$. Note that this set is non-empty since $(\mathbf{Y}^O(d), \mathbf{S}^O(d))$ produces non-trivial values only for $G = O$.

By Lemma A.2.1, the distribution function $P((Y(d), S(d), G)) \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}\times\{E,O\}}$ characterizes a selection in $Sel((\mathbf{Y}^O(d), \mathbf{S}^O(d), G))$ if and only if:

(A.11)

$$\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S} \times \{E, O\}) : \quad P((Y(d), S(d), G) \in B) \geq P((\mathbf{Y}^O(d), \mathbf{S}^O(d), G) \subseteq B)$$

By Molchanov and Molinari (2018, Theorem 2.33), (A.11) is equivalent to:

(A.12)

$$\begin{aligned} &\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \quad P((Y(d), S(d)) \in B|G) \geq P((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq B|G) \ P\text{-a.s.} \\ \Leftrightarrow &\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \quad P((Y(d), S(d)) \in B|G) \geq P_O((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq B) \end{aligned}$$

where the second line follows since experimental data are unobserved, and hence $P(G = O) = 1$.

For $A = \mathcal{Y} \times \mathcal{S}$, $P_O((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq A) = 1$.[6] For any other closed subset $B \subsetneq \mathcal{Y} \times \mathcal{S}$, the containment functional can be written as:

$$\begin{aligned} P_O((\mathbf{Y}^O(d), \mathbf{S}^O(d)) \subseteq B) &= P_O((Y(d), S(d)) \in B, D = d) \\ &= P_O((Y, S) \in B, D = d). \end{aligned}$$

where the second equality follows by definition of Y and S. Hence, the identified set for $P_{Y(d),S(d)}$ follows by (A.12) and (A.11). Sharpness follows by construction. For any

6. The support of a random variable X is the smallest closed set $\mathcal{X}$ such that $P(X \in \mathcal{X}) = 1$. Hence $\mathcal{Y} \times \mathcal{S} \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$.

$P_{Y(d),S(d)} \in \mathcal{H}^O(P_{Y(d),S(d)})$ there exist $(Y(d), S(d))$ that are consistent with the data and assumptions such that $(Y(d), S(d)) \stackrel{d}{=} P_{Y(d),S(d)}$. □

Lemma A.2.4. *Suppose that Assumptions RA and EV hold. Then the identified set* $\mathcal{H}(P_{Y(d),S(d)})$ *for the distribution function* $P_{Y(d),S(d)}$ *is:*

(A.13)

$$\mathcal{H}(P_{Y(d),S(d)}) =$$

$$\left\{ \begin{array}{l} \delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \ \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}) : \\ \delta(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y,S) \in B, D = d) \end{array} \right] \end{array} \right\}$$

Proof. The proof proceeds by extending arguments of Lemma A.2.3. Define the random set for $d \in \{0, 1\}$:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y,S)\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y} \times \{S\}, \text{ if } (D,G) = (d,E) \\ \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases} . \tag{A.14}$$

By definition $(\mathbf{Y}(d), \mathbf{S}(d))$ summarizes all information in the observed data on $(Y(d), S(d))$. Let $\tilde{I}$ be the set of triples random elements (E_1, E_2, E_3) such that $(E_1, E_2, E_3) \in \mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}$ and $(E_1, E_2) \perp\!\!\!\perp E_3$. Then all information in the data and assumptions can be expressed as $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$. If Assumptions RA and EV hold, $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I} \neq \emptyset$.

By Lemma A.2.1, the distribution function $P((Y(d), S(d), \tilde{Z})) \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}\times\tilde{\mathcal{Z}}}$ characterizes a selection in $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))$ if and only if:

$$\text{(A.15)} \qquad \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}\times\tilde{\mathcal{Z}}): \quad P((Y(d), S(d), \tilde{Z}) \in B) \geq P((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) \subseteq B)$$

By Molchanov and Molinari (2018, Theorem 2.33), (A.15) is equivalent to:

$$\text{(A.16)} \qquad \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \quad P((Y(d), S(d)) \in B|\tilde{Z}) \geq P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) \ P\text{-a.s.}$$

Possible forms that B can take are: 1) $B = \mathcal{Y}\times\mathcal{S}$; 2) $B = \mathcal{Y}\times B_S$ for some $B_S \subsetneq \mathcal{S}$; 3) $B \neq \mathcal{Y}\times B_S$ for any $B_S \subseteq \mathcal{S}$. Now consider the containment functional $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z})$ for each case.

For $B = \mathcal{Y}\times\mathcal{S}$, $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = 1$ P-a.s. If $B = \mathcal{Y}\times B_S$ for some $B_S \subsetneq \mathcal{S}$, then P-a.s.:

$$\begin{aligned} P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) &= P((Y, S) \in B, D = d|\tilde{Z}) \\ &= P(Y \in \mathcal{Y}, S \in B_S, D = d|\tilde{Z}) \\ &= P(S \in B_S, D = d|\tilde{Z}) \end{aligned}$$

where the first equality is by definition of the random set, the second is by the fact that $B = \mathcal{Y}\times B_S$, and the third by definition of $\mathcal{Y}$. Finally, if for all $B_S \subseteq \mathcal{S}: B \neq \mathcal{Y}\times B_S$:

$$P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = \begin{cases} 0, \text{ if } \tilde{Z} \in \mathcal{Z} \text{ (i.e. } G = E) \\ P_O((Y, S) \in B, D = d), \text{ if } \tilde{Z} \notin \mathcal{Z} \text{ (i.e. } G = O) \end{cases}.$$

To see why the first case holds, define the fiber of B at point s as $B_Y(s) = \{y : (y, s) \in B\}$. Observe that if for all $B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S$, then for some s it must be that $B_Y(s) \subsetneq \mathcal{Y}$. Therefore, whenever $G = E$ (or equivalently $\tilde{Z} \in \mathcal{Z}$), the random set $(\mathbf{Y}(d), \mathbf{S}(d)) = \mathcal{Y} \times \{S\} \not\subseteq B$. Hence, only if $G = O$ can the containment functional be positive, that is, when $\tilde{Z} \notin \mathcal{Z}$. That $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = P_O((Y, S) \in B, D = d)$ when $G = O$ is immediate by definitions of Y, S, $\tilde{Z}$ and the random set.

Collect the relevant cases to characterize the containment functional:

$$P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = \begin{cases} P(S \in B_S, D = d|\tilde{Z}), \text{ if } B = \mathcal{Y} \times B_S \text{ for some } B_S \subseteq \mathcal{S} \\ \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y, S) \in B, D = d), \text{ otherwise} \end{cases}$$

$$= \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) +$$

$$\mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y, S) \in B, D = d)$$

Hence, the distribution function $P((Y(d), S(d), \tilde{Z})) \in \mathcal{P}^{\mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}}$ characterizes a selection in $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ P-a.s.:

$$P((Y(d), S(d)) \in B|\tilde{Z})$$

$$\geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y, S) \in B, D = d) \end{array} \right].$$

Finally, to incorporate the fact that $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$, intersect $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$ which yields:

$$
\begin{aligned}
& P((Y(d), S(d)) \in B) \\
& \geq ess \sup_{\tilde{z}} \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d | \tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y, S) \in B, D = d) \end{array} \right]. \\
& = \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] ess \sup_{\tilde{Z}} P(S \in B_S, D = d | \tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y, S) \in B, D = d) \end{array} \right] \\
& = \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess \sup_Z P_E(S \in B_S, D = d | Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y, S) \in B, D = d) \end{array} \right]
\end{aligned}
$$

where the first line follows by the fact that $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$, the second by the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S]$ and $\mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S]$ refer to mutually exclusive deterministic events, and the third by definition of $\tilde{Z}$ and the fact that $P(G = g) > 0$ for $g \in \{O, E\}$. Hence, the identified set for $P_{Y(d),S(d)}$ follows by (A.16) and (A.15). Sharpness follows by construction. For any $P_{Y(d),S(d)} \in \mathcal{H}^O(P_{Y(d),S(d)})$ there exist $(Y(d), S(d))$ that are consistent with the data and assumptions such that $(Y(d), S(d)) \stackrel{d}{=} P_{Y(d),S(d)}$. □

Lemma A.2.5. *Let $\mathcal{H}^O(P_{Y(d)})$ and $\mathcal{H}(P_{Y(d)})$ be the sets of marginals of distributions in $\mathcal{H}^O(P_{Y(d),S(d)})$ and $\mathcal{H}(P_{Y(d),S(d)})$. Then:*

$$
\mathcal{H}^O(P_{Y(d)}) = \mathcal{H}(P_{Y(d)}) = \{\delta \in \mathcal{P}^{\mathcal{Y}} : \delta(B) \geq P_O(Y \in B, D = d) \ \forall B \in \mathcal{C}(\mathcal{Y})\}. \tag{A.17}
$$

Proof. For any Borel set $B \in \mathcal{B}(\mathbb{R})$, by definition of a marginal distribution function:

$$P(Y(d) \in B) = P(Y(d) \in B, S(d) \in \mathcal{S}) = P((Y(d), S(d)) \in B \times \mathcal{S}) \tag{A.18}$$

where the last line is by equivalence of events $\{Y(d) \in B, S(d) \in \mathcal{S}\}$ and $\{(Y(d), S(d)) \in B \times \mathcal{S}\}$. Lemma A.2.3 yields the identified set for joint distributions $P_{Y(d),S(d)}$ using only observational data:

$$\mathcal{H}^O(P_{Y(d),S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \delta(B) \geq P_O((Y,S) \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})\}. \tag{A.19}$$

Hence the identified set for marginals $P_{Y(d)}$ using only observational data is:

$$\begin{aligned}
&\mathcal{H}^O(P_{Y(d)}) \\
&= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : \exists \delta \in \mathcal{H}^O(P_{Y(d),S(d)}) \text{ s.t. } P(Y(d) \in B) = \delta(B \times \mathcal{S})\ \forall B \in \mathcal{B}(\mathbb{R})\} \\
&= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : P(Y(d) \in B) \geq P_O((Y,S) \in B \times \mathcal{S}, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y})\} \\
&= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : P(Y(d) \in B) \geq P_O(Y \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y})\}
\end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma A.2.3 and the third is by (A.18).

Lemma A.2.4 yields the identified set for joint distributions $P_{Y(d),S(d)}$ using combined data:

$$\mathcal{H}(P_{Y(d),S(d)}) =$$

$$\left\{ \begin{array}{l} \delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \ \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}) : \\ \delta(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S]\times \\ \max\left(ess\sup_Z P_E(S\in B_S, D=d|Z), P_O(S\in B_S, D=d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P_O((Y,S)\in B, D=d) \end{array} \right] \end{array} \right\} \tag{A.20}$$

Observe that the marginals are fully defined by Borel sets of the form $B \times \mathcal{S}$ with $B \subsetneq \mathcal{Y}$, which means that for all sets of interest $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] = 0$ in the expression above. Thus, the identified set for marginals $P_{Y(d)}$ using combined data is:

$$\begin{aligned} \mathcal{H}(P_{Y(d)}) &= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : \exists \delta \in \mathcal{H}(P_{Y(d),S(d)}) \text{ s.t } P(Y(d)\in B) = \delta(B\times\mathcal{S}) \ \forall B \in \mathcal{B}(\mathbb{R})\} \\ &= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B) \geq P_O((Y,S)\in B\times\mathcal{S}, D=d) \ \forall B \in \mathcal{C}(\mathcal{Y})\} \\ &= \{P_{Y(d)} \in \mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B) \geq P_O(Y\in B, D=d) \ \forall B \in \mathcal{C}(\mathcal{Y})\} \end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma A.2.4 and the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] = 0$, and the third is by (A.18). It is then immediate that $\mathcal{H}(P_{Y(d)}) = \mathcal{H}^O(P_{Y(d)})$ □

Remark A.2.1. The formulation of the identified sets $\mathcal{H}^O(P_{Y(d)})$ and $\mathcal{H}(P_{Y(d)})$ coincides by application of (A.3) to the random set:

$$\mathbf{Y}(d) = \begin{cases} \{Y\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}.$$

Lemma A.2.6. *Let $\mathcal{H}^O(P_{S(d)})$ and $\mathcal{H}(P_{S(d)})$ be the sets of marginals of distributions in $\mathcal{H}^O(P_{Y(d),S(d)})$ and $\mathcal{H}(P_{Y(d),S(d)})$. Then:*

(A.21)

$$\mathcal{H}^O(P_{S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{S}} : \delta(B) \geq P_O(S \in B, D = d) \ \forall B \in \mathcal{C}(\mathcal{S})\}$$

(A.22)

$$\mathcal{H}(P_{S(d)}) = \left\{ \begin{array}{l} \delta \in \mathcal{P}^{\mathcal{S}} : \ \forall B \in \mathcal{C}(\mathcal{S}) : \\ \delta(B) \geq \ \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \end{array} \right\}$$

Let $\mathcal{H}^E(P_{S(d)})$ be the identified set for $P_E(S(d))$ obtained using only experimental data. Then:

(A.23) $$\mathcal{H}^E(P_{S(d)}) = \{\delta \in \mathcal{P}^{\mathcal{S}} : ess\sup_Z P_E(S \in B, D = d|Z) \ \forall B \in \mathcal{C}(\mathcal{S})\}.$$

Proof. For any Borel set $B \in \mathcal{B}(\mathbb{R})$, by definition of a marginal distribution:

(A.24) $$P(S(d) \in B) = P(S(d) \in \mathcal{Y}, S(d) \in B) = P((Y(d), S(d)) \in \mathcal{Y} \times B)$$

where the last line is by equivalence of events $\{Y(d) \in \mathcal{Y}, S(d) \in B\}$ and $\{(Y(d), S(d)) \in \mathcal{Y} \times B\}$. Lemma A.2.3 yields the identified set for joint distributions $P_{Y(d),S(d)}$ using only

observational data:

$$\text{(A.25)} \qquad \mathcal{H}^O(P_{Y(d),S(d)}) = \{\delta \in \mathcal{P} : \delta(B) \geq P_O((Y,S) \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})\}.$$

Hence the identified set for marginals $P(S(d))$ using only observational data is:

$$\begin{aligned}
&\mathcal{H}^O(P_{S(d)}) \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : \exists \delta \in \mathcal{H}^O(P_{Y(d),S(d)}) \text{ s.t } P(S(d) \in B) = \delta(\mathcal{Y} \times B)\ \forall B \in \mathcal{B}(\mathbb{R})\} \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : P(S(d) \in B) \geq P_O((Y,S) \in \mathcal{Y} \times B, D = d)\ \forall B \in \mathcal{C}(\mathcal{S})\} \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : P(S(d) \in B) \geq P_O(S \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{S})\}
\end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma A.2.3 and the third is by (A.24).

Lemma A.2.4 yield the identified set for joint distributions $P_{Y(d),S(d)}$ using combined data:

$$\text{(A.26)} \qquad \mathcal{H}(P_{Y(d),S(d)}) = \left\{ \begin{array}{l} \delta \in \mathcal{P} :\ \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \\ \delta(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y,S) \in B, D = d) \end{array} \right] \end{array} \right\}$$

Observe that marginals are defined by Borel sets of the form $\mathcal{Y} \times B$, which means that for all sets of interest $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 1$ in the expression above. Thus, the

identified set for marginals $P(S(d))$ using combined data is:

$$\begin{aligned}\mathcal{H}(P_{S(d)}) &= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : \exists \delta \in \mathcal{H}(P_{Y(d),S(d)}) \text{ s.t } P(S(d) \in B) = \delta(\mathcal{Y} \times B)\ \forall B \in \mathcal{B}(\mathbb{R})\} \\ &= \{\delta \in \mathcal{P}^{\mathcal{S}} : \delta(B) \geq P_O(S \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{S})\} \\ &= \left\{ \begin{array}{l} \delta \in \mathcal{P}^{\mathcal{S}} : \ \forall B \in \mathcal{C}(\mathcal{S}) : \\ \delta(B) \geq \ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) \end{array} \right\}.\end{aligned}$$

where the first line is by definition of a marginal distribution, and the second is by Lemma A.2.4 and the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 1$.

For $\mathcal{H}^E(P_{S(d)})$ a simplified version of the argument for Lemma A.2.4 applies, and is therefore omitted. □

Remark A.2.2. The formulation of the identified sets $\mathcal{H}^O(P_{S(d)})$ and $\mathcal{H}(P_{S(d)})$ coincides by application of (A.3) to the random set:

$$(\mathbf{S}(d), \tilde{Z}) = \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases}$$

and finding the set of selections $Sel(\mathbf{S}(d), \tilde{Z}) \cap I_S$ where I_S is the set of random elements (E_1, E_2) such that $E_1 \perp\!\!\!\perp E_2$.

Lemma A.2.7. *Let $\mathcal{Y}$ be a compact set. If there exists $d \in \{0, 1\}$ such that $V_O[Y|S, D = d] > 0$ $P-$a.s., then $E_O[Y|S, D = d] \in (\inf \mathcal{Y}, \sup \mathcal{Y})$ $P-$a.s.*

Proof. I prove that $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s. and $E_O[Y|S, D = d] > \inf \mathcal{Y}$ $P-$a.s follows by a symmetric argument. Since $\mathcal{Y}$ is a compact set, both $\sup \mathcal{Y}$ and $\inf \mathcal{Y}$ are finite.

By contraposition suppose that $P(E_O[Y|S, D = d] \geq \sup \mathcal{Y}) > 0$. Then by definition of $\mathcal{Y}$, $P(E_O[Y|S, D = d] = \sup \mathcal{Y}) > 0$, so there exists a Borel subset $B \subseteq \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D = d) > 0$ such that $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Now I show that this implies $P(Y = \sup \mathcal{Y}|S \in B, D = d) = 1$. Suppose not, so that $P(Y = \sup \mathcal{Y}|S \in B, D = d) < 1$, then:

(A.27)

$$\begin{aligned}
E_O[Y|S \in B, D = d] =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d]P(Y = \sup \mathcal{Y}|S \in B, D = d)+ \\
& E_O[Y|Y \neq \sup \mathcal{Y}, S \in B, D = d]P_O(Y \neq \sup \mathcal{Y}|S \in B, D = d) \\
=& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d]P_O(Y = \sup \mathcal{Y}|S \in B, D = d)+ \\
& E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d]P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\
=& \sup \mathcal{Y} P_O(Y = \sup \mathcal{Y}|S \in B, D = d)+ \\
& E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d]P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\
<& \mathcal{Y}
\end{aligned}$$

where the first equality is by LIE, second is by definition of $\mathcal{Y}$, third by $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$, and the fourth by $E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] < \sup \mathcal{Y}$ and $P(Y = \mathcal{Y}|S \in B, D = d) < 1$. By assumption, $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Then (A.27) yields a contradiction, showing that $P(Y = \sup \mathcal{Y}|S \in B, D = d) = 1$. But then $V_O[Y|S \in B, D = d] = 0$ and $P_O(S \in B|D = d) > 0$, so $P(V_O[Y|S \in B, D = d] = 0) > 0$

which contradicts $V_O[Y|S \in B, D = d] > 0$ $P-$a.s. Thus $V_O[Y|S, D = d] > 0$ $P-$a.s. implies $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s.

□

Lemma A.2.8. *For any γ_d that is a distribution of a selection in $Sel((\mathbf{S}(d), \tilde{Z})) \cap I$, there exists a γ_d-integrable function π_{γ_d} such that for any measurable set $B \in \mathcal{B}(\mathcal{S})$:*

$$P_O(S \in B, D = d) = \int_B \pi_{\gamma_d} d\gamma_d. \tag{A.28}$$

Then for the propensity score functional $\pi_{\gamma_d} := \frac{dP_O(S,D=d)}{d\gamma_d}$ and any $\varsigma'_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ with $\varsigma'_d \overset{d}{=} \gamma'_d$:

$$P_O(D = d|\varsigma'_d) = \pi_{\gamma'_d}(\varsigma'_d) \text{ a.s.} \tag{A.29}$$

Proof. Fix any γ_d such that $\exists \varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\gamma_d \overset{d}{=} \varsigma_d$. Then for any $B \in \mathcal{B}(\mathcal{S})$:

$$P_O(\varsigma_d \in B, D = d) \leq P_O(\varsigma_d \in B) = P(\varsigma_d \in B) = \gamma_d(B)$$

where the inequality is by observation. For the first equality, recall that I is a set of random elements $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$, and observe that $(\varsigma_d, \tilde{Z}) \in I$. Therefore, $\varsigma_d \perp\!\!\!\perp G$, by definition of $\tilde{Z}$. For the second equality note that $\varsigma_d \overset{d}{=} \gamma_d$.

Next, note that that $P_O(\varsigma_d \in B, D = d) = P_O(S \in B, D = d)$ for any measurable set $B \in \mathcal{B}(\mathcal{S})$ because $\varsigma_d \in Sel(\mathbf{S}(d))$ and $P_O(\mathbf{S}(d) = \{S\}, D = d) = 1$. Therefore, $P_O(S \in B, D = d) \leq \gamma_d(B)$ for any $B \in \mathcal{B}(\mathcal{S})$. Hence, $P_O(S, D = d)$ is absolutely continuous with respect to γ_d. Then, by the Radon-Nikodym theorem there exists a

measurable function π_{γ_d} such that for any measurable set $B \in \mathcal{B}(\mathcal{S})$:

$$P_O(S \in B, D = d) = \int_B \pi_{\gamma_d} d\gamma_d$$

and $\pi_{\gamma_d} = dP_O(S, D = d)/d\gamma_d$.

Therefore, for any γ'_d that is a distribution of a selection in $Sel((\mathbf{S}(d), \tilde{Z})) \cap I$, there exists $\pi_{\gamma'_d} = dP_O(S, D = d)/d\gamma'_d$ $P_O(S \in B, D = d) = \int_B \pi_{\gamma'_d} d\gamma'_d$ for any measurable set $B \in \mathcal{B}(\mathcal{S})$. Hence, also $\pi_{\gamma'_d}(s) = P_O(D = d|\varsigma_d = s)$, γ'_d-a.e. $s \in \mathcal{S}$, which concludes the proof. □

Lemma A.2.9. *Suppose Assumption RA holds. Assume that there is perfect experimental compliance so $Z = D|G = E$ $P-$a.s. and define conditions:*

C.1 (Surrogacy) $Y \perp\!\!\!\perp D|S, G = E$;

C.2 (Comparability) $Y \perp\!\!\!\perp G|S$.

Then:

i) C.1 implies $E_E[Y(1)|S(1) = s] = E_E[Y(0)|S(0) = s]$ for all $s \in \mathcal{S}$;

ii) C.1 and EV imply $E_g[Y(1)|S(1) = s] = E_{g'}[Y(0)|S(0) = s]$ for all $s \in \mathcal{S}$ and $g, g' \in \{O, E\}$;

iii) C.2 implies $E_O[Y|S = s] = E_E[Y(1)|S(1) = s]P_E(D = 1|S = s) + E_E[Y(0)|S(0) = s]P_E(D = 0|S = s)$ for all $s \in \mathcal{S}$;

iv) C.2 and EV imply $E_O[Y|S = s] = E_g[Y(1)|S(1) = s]P_E(D = 1|S = s) + E_{g'}[Y(0)|S(0) = s]P_E(D = 0|S = s)$ for all $s \in \mathcal{S}$ and $g, g' \in \{O, E\}$;

v) C.1 and C.2 imply $E_O[Y|S = s] = E_E[Y(d)|S(d) = s]$ for all $s \in \mathcal{S}$;

vi) C.1, C.2 and EV imply $E_O[Y|S = s] = E_g[Y(d)|S(d) = s]$ *for all* $s \in \mathcal{S}$ *and* $g \in \{O, E\}$.

Proof. *i)* Write for any $d \in \{0, 1\}$:

$$E_E[Y|S] = E_E[Y|S, D = d] = E_E[Y(d)|S(d), D = d] = E_E[Y(d)|S(d)] \tag{A.30}$$

where the first equality is by surrogacy, second is by definition, and third is by random assignment and perfect compliance.

ii) Under Assumption EV, $E_E[Y(d)|S(d)] = E[Y(d)|S(d)]$. The result then follows from *i)*.

iii) Write:

$$\begin{aligned} & E_O[Y|S = s] \\ & = E_E[Y|S = s] \\ & = E_E[Y(1)|S(1) = s, D = 1]P_E(D = 1|S = s) \\ & + E_E[Y(0)|S(0) = s, D = 0]P_E(D = 0|S = s) \\ & = E_E[Y(1)|S(1) = s]P_E(D = 1|S = s) + E_E[Y(0)|S(0) = s]P_E(D = 0|S = s) \end{aligned} \tag{A.31}$$

where the first equality is by comparability, second is by LIE and definitions of Y and S, and the third is by random assignment and perfect compliance. *iv)* Under Assumption EV, $E_E[Y(d)|S(d)] = E[Y(d)|S(d)]$. The result then follows from *iii)*.

v) Immediate from *i)* and *iii)*.

vi) Immediate from *v)* under Assumption EV.

□

Subsection A.2.4. Proofs of Main Results

PROOF OF LEMMA 1.3.1. Let I be the set of random elements $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ such that $E_1 \perp\!\!\!\perp E_2$. Recalling the definition of $\tilde{Z}$ in (A.2), note that $S(d) \in Sel(\mathbf{S}(d)) \cap \bar{I}$ can be equivalently stated as $(S(d), \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z}) \cap I$.

The proof then proceeds through a series of steps:

(1) Find the set of (m_d, γ_d) which are consistent with the data, Assumptions RA and EV, and $E[|Y(d)|] < \infty$ in terms of measurable selections $(\upsilon_d, \varsigma_d)$ of $(\mathbf{Y}(d), \mathbf{S}(d))$;

(2) Equivalently characterize the set, removing redundant restrictions;

(3) Find the set of corresponding (m, γ) consistent with the data, Assumptions RA and EV, and $E[|Y(d)|] < \infty$;

(4) Collect all (m, γ) that satisfy Assumption MA to obtain $\mathcal{H}(m, \gamma)$.

Step 1: Restrictions on (m_d, γ_d) without the modeling assumption and integrability.

I use the random set:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y, S)\}, & \text{if } (D, G) = (d, O) \\ \mathcal{Y} \times \{S\}, & \text{if } (D, G) = (d, E) \\ \mathcal{Y} \times \mathcal{S}, & \text{otherwise} \end{cases} \tag{A.32}$$

which summarizes all information on $(Y(d), S(d))$ contained in the data, by definition. Recall from the proof of Lemma A.2.4 that all restrictions imposed by data and Assumptions RA and EV on $(Y(d), S(d))$ can be expressed as $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$ where $\tilde{I}$ is the set of all random elements $(E_1, E_2, E_3) \in \mathcal{S} \times \mathcal{Y} \times \tilde{\mathcal{Z}}$ such that $E_3 \perp\!\!\!\perp (E_1, E_2)$. Then, the set of (m_d, γ_d) consistent with the data and Assumptions RA and EV follows

by definition as:

(A.33)

$$\mathcal{H}^{EV/RA}(m_d,\gamma_d)=\left\{\begin{array}{l}(m_d,\gamma_d)\in\mathcal{M}_d\times\mathcal{P}^{\mathcal{S}}:\exists(\upsilon_d,\varsigma_d,\tilde{Z})\in Sel((\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z}))\cap\tilde{I},\\ \gamma_d\overset{d}{=}\varsigma_d,\ \ m_d(\varsigma_d)=E[\upsilon_d|\varsigma_d]\text{ a.s.}\end{array}\right\}$$

where $\mathcal{M}_d$ is the projection of $\mathcal{M}$ onto its first component. Next, recall the definition of random sets:

(A.34)

$$\mathbf{Y}(d)=\begin{cases}\{Y\},\ \text{if }(D,G)=(d,O)\\ \mathcal{Y},\ \text{otherwise}\end{cases},\ \ \mathbf{S}(d)=\begin{cases}\{S\},\ \text{if }(D,G)\in\{(d,E),(d,O)\}\\ \mathcal{S},\ \text{otherwise}\end{cases}$$

so that $(\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z})=\mathbf{Y}(d)\times\mathbf{S}(d)\times\{\tilde{Z}\}$ for any $d\in\{0,1\}$.

I now show that $Sel((\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z}))=Sel(\mathbf{Y}(d))\times Sel(\mathbf{S}(d))\times\{\tilde{Z}\}$. Fix an arbitrary $(\upsilon_d,\varsigma_d,\tilde{Z})\in Sel((\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z}))$. Then:

(A.35)

$$\begin{aligned}1&=P\Big((\upsilon_d,\varsigma_d,\tilde{Z})\in(\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z})\Big)\\&=P\Big(\upsilon_d\in\mathbf{Y}(d),\varsigma_d\in\mathbf{S}(d),\tilde{Z}\in\{\tilde{Z}\}\Big)\\&=P\Big(\upsilon_d\in\mathbf{Y}(d),\varsigma_d\in\mathbf{S}(d)\Big)\\&\leq P\Big(\upsilon_d\in\mathbf{Y}(d)\Big).\end{aligned}$$

where the first line follows since $(\upsilon_d,\varsigma_d,\tilde{Z})\in Sel((\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z}))$, the second is by $(\mathbf{Y}(d),\mathbf{S}(d),\tilde{Z})=\mathbf{Y}(d)\times\mathbf{S}(d)\times\{\tilde{Z}\}$, the third and fourth are by observation. Hence $P\Big(\upsilon_d\in\mathbf{Y}(d)\Big)=1$. By a similar argument, $P\Big(\varsigma_d\in\mathbf{S}(d)\Big)=1$. Therefore $(\upsilon_d,\varsigma_d,\tilde{Z})\in Sel(\mathbf{Y}(d))\times Sel(\mathbf{S}(d))\times\{\tilde{Z}\}$.

Next, fix an arbitrary $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$. Then:

$$
\begin{aligned}
1 &= P\Big(\upsilon_d \in \mathbf{Y}(d)\Big) \\
&= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d)\Big) + P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \notin \mathbf{S}(d)\Big) \\
&= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d), \tilde{Z} \in \{\tilde{Z}\}\Big) + P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \notin \mathbf{S}(d)\Big) \qquad \text{(A.36)} \\
&= P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \in \mathbf{S}(d), \tilde{Z} \in \{\tilde{Z}\}\Big) \\
&= P\Big((\upsilon_d, \varsigma_d, \tilde{Z}) \in (\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})\Big).
\end{aligned}
$$

where the first line is since $\upsilon_d \in Sel(\mathbf{Y}(d))$, second and third are by observation, fourth is since $P\Big(\upsilon_d \in \mathbf{Y}(d), \varsigma_d \notin \mathbf{S}(d)\Big) \leq P\Big(\varsigma_d \notin \mathbf{S}(d)\Big) = 0$ given that $\varsigma_d \in Sel(\mathbf{S}(d))$, and the last is by $(\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) = \mathbf{Y}(d) \times \mathbf{S}(d) \times \{\tilde{Z}\}$. Thus, $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) = Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$. Then write:

$$
\begin{aligned}
&\mathcal{H}^{EV/RA}(m_d, \gamma_d) \\
&= \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \exists (\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap \tilde{I}, \\ \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\} \\
&= \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \;\; \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \;\; (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\} \qquad \text{(A.37)} \\
&= \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \;\; \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \;\; (\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E_O[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\}
\end{aligned}
$$

where the first line holds by $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) = Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$, second is by rearrangement, and third is by $(\upsilon_d, \varsigma_d) \perp\!\!\!\perp \tilde{Z}$.

Step 2: Equivalent restrictions on (m_d, γ_d) without the modeling assumption.

I show that $\mathcal{H}^{EV/RA}(m_d, \gamma_d)$ is equivalent to:

$$\tilde{\mathcal{H}}^{EV/RA}(m_d, \gamma_d) = \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \ \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \ \ \gamma_d \overset{d}{=} \varsigma_d, \ \ m_d(\varsigma_d) = E_O[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\}$$

First fix $(m_d, \gamma_d) \in \mathcal{H}^{EV/RA}$. Then, there exist $(\upsilon_d, \varsigma_d)$ such that $m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d]$ a.s. and $\gamma_d \overset{d}{=} \varsigma_d$, $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\upsilon_d \in Sel(\mathbf{Y}(d))$. Hence $(m_d, \gamma_d) \in \tilde{\mathcal{H}}^{EV/RA}(m_d, \gamma_d)$.

Next, fix $(m_d, \gamma_d) \in \tilde{\mathcal{H}}^{EV/RA}(m, \gamma)$ and let $(\upsilon_d, \varsigma_d)$ be the corresponding selections in $Sel(\mathbf{Y}(d)) \times \left[Sel((\mathbf{S}(d), \tilde{Z})) \cap I \right]$ that generate them. I show that there exist $(\upsilon'_d, \varsigma'_d)$ such that: 1)$(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\upsilon'_d \in Sel(\mathbf{Y}(d))$; 2) $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ and $\varsigma'_d \overset{d}{=} \gamma_d$; 3) $(\upsilon'_d, \varsigma'_d) \perp\!\!\!\perp \tilde{Z}$.

Let $P_{\upsilon'_d, \varsigma'_d}$ be a distribution such that $\forall z \in \tilde{\mathcal{Z}}$, $P_{\upsilon'_d, \varsigma'_d}(\cdot|\varsigma'_d = s, \tilde{Z} = z) = P_{\upsilon_d, \varsigma_d}(\cdot|\varsigma_d = s, G = O)$ $\forall s \in \mathcal{S}$, and $P_{\varsigma'_d}(\cdot|\tilde{Z} = z) = P_{\varsigma_d}(\cdot)$. Note that these conditions fully specify $P_{\upsilon'_d, \varsigma'_d}$. I first show that there exist $(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\upsilon'_d \in Sel(\mathbf{Y}(d))$ such that $(\upsilon'_d, \varsigma'_d) \overset{d}{=} P_{\upsilon'_d, \varsigma'_d}$. I then show that $(\upsilon'_d, \varsigma'_d)$ fulfill conditions 2) $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ and $\varsigma'_d \overset{d}{=} \gamma_d$; and 3) $(\upsilon'_d, \varsigma'_d) \perp\!\!\!\perp \tilde{Z}$.

Recall that, as in the proof of Lemma A.2.4, by Lemma A.2.1 and Molchanov and Molinari (2018, Theorem 2.33), $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\}$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ P-a.s.:

$$P_{\upsilon_d, \varsigma_d}(B|\tilde{Z}) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y, S) \in B, D = d) \end{array} \right]. \tag{A.38}$$

Since $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times \left[Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap I\right]$, it must be that $P_{\varsigma_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$ $P-$a.s.. This a restriction on the marginal of $P_{\upsilon_d,\varsigma_d}$, hence for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ such that $B = \mathcal{Y} \times B_S$ for some $B_S \subseteq \mathcal{S}$:[7]

$$P_{\upsilon_d,\varsigma_d}(B|\tilde{Z}) = P_{\varsigma_d}(B_S|\tilde{Z}) = P_{\varsigma_d}(B_S) = P_{\upsilon_d,\varsigma_d}(B) \tag{A.39}$$

Where the first equality is by definition of a marginal distribution and $B = \mathcal{Y} \times B_S$, the second is because $P_{\varsigma_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$ $P-$a.s., and the third is by definition of a marginal distribution and $B = \mathcal{Y} \times B_S$.

By (A.38) and (A.39), $(\upsilon_d, \varsigma_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times \left[Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap I\right]$ only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ P-a.s.:

$$\begin{aligned} P_{\upsilon_d,\varsigma_d}(B|\tilde{Z}) &\geq \begin{bmatrix} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] ess\sup_{\tilde{Z}} P(S \in B_S, D = d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S]\mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y,S) \in B, D = d) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S]\mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P_O((Y,S) \in B, D = d) \end{bmatrix}. \end{aligned} \tag{A.40}$$

Observe that by (A.40) $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$P_{\upsilon_d,\varsigma_d}(B|G = O) \geq \begin{bmatrix} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y,S) \in B, D = d) \end{bmatrix}. \tag{A.41}$$

7. Note that the condition need not hold for *every* $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$, only for B s.t. $B = \mathcal{Y} \times B_S$ for some $B_S \subseteq \mathcal{S}$.

Then for any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ $P-$a.s.:

$$P_{\upsilon'_d,\varsigma'_d}(B) = P_{\upsilon'_d,\varsigma'_d}(B|\tilde{Z}) = P_{\upsilon_d,\varsigma_d}(B|G = O) \geq P_O((Y,S) \in B, D = d) \tag{A.42}$$

where the first equality is by $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d, \tilde{Z}) = P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d)$ and $P_{\varsigma'_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot)$, the second is by $P_{\upsilon'_d,\varsigma'_d}(\cdot|\varsigma'_d = s, \tilde{Z} = z) = P_{\upsilon_d,\varsigma_d}(\cdot|\varsigma_d = s, G = O)$, and the inequality is by (A.40).

For any $B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ such that $B = \mathcal{Y} \times B_S$ for some $B_S \subset \mathcal{S}$:

$$\begin{aligned} P_{\upsilon'_d,\varsigma'_d}(B|\tilde{Z}) &= P_{\varsigma'_d}(B_S|\tilde{Z}) = P_{\varsigma_d}(B_S) = P_{\upsilon_d,\varsigma_d}(B) \\ &\geq \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right). \end{aligned} \tag{A.43}$$

where the first equality follows by definition of a marginal distribution and $B = \mathcal{Y} \times B_S$, the second is by $P_{\varsigma'_d}(B_S|\tilde{Z}) = P_{\varsigma_d}(\cdot)$, third is by definition of a marginal distribution and $B = \mathcal{Y} \times B_S$, and the inequality is by (A.40).

By (A.42) and (A.43) $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$:

$$P_{\upsilon'_d,\varsigma'_d}(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P_O((Y,S) \in B, D = d) \end{array}\right]. \tag{A.44}$$

Then recall that by Lemma A.2.4, $\exists(\upsilon'_d, \varsigma'_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap \tilde{I}$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ (A.44) holds. Therefore, there exist $(\varsigma'_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\upsilon'_d \in Sel(\mathbf{Y}(d))$ such that $(\upsilon'_d, \varsigma'_d) \overset{d}{=} P_{\upsilon'_d,\varsigma'_d}$.

Next, note that since $P_{\upsilon'_d,\varsigma'_d}(\cdot) = P_{\upsilon_d,\varsigma_d}(\cdot|G=O)$, then $m_d(\varsigma'_d) = E_O[\upsilon'_d|\varsigma'_d]$ a.s. Because $P_{\varsigma'_d}(\cdot|\tilde{Z}) = P_{\varsigma_d}(\cdot) = P_{\varsigma'_d}(\cdot)$, $\varsigma'_d \overset{d}{=} \varsigma_d \overset{d}{=} \gamma_d$. Finally, because $(\upsilon'_d, \varsigma'_d, \tilde{Z}) \in Sel(\mathbf{Y}(d)) \times Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap \tilde{I}$, $(\upsilon'_d, \varsigma'_d) \perp\!\!\!\perp \tilde{Z}$. Therefore, if $(m_d, \gamma_d) \in \tilde{\mathcal{H}}^{EV/RA}(m_d, \gamma_d)$, then $(m_d, \gamma_d) \in \mathcal{H}^{EV/RA}(m_d, \gamma_d)$.

Hence:

$$\text{(A.45)} \quad \mathcal{H}^{EV/RA}(m_d, \gamma_d) = \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \ \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel(\mathbf{Y}(d)), \ \gamma_d \overset{d}{=} \varsigma_d, \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.}. \end{array} \right\}$$

Finally, impose $E[|Y(d)|] < \infty$. This can equivalently be restated as $Y(d) \in Sel^1(\mathbf{Y}(d))$. Then the identified set for (m_d, γ_d) under Assumptions RA and EV, and $E[|Y(d)|] < \infty$ is:

(A.46)

$$\mathcal{H}^{EV/RA/Int}(m_d, \gamma_d) = \left\{ \begin{array}{l} (m_d, \gamma_d) \in \mathcal{M}_d \times \mathcal{P}^{\mathcal{S}} : \ \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)), \ \gamma_d \overset{d}{=} \varsigma_d, \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}.$$

Step 3: Restrictions on (m, γ) without the modeling assumption.

Since the data never reveal $(S(0), Y(0))$ and $(S(1), Y(1))$ jointly, Assumptions RA and EV, and $E[|Y(d)|] < \infty$ do not impose cross-restrictions on them. Then the set of all (m, γ) consistent with the data Assumptions RA and EV, and $E[|Y(d)|] < \infty$ is:

$$\begin{aligned} &\mathcal{H}^{EV/RA/Int}(m, \gamma) = \mathcal{H}^{EV/RA/Int}(m_0, \gamma_0) \times \mathcal{H}^{EV/RA/Int}(m_1, \gamma_1) \\ &= \left\{ \begin{array}{l} (m, \gamma_d) \in \mathcal{M} \times (\mathcal{P}^{\mathcal{S}}))^2 : \ \forall d \in \{0,1\}, \ \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)), \ \gamma_d \overset{d}{=} \varsigma_d, \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}. \end{aligned} \tag{A.47}$$

Step 4: Identified set $\mathcal{H}(m, \gamma)$.

It only remains to impose Assumption MA. To do so, observe that a valid identified set is:

$$\begin{aligned}\mathcal{H}(m,\gamma) &= \mathcal{H}^{EV/RA/Int}(m,\gamma) \cap (\mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2) \\ &= \left\{ \begin{array}{l} (m,\gamma_d) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}}))^2 : \ \forall d \in \{0,1\}, \ \exists (\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)), \ \ \gamma_d \overset{d}{=} \varsigma_d, \ \ m_d(\varsigma_d) = E_O[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\}. \end{aligned} \tag{A.48}$$

Next note that for every $(m,\gamma) \in \mathcal{H}(m,\gamma)$, there exist selections $(\varsigma_0, \varsigma_1, \upsilon_0, \upsilon_1)$ that generate them and that are consistent with the data, modeling assumption, Assumptions RA and EV, and $E[|Y(d)|] < \infty$. Therefore, $\mathcal{H}(m,\gamma)$ is sharp. □

Proof of Theorem 1.3.1. The proof proceeds through a series of steps:

(1) Restrictions on m_d given a selection $\varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ are equivalently stated using the conditional Aumann expectation;

(2) The restrictions on m_d given ς_d are restated using the support function of the conditional Aumann expectation via the convexification property on non-atomic probability spaces;

(3) Restrictions on γ_d given a selection $\varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ are stated using Artstein's theorem;

(4) Restrictions on (m,γ) are shown to be invariant to the selection ς_d.

Steps 1 and 2 remove the need to search over selections $\upsilon_d \in Sel^1(\mathbf{Y}(d))$. Steps 3 allows the removal of search over selections $\varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$. This is formalized in Step 4.

Step 1: Restrictions on m_d given ς_d as a conditional Aumann expectation.

Fix an arbitrary $d \in \{0,1\}$ and ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap I$. Let $\sigma(\varsigma_d|G = O)$ be the sub-σ-algebra generated by ς_d given $\{\omega \in \Omega : G = O\}$. Let $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] := cl\{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$, where the closure is taken in L^1 space of all $\sigma(\varsigma_d|G = O)$-measurable functions. $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ exists, is a unique random set, and has at least one integrable selection. Since $\mathbf{Y}(d)$ is integrably bounded, so is $\mathbb{E}_O[\mathbf{Y}(d)|S(d)]$ (Molchanov (2017, Theorem 2.1.71)). ς_d is a measurable selection, hence a random vector. Therefore, the conditioning sub-σ-algebra $\sigma(\varsigma_d|G = O)$ of $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ is generated by a random vector and is thus countably generated. Then since $\mathbf{Y}(d)$ is integrably bounded and defined on $\mathbb{R}$, $\{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$ is a closed set, so $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] = \{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$ (Li and Ogura (1998, Theorem 1), Molchanov (2017, Theorem Section 2.1.6)).

Since $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] = \{E_O[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$, it is then immediate that:

$$\exists \upsilon_d \in Sel^1(\mathbf{Y}(d)) : \ \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \Leftrightarrow m_d(\varsigma_d) \in \mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] \text{ a.s.} \tag{A.49}$$

Therefore:

(A.50)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)),\ \ \gamma_d \overset{d}{=} \varsigma_d\ \ m_d(\varsigma_d) = E_O[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}$$

(A.51)

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \gamma_d \overset{d}{=} \varsigma_d,\ m_d(\varsigma_d) \in Sel^1(\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]) \end{array} \right\}$$

(A.52)

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \gamma_d \overset{d}{=} \varsigma_d,\ m_d(\varsigma_d) \in Sel(\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]) \end{array} \right\}$$

where the first line is by Lemma 1.3.1, the second is by (A.49) and the third follows since $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ is integrably bounded.

Step 2: Representation of restrictions on m_d given ς_d using the support function.

By assumption, the probability space is non-atomic. By Lemma A.2.2, P has no atoms over $\sigma(\varsigma_d|G = O)$ for any measurable selection ς_d. Since $E[|Y(d)|] < \infty$ for all $d \in \{0,1\}$, $\mathbf{Y}(d)$ is integrable. Thus, $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ is almost surely convex and equal to $\mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d]$ (Molchanov (2017, Theorem 2.1.77)). Therefore, $h_{\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]}(u) = h_{\mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d]}(u)$ a.s. for all $u \in \mathbb{R}$ by definition of the support function h. By $\mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d] = \mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$ and integrability of the latter, the former set is also integrable. It then follows that $h_{\mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d]}(u) = E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$ (Molchanov (2017, Theorem 2.1.72)). Hence, recalling that $\mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d] = \mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$, also $h_{\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]}(u) = \mathbb{E}_O[co(\mathbf{Y}(d))|\varsigma_d] = E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$.

Fix an arbitrary ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ and $\varsigma_d \overset{d}{=} \gamma_d$. Then:

$$
\begin{aligned}
& m_d(\varsigma_d) \in \mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d] \text{ a.s.} \\
\Leftrightarrow & \forall u \in \{-1, 1\} : \ \ u m_d(\varsigma_d) \leq h_{\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]}(u) \text{ a.s.} \\
\Leftrightarrow & \forall u \in \{-1, 1\} : \ \ u m_d(\varsigma_d) \leq E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s.}
\end{aligned} \tag{A.53}
$$

where the second line is by Rockafellar (1970, Theorem 13.1) and almost sure convexity of $\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]$, and the third is by $h_{\mathbb{E}_O[\mathbf{Y}(d)|\varsigma_d]}(u) = E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$. Moreover:

$$
\begin{aligned}
E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] &= E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d, D = d] P_O(D = d|\varsigma_d) \\
&\quad + E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d, D \neq d] P_O(D \neq d|\varsigma_d) \\
&= u E_O[Y|\varsigma_d, D = d] P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u) P_O(D \neq d|\varsigma_d) \\
&= u E_O[Y|S, D = d] P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u) P_O(D \neq d|\varsigma_d) \\
&= u \mu_d(\varsigma_d) P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u) P_O(D \neq d|\varsigma_d) \\
&= u \mu_d(\varsigma_d) \pi_{\gamma_d}(\varsigma_d) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(\varsigma_d))
\end{aligned} \tag{A.54}
$$

where the first equality is by LIE. The second follows because $co(\mathbf{Y}(d)) = \{Y\}$ whenever $D = d$, $h_{\{Y\}}(u) = uY$, and $co(\mathbf{Y}(d)) = co(\mathcal{Y})$ when $D \neq d$. The third is by observing that $P_O(\varsigma_d = S|D = d) = 1$ since $\varsigma_d \in Sel(\mathbf{S}(d))$ and $\mathbf{S}(d) = \{S\}$ when $D = d$. The fourth is by definition of μ_d and $P_O(\varsigma_d = S|D = d) = 1$. The final equality is by Lemma A.2.8. Then

observe that:

$$\begin{aligned} &\forall u \in \{-1,1\}: \quad u m_d(\varsigma_d) \leq E_O[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s.} \\ \Leftrightarrow &\forall u \in \{-1,1\}: \quad u m_d(\varsigma_d) \leq u\mu_d(\varsigma_d)\pi_{\gamma_d}(\varsigma_d) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(\varsigma_d)) \text{ a.s.} \\ \Leftrightarrow &\forall u \in \{-1,1\}: \quad u m_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)) \ \gamma_d\text{−a.e.} \end{aligned} \tag{A.55}$$

where the second line follows by (A.54) and the third by $\varsigma_d \overset{d}{=} \gamma_d$. Therefore:

(A.56)

$$\begin{aligned} &\mathcal{H}(m,\gamma) \\ &= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \ \gamma_d \overset{d}{=} \varsigma_d, \\ \forall u \in \{-1,1\}: \ u m_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1-\pi_{\gamma_d}(s)), \ \gamma_d\text{−a.e.} \end{array} \right\} \end{aligned}$$

Step 3: Representation of restriction on γ_d given ς_d using Artstein's theorem.

Note that for any $(m,\gamma) \in \mathcal{H}(m,\gamma)$, there exists $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I$ such that $\gamma_d \overset{d}{=} \varsigma_d$. I follow similar steps to those in the proof of Lemma A.2.4 to characterize restrictions imposed on γ_d by this condition.

By Lemma A.2.1, a distribution function characterizes a selection in $Sel((\mathbf{S}(d), \tilde{Z}))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S} \times \tilde{\mathcal{Z}}): \quad P((S(d), \tilde{Z}) \in B) \geq P((\mathbf{S}(d), \tilde{Z}) \subseteq B) \tag{A.57}$$

$$\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B|\tilde{Z}) \geq P(\mathbf{S}(d) \subseteq B|\tilde{Z}) \text{ a.s.} \tag{A.58}$$

where the second line follows by Molchanov and Molinari (2018, Theorem 2.33). Now consider the containment functional $P(\mathbf{S}(d) \subseteq B|\tilde{Z})$. If $B = \mathcal{S}$, $P(\mathbf{S}(d) \subseteq B|\tilde{Z}) = 1$. If

$B \subseteq \mathcal{S}$, then $P(\mathbf{S}(d) \subseteq B|\tilde{Z}) = P(S \subseteq B, D = d|\tilde{Z})$. Hence, $\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z}))$ such that $\gamma_d \in \mathcal{P}^{\mathcal{S}}$ and $\gamma_d \overset{d}{=} \varsigma_d$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B|\tilde{Z}) \geq P(S \subseteq B, D = d|\tilde{Z}) \text{ a.s.} \tag{A.59}$$

Since $(\varsigma_d, \tilde{Z}) \in I$, (A.59) is equivalent to:

(A.60)

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B) \geq ess\sup_{\tilde{Z}} P(S \subseteq B, D = d|\tilde{Z})$$

$$= \max\left(ess\sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right) \tag{A.61}$$

where the first line follows since, by definition of I, $\varsigma_d \perp\!\!\!\perp \tilde{Z}$. The second is by definition of $\tilde{Z}$ and $P(G = O) > 0$ given that two datasets are observed.

Therefore, write:

(A.62)

$$\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I \text{ s.t. } \gamma_d \overset{d}{=} \varsigma_d$$

$$\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}): \quad P(\varsigma_d \in B) \geq \max\left(ess\sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right)$$

By definition, if $\mathfrak{C}$ is a core determining class, (A.62) is equivalent to:

(A.63)

$$\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I \text{ s.t. } \gamma_d \overset{d}{=} \varsigma_d$$

$$\Leftrightarrow \forall B \in \mathfrak{C}: \quad P(\varsigma_d \in B) \geq \max\left(ess\sup_{Z} P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right)$$

Recall that for any $(m,\gamma) \in \mathcal{H}(m,\gamma)$, there exists $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I$ such that $\gamma_d \overset{d}{=} \varsigma_d$. Then each such γ_d must satisfy the conditions (A.62). Hence, by the characterization of $\mathcal{H}(m,\gamma)$ in (A.56) it follows that:

(A.64)

$$\mathcal{H}(m,\gamma)$$
$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I,\ \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right),\ \gamma_d \overset{d}{=} \varsigma_d, \\ \forall u \in \{-1,1\} :\ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)),\ \gamma_d\text{−a.e.} \end{array} \right\}$$

Step 4: Removing search over selections ς_d.

It remains to show that $\mathcal{H}(m,\gamma) = \mathcal{H}^I$ where:

(A.65)

$$\mathcal{H}^I = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\} :\ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)),\ \gamma_d\text{−a.e.} \end{array} \right\}$$

First, pick $(m,\gamma) \in \mathcal{H}(m,\gamma)$. Then, since the conditions imposed on elements of $\mathcal{H}^I$ is a strict subset of those imposed on elements of $\mathcal{H}(m,\gamma)$, it must be that $(m,\gamma) \in \mathcal{H}^I$. Conversely, pick $(m,\gamma) \in \mathcal{H}^I$. By (A.62) and Lemma A.2.1, for every $d \in \{0,1\}$ there exists ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d), \tilde{Z})) \cap I$, $\gamma_d \overset{d}{=} \varsigma_d$. Therefore $(m,\gamma) \in \mathcal{H}(m,\gamma)$ and thus $\mathcal{H}^I = \mathcal{H}(m,\gamma)$.

If $\mathfrak{C}$ is a core determining class, then by similar arguments and (A.63) it follows that:

(A.66)

$$\mathcal{H}^I = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \forall B \in \mathfrak{C}, \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\}: \ \ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)), \ \gamma_d\text{−a.e.} \end{array} \right\}$$

□

Proof of Theorem 1.3.2. The proof proceeds through a series of steps:

(1) Characterizing $\mathcal{H}(m,\gamma)$;

(2) Proving that $T(m,\gamma)$ is jointly continuous;

(3) Proving that $\mathcal{H}(m,\gamma)$ is convex;

(4) Proving that $\mathcal{H}(m,\gamma)$ is compact;

(5) Proving that $\mathcal{H}(\tau)$ is an interval.

Step 1: Characterizing $\mathcal{H}(m,\gamma)$.

For any selection $\upsilon_d \in \mathbf{Y}(d)$, $\upsilon_d \in \mathcal{Y}$ so $E[|\upsilon_d|] \leq |\sup \mathcal{Y}| < \infty$ where the strict inequality follows by boundedness of $\mathcal{Y}$. Hence, $\mathbf{Y}(d)$ is integrably bounded. Since $\mathcal{S} = \{1, 2, \ldots, k\}$, represent γ_d as an element of the k−dimensional simplex $\Delta(k)$ and $m_d \in \mathcal{Y}^k$. Let $\gamma_d(s)$ and $m_d(s)$ denote the s−th element of the corresponding vectors.

Then:

(A.67)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \forall u \in \{-1,1\}, \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ um_d(s) \leq u\mu_d(s)\pi_{\gamma_d}(s) + h_{co(\mathcal{Y})}(u)(1 - \pi_{\gamma_d}(s)) \ \gamma_d\text{−a.e.} \end{array} \right\}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\}, \forall B \in \mathcal{C}(\mathcal{S}), \forall u \in \{-1,1\}, \\ \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ um_d(s) \leq u\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right) \ \gamma_d\text{−a.e.} \end{array} \right\}$$

where the first line is by Theorem 1.3.1. The second is by definition of $\pi_{\gamma_d}(s)$ and γ_d being supported on $\mathcal{S}$ with $|\mathcal{S}| < \infty$.

$\mathcal{S}$ is closed by definition. Since it is finite, it is bounded. Hence, $\mathbf{S}(d)$ is almost surely compact, by definition. Then, by Beresteanu, Molchanov, and Molinari (2012, Lemma B.1) $\{\{s\} : s \in \mathcal{S}\}$ is a core-determining class for the containment functional of $\mathbf{S}(d)$.

Therefore:

(A.68)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S},\, \forall u \in \{-1,1\}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ u m_d(s) \leq u\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \end{array} \right\}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ -m_d(s) \leq -\mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} - Y_L\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \\ m_d(s) \leq \mu_d(s)\frac{P_O(S=s,D=d)}{\gamma_d(s)} + Y_U\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right)\ \gamma_d\text{−a.e.} \end{array} \right\}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\, \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ (m_d(s) - Y_L)\,\gamma_d(s) \geq (E_O[Y|S=s,D=d] - Y_L)\,P_O(S=s,D=d)\ , \\ (Y_U - m_d(s))\,\gamma_d(s) \geq (Y_U - E_O[Y|S=s,D=d])\,P_O(S=s,D=d) \end{array} \right\}.$$

where the first line is by Theorem 1.3.1 and (A.67), the second line is by definition of $h_{co(\mathcal{Y})}(u)$, and the third is by rearrangement and definition of $\mu_d(s)$.

Step 2: T is jointly continuous.

Endow the set of reals with its natural topology, making it a locally convex topological vector space (t.v.s.). By bilinearity of the Riemann-Stieltjes integral in the integrand and integrator, $T(m,\gamma)$ is a bilinear map. Since T is a bilinear map in a finite-dimensional space, it is separately continuous in each argument. Note that $T : \mathbb{R}^{2d_s} \times \mathbb{R}^{2d_s} \to \mathbb{R}$ and that $\mathbb{R}^{2d_s}$ is Polish (separable and completely metrizable), and hence metrizable. By a corollary of the first Baire category theorem, every Polish space is a Baire space, so $\mathbb{R}^{2d_s}$ is

a Baire space (Willard (2004, Corollary 25.4)). By Corollary A.2.1, T is jointly continuous since every separately continuous bilinear map from a product of a Baire space and a metrizable space to a locally convex t.v.s. is jointly continuous.

Step 3: $\mathcal{H}(m,\gamma)$ is convex.

Define the following set:

(A.69)

$$\mathcal{H}^{o}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (int(\Delta(k)))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ (m_d(s) - Y_L)\,\gamma_d(s) \geq (E_O[Y|S=s,D=d] - Y_L)\,P_O(S=s,D=d)\ , \\ (Y_U - m_d(s))\,\gamma_d(s) \geq (Y_U - E_O[Y|S=s,D=d])\,P_O(S=s,D=d) \end{array} \right\}$$

noting that $\gamma \in (int(\Delta(k)))^2$ so $\gamma_d(s) > 0$ for any $s \in \mathcal{S}$ and $d \in \{0,1\}$. It is immediate that $cl\left(\mathcal{H}^{o}(m,\gamma)\right) = \mathcal{H}(m,\gamma)$. I first prove that $\mathcal{H}^{o}(m,\gamma)$ is convex, which is sufficient for $\mathcal{H}(m,\gamma)$ to be convex since the closure of a convex set is convex. Pick any $(m,\gamma),(m',\gamma') \in \mathcal{H}^{o}(m,\gamma)$ and fix $a \in (0,1)$. It remains to show that $a(m,\gamma) + (1-a)(m',\gamma') \in \mathcal{H}^{o}(m,\gamma)$.

$\mathcal{M}^A$ is convex by assumption so $am + (1-a)m' \in \mathcal{M}^A$. $\Delta(k)$ is the k-dimensional simplex, and thus convex. The interior of a convex set is convex, so $int(\Delta(k))$ and $(int(\Delta(k)))^2$ are convex. Therefore, $a\gamma + (1-a)\gamma \in (int(\Delta(k)))^2$. Observe that for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$a\gamma_d(s) + (1-a)\gamma'_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right) \quad \text{(A.70)}$$

since both γ_d and γ'_d satisfy the same condition. Next, note that for any $d \in \{0,1\}$ and $s \in \mathcal{S}$, recalling that $\gamma_d(s) > 0$ and $\gamma'_d(s) > 0$:

$$\begin{aligned}
&\frac{a\gamma'_d(s) + (1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)} - \frac{1}{a\gamma_d(s) + (1-a)\gamma'_d(s)} \\
&= \frac{(a\gamma'_d(s) + (1-a)\gamma_d(s))(a\gamma_d(s) + (1-a)\gamma'_d(s)) - \gamma_d(s)\gamma'_d(s)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \\
&= \frac{(a^2 + (1-a)^2 - 1)\gamma'_d(s)\gamma_d(s) + a(1-a)(\gamma'_d(s)^2 + \gamma_d(s)^2)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \\
&= \frac{2a(a-1)\gamma'_d(s)\gamma_d(s) + a(1-a)(\gamma'_d(s)^2 + \gamma_d(s)^2)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \\
&= \frac{a(1-a)(\gamma_d(s) - \gamma_d(s)')^2}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \geq 0.
\end{aligned} \tag{A.71}$$

Then for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned}
&am_d(s) + (1-a)m'_d(s) - Y_L \\
&\geq (E_O[Y|S=s, D=d] - Y_L)\, P_O(S=s, D=d) \left(\frac{a}{\gamma_d(s)} + \frac{1-a}{\gamma'_d(s)}\right) \\
&= (E_O[Y|S=s, D=d] - Y_L)\, P_O(S=s, D=d) \left(\frac{a\gamma'_d(s) + (1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)}\right) \\
&\geq (E_O[Y|S=s, D=d] - Y_L)\, \frac{P_O(S=s, D=d)}{a\gamma_d(s) + (1-a)\gamma'_d(s)}.
\end{aligned} \tag{A.72}$$

where the first line follows by $(m,\gamma), (m',\gamma') \in \mathcal{H}^o(m,\gamma)$ and $\gamma_d(s) > 0$ and $\gamma'_d(s) > 0$, second is by observation and the third is by (A.71). Therefore:

$$\begin{aligned}
&(am_d(s) + (1-a)m'_d(s) - Y_L)\,(a\gamma_d(s) + (1-a)\gamma'_d(s)) \\
&\geq (E_O[Y|S=s, D=d] - Y_L)\, P_O(S=s, D=d)
\end{aligned} \tag{A.73}$$

Similarly, for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} & Y_U - a m_d(s) - (1-a) m'_d(s) \\ & \geq \left(Y_U - E_O[Y|S=s, D=d]\right) P_O(S=s, D=d) \left(\frac{a}{\gamma_d(s)} + \frac{1-a}{\gamma'_d(s)}\right) \\ & = \left(Y_U - E_O[Y|S=s, D=d]\right) P_O(S=s, D=d) \left(\frac{a\gamma'_d(s) + (1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)}\right) \\ & \geq \left(Y_U - E_O[Y|S=s, D=d]\right) \frac{P_O(S=s, D=d)}{a\gamma_d(s) + (1-a)\gamma'_d(s)}. \end{aligned} \tag{A.74}$$

where the first line follows by $(m,\gamma), (m',\gamma') \in \mathcal{H}^o(m,\gamma)$ and $\gamma_d(s) > 0$ and $\gamma'_d(s) > 0$, second is by observation and the third is by (A.71). Hence:

$$\begin{aligned} & \left(Y_U - a m_d(s) - (1-a) m'_d(s)\right) \left(a\gamma_d(s) + (1-a)\gamma'_d(s)\right) \\ & \geq \left(Y_U - E_O[Y|S=s, D=d]\right) P_O(S=s, D=d) \end{aligned} \tag{A.75}$$

By verifying all conditions, $a(m,\gamma) + (1-a)(m',\gamma') \in \mathcal{H}^o(m,\gamma)$, and $\mathcal{H}^o(m,\gamma)$ is convex. Since closure preserves convexity, $\mathcal{H}(m,\gamma) = cl\left(\mathcal{H}^o(m,\gamma)\right)$ is convex.

Step 4: $\mathcal{H}(m,\gamma)$ is compact

$\mathcal{Y}$ is a bounded and $\mathcal{S}$ is finite, by assumption. It is then immediate that $\mathcal{H}(m,\gamma)$ is bounded since $\mathcal{H}(m,\gamma) \subseteq \mathcal{Y}^k \times (\Delta(k))^2$. That $\mathcal{H}(m,\gamma)$ is closed follows by definition of a closure and the fact that $\mathcal{H}(m,\gamma) = cl\left(\mathcal{H}^o(m,\gamma)\right)$. Then, $\mathcal{H}(m,\gamma)$ is compact.

Step 5: $\mathcal{H}(\tau)$ is an interval.

T was shown to be a continuous map, so it preserves connectedness. Hence, $\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\}$ is a connected set. Since $\mathcal{H}(\tau) \subseteq \mathbb{R}$, it is an interval.

Continuous images preserve compactness, so the $\mathcal{H}(\tau)$ is a compact interval, and:

$$\mathcal{H}(\tau) = \left[\inf_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma), \sup_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma) \right] \tag{A.76}$$

$$= \left[\min_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma), \max_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m,\gamma) \right] \tag{A.77}$$

where the second line follows by continuity of T and compactness of $\mathcal{H}(m,\gamma)$. □

Proof of Corollary 1.3.1. By observation, it is immediate that iterated and joint minima and maxima search over all values of the set $\mathcal{H}(m,\gamma)$. Thus:

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}) \right] \tag{A.78}$$

$$= \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \min_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \max_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}) \right]. \tag{A.79}$$

By definition of L_γ and U_γ:

$$\forall \tilde{m} \in \mathcal{H}(m|\gamma): \quad T(L_\gamma,\gamma) \leq T(\tilde{m},\gamma) \leq T(U_\gamma,\gamma). \tag{A.80}$$

Therefore:

$$\begin{aligned} \min_{\tilde{m}\in\mathcal{H}(m|\gamma)} T(\tilde{m},\gamma) &= T(L_\gamma,\gamma) \\ \max_{\tilde{m}\in\mathcal{H}(m|\gamma)} T(\tilde{m},\gamma) &= T(U_\gamma,\gamma). \end{aligned} \tag{A.81}$$

□

Proof of Theorem 1.3.3. Consider the following two estimators.

$$\mathcal{H}_n(\tau) = \left[\min_{\tilde{\gamma} \in \mathcal{H}_n(\gamma)} \min_{\tilde{m} \in \mathcal{H}_n(m|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}_n(\gamma)} \max_{\tilde{m} \in \mathcal{H}_n(\tilde{m}|\tilde{\gamma})} T(m, \tilde{\gamma}) \right]. \tag{A.82}$$

$$\tilde{\mathcal{H}}_n(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m}, \tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) \right]. \tag{A.83}$$

By the proof of Corollary 1.3.1, $\mathcal{H}_n(\tau)$ and $\tilde{\mathcal{H}}_n(\tau)$ are numerically equivalent. Similarly, the estimator $\left[\min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(L_{\tilde{\gamma}}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(U_{\tilde{\gamma}}, \tilde{\gamma})\right]$ obtained using minimal and maximal selectors in (A.82) is numerically equivalent to $\mathcal{H}_n(\tau)$ and hence numerically equivalent to $\tilde{\mathcal{H}}_n(\tau)$. Therefore, consistency of $\tilde{\mathcal{H}}_n(\tau)$ yields consistency of $\mathcal{H}_n(\tau)$ both with and without using minimal and maximal selectors.

It remains to prove $d_H(\tilde{\mathcal{H}}_n(\tau), \mathcal{H}(\tau)) \xrightarrow{\text{p}} 0$. For this, note that $\mathcal{H}(\tau)$ and $\tilde{\mathcal{H}}_n(\tau)$ are both closed intervals by Theorem 1.3.2 and the definition of the latter. Then, by definition of the Hausdorff distance, it is sufficient to show that boundaries of $\tilde{\mathcal{H}}_n(\tau)$ converge in probability to boundaries of $\mathcal{H}(\tau)$ as $n \longrightarrow \infty$.

Considering the upper bounds of $\mathcal{H}(\tau)$ and $\tilde{\mathcal{H}}_n(\tau)$, I show that for any $\varepsilon > 0$:

$$\limsup_{n \longrightarrow \infty} P\left(\left| \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) - \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) \right| > \varepsilon \right) = 0 \tag{A.84}$$

and the argument for the lower bounds is symmetric. Fix any $\varepsilon > 0$ and note that:

$$\begin{aligned} &\left| \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}_n(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) - \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) \right| \\ &\leq \sup_{||(m,\gamma)-(m',\gamma')|| \leq d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma))} |T(m,\gamma) - T(m',\gamma')|. \end{aligned} \tag{A.85}$$

$\mathcal{M}^A$ is a set of finite-dimensional vectors, defined by finitely many linear equality and inequality constraints, thus a polytope, and therefore compact. $\mathcal{M}^A \times (\Delta(k))^2$ is then also compact. Proof of Theorem 1.3.2 shows that T is a jointly continuous functional under maintained assumptions. Hence, $T : \mathcal{M}^A \times (\Delta(k))^2 \to \mathbb{R}$ is uniformly continuous over its domain by the Heine-Cantor theorem. For the fixed ε, let $\varepsilon' = 2\varepsilon > 0$. By uniform continuity, there exists a $\delta' > 0$ such that $||(m,\gamma) - (m',\gamma')|| < \delta'$ implies $|T(m,\gamma) - T(m',\gamma')| < \varepsilon'$. Let $\delta = \delta'/2 > 0$. If $|T(m,\gamma) - T(m',\gamma')| \geq \varepsilon' > \varepsilon$ it must be that $d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) \geq \delta' > \delta$. Therefore:

(A.86)
$$P\left(\left|\max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}_n(m,\gamma)} T(\tilde{m},\tilde{\gamma}) - \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right| > \varepsilon\right) \leq P\left(d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) > \delta\right)$$

Thus, to prove (A.84), it is sufficient to show that given $\delta > 0$:

(A.87)
$$\limsup_{n\to\infty} P(d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) > \delta) = 0.$$

Therefore to prove (A.84), it is sufficient to prove $d_H(\mathcal{H}(m,\gamma), \mathcal{H}_n(m,\gamma)) \xrightarrow{p} 0$. To do so, I adapt the arguments in Russell (2021, Theorem 2). Let μ_d be a k-dimensional vector with components $\mu_d(s) = E_O[Y|S = s, D = d]$. Let η_d be a $k \times |\tilde{\mathcal{Z}}|$ matrix with the element $(s,\tilde{z})$ being $\eta_d(s,\tilde{z}) = P(S = s, D = d|\tilde{Z} = z)$. Finally, collect $\beta = (\mu_0, \mu_1, \eta_0, \eta_1) \in \mathfrak{B}$. Let β_n be the plug-in estimator of β. By elementary arguments $\beta_n \xrightarrow{p} \beta$ as $n \to \infty$.

By Assumption E *iii)*, $\mathcal{M}^A = \{m \in \mathcal{M} : h(m,\beta) \geq 0,\ \ g(m,\beta) = 0\}$ for some known linear functions g and h. Then:

(A.88)

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z\in\mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s,D=d)\right), \\ (m_d(s) - Y_L)\gamma_d(s) \geq \left(E_{O,n}[Y|S=s,D=d] - Y_L\right) P_{O,n}(S=s,D=d), \\ \left(Y_U - m_d(s)\right)\gamma_d(s) \geq \left(Y_U - E_{O,n}[Y|S=s,D=d]\right) P_{O,n}(S=s,D=d) \end{array} \right\}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : \\ h(m,\beta) \geq 0,\ \ g(m,\beta) = 0,\ \ \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\max_{z\in\mathcal{Z}} P_E(S=s, D=d|Z=z), P_O(S=s,D=d)\right), \\ (m_d(s) - Y_L)\gamma_d(s) \geq \left(E_{O,n}[Y|S=s,D=d] - Y_L\right) P_{O,n}(S=s,D=d), \\ \left(Y_U - m_d(s)\right)\gamma_d(s) \geq \left(Y_U - E_{O,n}[Y|S=s,D=d]\right) P_{O,n}(S=s,D=d) \end{array} \right\}$$

$$= \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : h(m,\beta) \geq 0,\ \ g(m,\beta) = 0, \\ \forall d \in \{0,1\},\ \forall s \in \mathcal{S},\ \forall \tilde{z} \in \tilde{\mathcal{Z}}, \\ \gamma_d(s) \geq \eta_d(s,\tilde{z}), \\ (m_d(s) - Y_L)\gamma_d(s) \geq \mu_d(s)\eta_d(s, \sup\tilde{\mathcal{Z}} + 1)\ , \\ \left(Y_U - m_d(s)\right)\gamma_d(s) \geq (Y_U - \mu_d(s))\eta_d(s, \sup\tilde{\mathcal{Z}} + 1) \end{array} \right\}$$

where the first line follows by Theorem 1.3.2 and Assumption E *ii)*, the second line is by definition of $\mathcal{M}^A$, and the third is by definition η_d and μ_d and $\tilde{Z}$. Hence, $\mathcal{H}(m,\gamma)$ can be equivalently represented through a set of equality and inequality constraints as:

$$\mathcal{H}(m,\gamma) = \left\{ \ (m,\gamma) \in \mathcal{Y}^{2k} \times (\Delta(k))^2 : \tilde{h}(m,\gamma,\beta) \geq 0,\ \ g(m,\beta) = 0 \ \right\}. \tag{A.89}$$

where $\tilde{h}(m, \gamma, \beta)$ collects all linear inequality restrictions $h(m, \beta) \geq 0$ and remaining linear and bilinear inequality constraints in (A.88).

Next, convert all inequality constraints $\tilde{h}(m, \gamma, \beta)$ to equality constraints by introducing slackness parameters $\lambda_t \in [0, 1]$ for each inequality constraint, as in Shi and Shum (2015, Remark pp. 497).[8] Denote by λ the vector of all slackness parameters, and let $\theta = (m, \gamma, \lambda) \in \mathfrak{T}$ be a vector of dimension $d_\theta \times 1$. Write all converted equality constraints and existing equality constraints $g(m, \beta)$ as $\tilde{g}(\theta, \beta) = 0$. Define also the inequality constraints $h_\lambda(\theta) \geq 0$, which collect non-negativity constraints $\lambda_t \geq 0$. Now define $\Theta = \{\theta : \tilde{g}(\theta, \beta) = 0, \ \ h_\lambda(\theta) \geq 0.\}$. Under the assumptions, both $\mathcal{H}(m, \gamma)$ and Θ are non-empty. Therefore, equivalently write:

$$\Theta = \underset{\theta \in \mathfrak{T}: h_\lambda(\theta) \geq 0}{\arg\min} \ \tilde{g}(\theta, \beta)' \tilde{g}(\theta, \beta) \tag{A.90}$$

and let the corresponding estimator be:

$$\Theta_n = \underset{\theta \in \mathfrak{T}: h_\lambda(\theta) \geq 0}{\arg\min} \ \tilde{g}(\theta, \beta_n)' \tilde{g}(\theta, \beta_n). \tag{A.91}$$

Note that for any $(m, \gamma) \in \mathcal{H}(m, \gamma)$ if and only if $(m, \gamma, \lambda) \in \Theta$ for some feasible λ. Then, the projection of Θ onto (m, γ) two components is $\mathcal{H}(m, \gamma)$. Therefore, whenever $\tilde{\mathcal{H}}_n(m, \gamma) \neq \emptyset$, $\tilde{\mathcal{H}}_n(m, \gamma)$ is numerically equivalent to the projection of Θ_n onto (m, γ). Moreover, since $\beta_n \xrightarrow{p} \beta$ as $n \to \infty$, $P(\tilde{\mathcal{H}}_n(m, \gamma) \neq \emptyset) \to 1$ (see Yildiz (2012, Footnote 10)). Thus, for (A.87) and therefore (A.84), it is sufficient to show that $d_H(\Theta_n, \Theta) \xrightarrow{p} 0$.

8. Note that for proofs of consistency, it is sufficient to just add slackness parameters to each inequality constraint.

This follows immediately by verifying the conditions of Shi and Shum (2015, Theorem 2.1).

First, the preceding arguments argue that $\beta_n \xrightarrow{p} \beta$. Second, for $d \in \{0,1\}$, $\mu_d, m_d \in [0,1]^k$, $\eta_d, \gamma_d \in \Delta(k)$, $\lambda_t \in [0,1]$ for all $t < \infty$, hence the parameter spaces for $\mathfrak{T}$ and $\mathfrak{B}$ are compact. Third, $\tilde{g}(\cdot, \beta)$ is continuously differentiable for $\beta \in \mathfrak{B}$ as it is bilinear in θ; $h_\lambda(\cdot)$ is linear in θ and hence continuous. Applying identical arguments of Step 4 in the proof Russell (2021, Theorem 2) then yields $d_H(\Theta_n, \Theta) \xrightarrow{p} 0$.

□

Proof of Proposition 1.4.1. I show that $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$, which immediately yields $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$.

The data never reveal $(Y(0), Y(1))$ jointly, so the data and assumptions do not impose cross-restrictions on $Y(0)$ and $Y(1)$. Then the identified set for $P_{Y(0),Y(1)}$ *given* $P_{Y(1)}$ and $P_{Y(0)}$ is the set of all joint distributions consistent with the marginals $P_{Y(1)}$ and $P_{Y(0)}$. The identified set for $P_{Y(0),Y(1)}$ is the union of such sets over all possible $(P_{Y(0)}, P_{Y(1)})$.

To that end, let $\Pi(\nu_0, \nu_1)$ be the set of couplings of probability measures ν_0 and ν_1 defined as (Villani et al. (2009, Definition 1.1)):

$$\Pi(\nu_0, \nu_1) = \left\{ \delta \in \mathcal{P}^{\mathcal{Y}} \times \mathcal{P}^{\mathcal{Y}} : \ \forall A \subseteq \mathcal{Y} \quad \begin{array}{l} \delta(A \times \mathcal{Y}) = \nu_0(A), \\ \delta(\mathcal{Y} \times A) = \nu_1(A) \end{array} \right\}. \tag{A.92}$$

$\Pi(\nu_0, \nu_1)$ is always non-empty (Galichon (2018, Section 2.1)). Equivalently, the identified set for $P_{Y(0),Y(1)}$ *given* $P_{Y(1)}$ and $P_{Y(0)}$ is $\Pi(P_{Y(0)}, P_{Y(1)})$. Using the identified sets for the marginals $P_{Y(d)} \in \mathcal{H}^O(P_{Y(d)})$ for $d \in \{0,1\}$, the identified set $\mathcal{H}^O(P_{Y(0),Y(1)})$ is then

the union of all possible couplings:

$$\mathcal{H}^O(P_{Y(0),Y(1)}) = \bigcup_{(\nu_0,\nu_1)\in\mathcal{H}^O(P_{Y(0)})\times\mathcal{H}^O(P_{Y(1)})} \Pi(\nu_0,\nu_1). \tag{A.93}$$

Similarly for $\mathcal{H}(P_{Y(0),Y(1)})$:

$$\mathcal{H}(P_{Y(0),Y(1)}) = \bigcup_{(\nu_0,\nu_1)\in\mathcal{H}(P_{Y(0)})\times\mathcal{H}(P_{Y(1)})} \Pi(\nu_0,\nu_1). \tag{A.94}$$

Lemma A.2.5 shows that $\mathcal{H}^O(P_{Y(d)}) = \mathcal{H}(P_{Y(d)})$ for any $d \in \{0,1\}$. That $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$ follows.

Next, observe that τ is a functional of $P_{Y(0),Y(1)}$. It is then immediate that $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$,

Remark A.2.3. The same result may be obtained directly by defining the random set:

$$(\mathbf{Y},\mathbf{S}) = \begin{cases} \mathcal{S}\times\{S\}\times\mathcal{Y}\times\{Y\}, \text{ if } (D,G)=(1,O) \\ \mathcal{S}\times\{S\}\times\mathcal{Y}\times\mathcal{Y}, \text{ if } (D,G)=(1,E) \\ \{S\}\times\mathcal{S}\times\{Y\}\times\mathcal{Y}, \text{ if } (D,G)=(0,O) \\ \{S\}\times\mathcal{S}\times\mathcal{Y}\times\mathcal{Y}, \text{ if } (D,G)=(0,E) \end{cases} \tag{A.95}$$

which summarizes all information on $(S(0),S(1),Y(0),Y(1))$, and retracing the steps of Lemmas A.2.3, A.2.4 and A.2.5 for the joint distribution $P_{Y(0),Y(1)}$.

□

PROOF OF LEMMA 1.4.1.

$$d(\tilde{\mathcal{H}}^{O/A},\tau) = \inf\left\{\|t-\tau\| : t \in \tilde{\mathcal{H}}^{O/A}\right\} \leq \inf\left\{\|t-\tau\| : t \in \tilde{\mathcal{H}}\right\} = d(\tilde{\mathcal{H}},\tau) \tag{A.96}$$

where the inequality follows by $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$. □

Lemma A.2.10. *Let Assumptions RA, and EV hold. Suppose that $\mathcal{S}$ is a finite set.*

i) Suppose that Assumption LIV holds. Then for $s \in \mathcal{S}$:

$$\begin{aligned}
L_\gamma(s) = \Bigg(& \min_{s' \geq s} E_O[Y|S=s', D=0]\frac{P_O(S=s', D=0)}{\gamma_0(s')} + 1 - \frac{P_O(S=s', D=0)}{\gamma_0(s')}, \\
& \max_{s' \leq s} E_O[Y|S=s', D=1]\frac{P_O(S=s', D=1)}{\gamma_1(s')}\Bigg), \\
U_\gamma(s) = \Bigg(& \max_{s' \leq s} E_O[Y|S=s', D=0]\frac{P_O(S=s', D=0)}{\gamma_0(s')}, \\
& \min_{s' \geq s} E_O[Y|S=s', D=1]\frac{P_O(S=s', D=1)}{\gamma_1(s')} + 1 - \frac{P_O(S=s', D=1)}{\gamma_1(s')}\Bigg),
\end{aligned}$$

ii) Suppose that Assumption TI holds. Then $L_{\gamma'} = (m^{TI,L,\gamma'}, m^{TI,L,\gamma'})$ and $U_{\gamma'} = (m^{TI,U,\gamma'}, m^{TI,U,\gamma'})$ for:

$$\begin{aligned}
m(s)^{TI,L,\gamma'} &= m(s)^{L,\gamma'}\mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] + m(s)^{U,\gamma'}\mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] \\
m(s)^{TI,U,\gamma'} &= m(s)^{L,\gamma'}\mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] + m(s)^{U,\gamma'}\mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)],
\end{aligned} \tag{A.97}$$

where:

(A.98)
$$m(s)^{L,\gamma'} = \min_{d\in\{0,1\}} E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma'_d(s)},$$
$$m(s)^{U,\gamma'} = \max_{d\in\{0,1\}} E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma'_d(s)} + 1 - \frac{P_O(S=s,D=d)}{\gamma'_d(s)}.$$

Proof. *i)*

Fix any γ' such that there exists $(m,\gamma') \in \mathcal{H}(m,\gamma)$. Then $\mathcal{H}(m|\gamma') \neq \emptyset$. By bounded $\mathcal{Y}$, $h_{co(\mathcal{Y})}(-1) = 0$ and $h_{co(\mathcal{Y})}(1) = 1$. By Theorem 1.3.2, $\forall s \in \mathcal{S}$ restrictions imposed by data on $m_d(s)$ can then be equivalently stated for $d \in \{0,1\}$ as:

(A.99)
$$m_d(s) \in \left[E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma'_d(s)}, \right.$$
$$\left. E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma'_d(s)} + 1 - \frac{P_O(S=s,D=d)}{\gamma'_d(s)}\right].$$

By Manski and Pepper (2000a, Proposition 1) under Assumption LIV the sharp bound on $m_d(s)$ is:

(A.100)
$$m_d(s) \geq m_d(s)^{LIV,L,\gamma'} := \sup_{s'\leq s} E_O[Y|S=s',D=d]\frac{P_O(S=s',D=d)}{\gamma'_d(s')}$$
$$m_d(s) \leq m_d(s)^{LIV,U,\gamma'} := \inf_{s'\geq s} E_O[Y|S=s',D=d]\frac{P_O(S=s',D=d)}{\gamma'_d(s')}$$
$$+ 1 - \frac{P_O(S=s',D=d)}{\gamma'_d(s')}.$$

First, note that both $m_d^{LIV,L,\gamma'}$ and $m_d^{LIV,U,\gamma'}$ are non-decreasing in s by definition for all $d \in \{0,1\}$. Thus, $L_{\gamma'} := (m_0^{LIV,U,\gamma'}, m_1^{LIV,L,\gamma'}) \in \mathcal{M}^A$ and $U_{\gamma'} := (m_0^{LIV,L,\gamma'}, m_1^{LIV,U,\gamma'}) \in \mathcal{M}^A$. Hence $(L_{\gamma'},\gamma'), (U_{\gamma'},\gamma') \in \mathcal{H}(m,\gamma)$. Since γ' was arbitrary, $(L_{\gamma'},\gamma'), (U_{\gamma'},\gamma') \in \mathcal{H}(m,\gamma)$

for any $\gamma' \in \mathcal{H}(\gamma)$. Therefore, $L_{\gamma'}$ and $U_{\gamma'}$ are selectors of $\mathcal{H}(m|\cdot)$. Then, observe that T is non-decreasing in $m_1(s)$ and non-increasing in $m_0(s)$ for each $s \in \mathcal{S}$. Therefore, $\forall m \in \mathcal{H}(m|\gamma')\ T(L_{\gamma'}, \gamma') \leq T(m, \gamma')$, so $L_{\gamma'}$ is a minimal selector with respect to T. Similarly, $\forall m \in \mathcal{H}(m|\gamma')\ T(U_{\gamma'}, \gamma') \geq T(m, \gamma')$, so $U_{\gamma'}$ is a maximal selector with respect to T. Since $|\mathcal{S}|$ is a compact set, infima and suprema may be replaced by minima and maxima.

ii)

As in proof of *i)*, fix any γ' such that there exists $(m, \gamma') \in \mathcal{H}(m, \gamma)$, so $\mathcal{H}(m|\gamma') \neq \emptyset$. Assumption TI maintains that $m_1 = m_0$. Then write for any $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$T(m, \gamma') = \int_{\mathcal{S}} m_1(s) d\gamma_1'(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0'(s) = \int_{\mathcal{S}} m_d(s)(d\gamma_1'(s) - d\gamma_0'(s)). \tag{A.101}$$

Define:

(A.102)

$$m(s)^{L,\gamma'} := \min_{d \in \{0,1\}} E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)},$$
$$m(s)^{U,\gamma'} := \max_{d \in \{0,1\}} E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)} + 1 - \frac{P_O(S = s, D = d)}{\gamma_d'(s)}.$$

Next let for any $s \in \mathcal{S}$:

(A.103)

$$m(s)^{TI,L,\gamma'} := m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)]$$
$$m(s)^{TI,U,\gamma'} := m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)]$$

and $L_{\gamma'} := (m^{TI,L,\gamma'}, m^{TI,L,\gamma'})$ and $U_{\gamma'} := (m^{TI,U,\gamma'}, m^{TI,U,\gamma'})$.

By Theorem 1.3.2, it is immediate that for $\mathcal{H}(m|\gamma') = \{m \in \mathcal{M} : m_1 = m_0,\ \forall d \in \{0,1\},\ \forall s \in \mathcal{S},\ m_d(s) \geq m(s)^{L,\gamma'},\ m_d(s) \leq m_d(s)^{U,\gamma'}\}$. Hence $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m|\gamma')$. Since γ' was arbitrary, $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m|\gamma')$ for any $\gamma' \in \mathcal{H}(\gamma)$. Therefore, $L_{\gamma'}$ and $U_{\gamma'}$ are selectors of $\mathcal{H}(m|\cdot)$.

Then observe that by (A.101), $\forall m \in \mathcal{H}(m|\gamma')\ T(L_{\gamma'}, \gamma') \leq T(m, \gamma')$, so $L_{\gamma'}$ is a minimal selector with respect to T. Similarly, $\forall m \in \mathcal{H}(m|\gamma')\ T(U_{\gamma'}, \gamma') \geq T(m, \gamma')$, so $U_{\gamma'}$ is a maximal selector with respect to T. □

Proof of Corollary A.1.1. Suppose first that Assumptions RA and EV hold. Assume that $\mathcal{Y} = \mathbb{R}$ and pick an arbitrary $\tilde{c} \in \mathbb{R}$. I show that $\tilde{c} \in \mathcal{H}^O(\tau)$ which is equivalent to $\tilde{c} \in \mathcal{H}(\tau)$ by Proposition 1.4.1.

Define a distribution function for any $(a, d) \in \mathbb{R} \times \{0, 1\}$:

$$\gamma_{a|d}(B) = P_O(Y \in B, D = d) + \mathbb{1}[a \in B] P_O(D \neq d)$$

for any Borel set $B \in \mathcal{B}(\mathcal{Y})$. Recall from Lemma A.2.5 that:

$$\mathcal{H}^O(P_{Y(d)}) = \{\gamma \in \mathcal{P}^{\mathcal{Y}} : \gamma(B) \geq P_O(Y \in B, D = d)\ \forall B \in \mathcal{C}(\mathcal{Y})\}.$$

Since $\mathcal{C}(\mathcal{Y}) \subseteq \mathcal{B}(\mathcal{Y})$, then $\gamma_{a|d}(B) \in \mathcal{H}^O(P_{Y(d)})$ for any $(a, d) \in \mathbb{R} \times \{0, 1\}$. Note also that any coupling of $\gamma_{a|1} \in \mathcal{H}^O(P_{Y(1)})$ and $\gamma_{a'|0} \in \mathcal{H}^O(P_{Y(0)})$ is compatible with the observed data.

Next, observe that $\gamma_{a|d}$ is a pushforward measure of the random variable $Y\mathbb{1}[D = d] + a\mathbb{1}[D \neq d]$, which has the expectation of $E[Y\mathbb{1}[D = d]] + aP_O(D \neq d)$. Let $c = \frac{\tilde{c} - E_O[YD] + E_O[Y(1-D)]}{P_O(D \neq d)} \in \mathbb{R}$. Then $\gamma_{c|1}$ yields the expected value:

$$E[YD] + \frac{\tilde{c} - E_O[YD] + E_O[Y(1-D)]}{P_O(D=0)} P_O(D=0) = \tilde{c} + E_O[Y(1-D)].$$

Similarly, $\gamma_{0|0}$ yields the expected value $E[Y(1-D)]$.

Now take $\gamma_{c|1} \in \mathcal{H}^O(P_{Y(1)})$ and $\gamma_{0|0} \in \mathcal{H}^O(P_{Y(0)})$ as distribution functions of $Y(1)$ and $Y(0)$, recalling that any coupling of $\gamma_{c|1}$ and $\gamma_{c|0}$ is compatible with the observed data. It follows that $\tau = E[Y(1) - Y(0)] = \tilde{c}$. Since $\tilde{c}$ was arbitrary, $\mathcal{H}^O(\tau) = \mathbb{R}$. By Proposition 1.4.1, $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$.

Next, let $\mathcal{Y} = [0,1]$. Since Proposition 1.4.1 holds for any $\mathcal{Y} \subseteq \mathbb{R}$, I can again recover $\mathcal{H}(\tau)$ by using only distributions in $\mathcal{H}^O(P_{Y(d)})$. Equivalently, I can find $\mathcal{H}(\tau)$ by utilizing only information in the observational data. Then, by elementary arguments as in Manski (1990), the bounds in (A.1) follow. □

Proof of Proposition A.1.1. I prove the claims in order.

i)

$\mathcal{Y}$ is closed by definition. Since it is bounded, it is a compact set. Then $\sup \mathcal{Y} < \infty$ and $\inf \mathcal{Y} > -\infty$. Using arguments of Manski (1990), the sharp upper bound of $\mathcal{H}^O(\tau)$ is:

$$\tau \leq E_O[Y(2D-1)] + \sup \mathcal{Y} P_O(D=1) - \inf \mathcal{Y} P_O(D=0) = \sup \mathcal{H}^O(\tau). \tag{A.104}$$

By Lemma A.2.7 $V_O[Y|S, D=d] > 0$ $P-$a.s. implies $E_O[Y|S, D=d] < \sup \mathcal{Y}$ $P-$a.s. If there exists $d \in \{0,1\}$ s.t. $V[Y|S, D=d] > 0$ $P-$a.s., then it must be that for every Borel subset $B \subseteq \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D=d) > 0$ we have $E_O[Y|S \in B, D=d] < \sup \mathcal{Y}$.

Under Assumption LUC then:

$$\begin{aligned} E_O[Y(d)|D \neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d]P_O(D \neq d) \\ &= E_O[E_O[Y(d)|S(d), D = d|D \neq d]]P_O(D \neq d) \\ &= E_O[E_O[Y|S, D = d]|D \neq d]P_O(D \neq d) \\ &< \sup \mathcal{Y} P_O(D \neq d) \end{aligned} \tag{A.105}$$

where the first line is by LIE, second by Assumption LUC, third by definition, and the fourth since $E[Y|S \in B, D = d] < \sup \mathcal{Y}$ for every Borel set B of positive measure. Then under LUC:

$$\begin{aligned} E[Y(d)] &= E_O[Y\mathbb{1}[D = d]] + E[Y(d)|D \neq d]P_O(D \neq d) \\ &< E_O[Y\mathbb{1}[D = d]] + \sup \mathcal{Y} P_O(D \neq d). \end{aligned} \tag{A.106}$$

Therefore, under Assumption LUC:

$$\begin{aligned} \tau &= E[Y(1) - Y(0)] \\ &= E_O[YD] + E[Y(1)|D = 0]P_O(D = 0) - E[Y(1 - D)] - E[Y(0)|D = 1]P_O(D = 1) \\ &< E_O[Y(2D - 1)] + \sup \mathcal{Y} P_O(D = 1) - \inf \mathcal{Y} P_O(D = 0) = \sup \mathcal{H}^O(\tau) \end{aligned}$$

where the inequality follows by (A.106). Thus $\sup \mathcal{H}^{O/LUC}(\tau) < \sup \mathcal{H}^O(\tau)$. So there must exist a point in $\mathcal{H}(\tau)$ which is not contained in $\mathcal{H}^{O/LUC}(\tau)$. Conclude that $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.

ii)

Suppose that for every $d \in \{0,1\}$ $E_O[Y|S, D = d]$ is a trivial measurable function. Hence there exists a $y \in \mathcal{Y}$ such that $E_O[Y|S, D] = y$ $P-$a.s.

Then, following the same steps as in (A.105):

$$\begin{aligned} E_O[Y(d)|\neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d]P_O(D \neq d) \\ &= E_O[E_O[Y(d)|S(d), D = d|D \neq d]]P_O(D \neq d) \\ &= E_O[E_O[Y|S, D = d]|D \neq d]P_O(D \neq d) \\ &= yP_O(D \neq d) \end{aligned} \tag{A.107}$$

where the final line follows since $E_O[Y|S, D] = y$ $P-$a.s. and $Supp(\mathcal{S}(d)) = \mathcal{S}$. Given that y is identified by the data, then $E_O[Y(d)]$ is identified for every $d \in \{0,1\}$, so τ is too. It is also immediate that $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$ since for every $d \in \{0,1\}$ and any $\gamma_d \in \mathcal{P}^{\mathcal{S}}$, we have that $E[Y(d)] = \int_{\mathcal{S}} y d\gamma_d(s) = y$. Since experimental data only affect the feasible γ_d, the result follows.

□

APPENDIX B

Appendix to Chapter 2

Section B.1. Bounding Predictive Values

Positive predictive value (PPV) is the probability that a patient is diseased conditional on receiving a positive test result. Negative predictive value (NPV) is the probability that a patient who has tested negative is truly healthy. Clinicians are usually more concerned with knowing predictive values of a test t than its sensitivity and specificity.

The probability of the patient being diseased prior to observing a test result is referred to as a pre-test probability. For a known pre-test probability, sensitivity and specificity are often extrapolated from test performance studies to find predictive values using Bayes' theorem. With this in mind, let $Q(t, y)$ be the distribution of the clinical population of interest and suppose (θ_1, θ_0) extrapolate to this population in the sense of Assumption 2.3.1. As in Section 2.3, we use Q to emphasize that test performance is extrapolated and that $Q(y = 1)$ may differ from the prevalence in the performance study population. Clinicians settle on a pre-test probability $\pi = Q(y = 1)$ using the knowledge of local rates of infection and patients' symptoms and characteristics. (Watson, Whiting, and Brush (2020b))

Manski (2020) provides bounds on predictive values for COVID-19 antibody tests using point identified values of θ_1 and θ_0, when the pre-test probability π is bounded. The author notes that the analysis can be generalized to take bounds rather than exact values

of θ_1 and θ_0 as inputs. Ziegler (2021) extends the analysis of predictive values when θ_1 and θ_0 are partially identified due to an imperfect reference test, assuming that $s_0 = 1$. The bounds below do not require that $s_0 = 1$ in the performance study.

Predictive values are defined as:

$$\begin{aligned} PPV &= Q(y = 1|t = 1) = \frac{\theta_1 \pi}{\theta_1 \pi + (1 - \theta_0)(1 - \pi)} \\ NPV &= Q(y = 0|t = 0) = \frac{\theta_0 (1 - \pi)}{\theta_0 (1 - \pi) + (1 - \theta_1)\pi}. \end{aligned} \tag{B.1}$$

Assume that the sharp identification region $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ for (θ_1, θ_0) and the pre-test probability of the clinician π are known. From (B.1), it can be seen that both PPV and NPV increase with θ_1 and θ_0. Thus, the sharp bounds are:

$$\begin{aligned} PPV &\in \left[\frac{\theta_1^L \pi}{\theta_1^L \pi + (1 - \theta_0^L)(1 - \pi)}, \frac{\theta_1^U \pi}{\theta_1^U \pi + (1 - \theta_0^U)(1 - \pi)} \right] \\ NPV &\in \left[\frac{\theta_0^L (1 - \pi)}{\theta_0^L (1 - \pi) + (1 - \theta_1^L)\pi}, \frac{\theta_0^U (1 - \pi)}{\theta_0^U (1 - \pi) + (1 - \theta_1^U)\pi} \right]. \end{aligned} \tag{B.2}$$

If the clinician is not willing to settle on a single value of π, rather on a range of values $\pi \in [\pi_L, \pi_H]$, the bounds are simply:

$$\begin{aligned} PPV &\in \left[\frac{\theta_1^L \pi_L}{\theta_1^L \pi_L + (1 - \theta_0^L)(1 - \pi_L)}, \frac{\theta_1^U \pi_H}{\theta_1^U \pi_H + (1 - \theta_0^U)(1 - \pi_H)} \right] \\ NPV &\in \left[\frac{\theta_0^L \pi_H}{\theta_0^L \pi_H + (1 - \theta_1^L)(1 - \pi_H)}, \frac{\theta_0^U \pi_L}{\theta_0^U \pi_L + (1 - \theta_1^U)(1 - \pi_L)} \right]. \end{aligned} \tag{B.3}$$

The bounds are generalizable analogously to the previously outlined case for bounding prevalence when the identification region $\mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$ is expanded to $\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})$:

$$
\begin{aligned}
PPV \in \Bigg[& \min_{(\theta_1,\theta_0)\in\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})} \left\{ \frac{\theta_1\pi_L}{\theta_1\pi_L + (1-\theta_0)(1-\pi_L)} \right\}, \\
& \max_{(\theta_1,\theta_0)\in\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})} \left\{ \frac{\theta_1\pi_H}{\theta_1\pi_H + (1-\theta_0)(1-\pi_H)} \right\} \Bigg] \\
NPV \in \Bigg[& \min_{(\theta_1,\theta_0)\in\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})} \left\{ \frac{\theta_0\pi_H}{\theta_0\pi_H + (1-\theta_1)(1-\pi_H)} \right\}, \\
& \max_{(\theta_1,\theta_0)\in\mathcal{G}_{(\theta_1,\theta_0)}(\mathcal{S})} \left\{ \frac{\theta_0\pi_L}{\theta_0\pi_L + (1-\theta_1)(1-\pi_L)} \right\} \Bigg].
\end{aligned} \tag{B.4}
$$

Section B.2. Additional Moment Functions

This section defines moment functions for remaining identified sets in Propositions 2.2.1 and 2.2.2 when the tests have a tendency to wrongly agree only for $y = 0$, and for both $y = 1$ and $y = 0$. All proofs are collected in Appendix B.5.

Focus first on the bounds on θ_1 from Proposition 2.2.1. Following the reasoning in Section 2.4, we decompose the bounds on θ_1 to construct the appropriate moment inequalities. Note that there are four values determined by the population parameters that are all lower bounds, and four values that are all upper bounds on θ_1 given (s_1, s_0). One lower and one upper bound are trivial since they state that $\theta_1 \geq 0$ and $\theta_1 \leq 1$. Both can be omitted since $\theta_1 \in [0, 1]$ by definition. We can then represent the bound on θ_1 using six moment inequalities, corresponding to the six non-trivial boundary values of the identified set. One additional moment equality is needed to represent the joint identification region.

Proposition B.2.1. *Let the moment function m be:*

$$m(W_i,\theta) = \begin{pmatrix} m_1(W_i,\theta) \\ m_2(W_i,\theta) \\ m_3(W_i,\theta) \\ m_4(W_i,\theta) \\ m_5(W_i,\theta) \\ m_6(W_i,\theta) \\ m_7(W_i,\theta) \end{pmatrix} = \begin{pmatrix} (-\theta_1+s_1)\frac{r_i-1+s_0}{s_1-1+s_0} + (t_i-1)r_i \\ (-\theta_1+1-s_1)\frac{r_i-1+s_0}{s_1-1+s_0} + (r_i-1)(1-t_i) \\ (-\theta_1+1)\frac{r_i-1+s_0}{s_1-1+s_0} + (t_i-1) \\ \theta_1\frac{r_i-1+s_0}{s_1-1+s_0} - t_i \\ (\theta_1-s_1)\frac{r_i-1+s_0}{s_1-1+s_0} - t_i(1-r_i) \\ (\theta_1-1+s_1)\frac{r_i-1+s_0}{s_1-1+s_0} - t_i r_i \\ (\theta_0-1)(1-\frac{r_i-1+s_0}{s_1-1+s_0}) - \theta_1\frac{r_i-1+s_0}{s_1-1+s_0} + t_i \end{pmatrix}. \tag{B.5}$$

Identification region $\boldsymbol{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \Big(\mathcal{H}_{(\theta_1,\theta_0)}(s_1,s_0) \times \{(s_1,s_0)\}\Big)$ *with* $\mathcal{H}_{(\theta_1,\theta_0)}(s_1,s_0)$ *defined in Proposition 2.2.1 is represented by the moment function* m*. For each* $\theta \in [0,1]^2 \times \mathcal{S}$ *such that* $E_P\big(m_j(W_i,\theta)\big) \leq 0$ *for* $j = 1,\ldots,6$ *and* $E_P\big(m_7(W_i,\theta)\big) = 0$*, it must be that* $\theta \in \boldsymbol{\Theta}(P)$*. Conversely, if* $\theta \in \boldsymbol{\Theta}(P)$*, then* $E_P\big(m_j(W_i,\theta)\big) \leq 0$ *for* $j = 1,\ldots,6$ *and* $E_P\big(m_7(W_i,\theta)\big) = 0$*.*

The same reasoning applies for other bounds. Assume that the index and reference tests have a tendency to wrongly agree only for $y = 0$. As in the case when the tests have a tendency to wrongly agree only for $y = 1$, the three non-trivial lower-bound values are identical to the ones when there is no tendency to wrongly agree for any y. There are four cases for the upper bound, one of which is:

$$\theta_0 \leq \left(\frac{P_{s_1,s_0}(r=1,y=0)}{2} + P_{s_1,s_0}(r=0,y=0)\right)\frac{1}{P_{s_1,s_0}(y=0)} = \frac{1+s_0}{2} \tag{B.6}$$

Again, this is a restriction on the parameter space, since it states only that $\theta_0 \in [0, \frac{1+s_0}{2}]$. The relevant parameter space for θ when the two tests have a tendency to wrongly agree

for $y = 0$ is $\theta \in \bigcup_{(s_1,s_0)\in\mathcal{S}}[0,1] \times [0, \frac{1+s_0}{2}] \times \{(s_1, s_0)\}$. The restriction allows $\theta_0 > s_0$, but not by more than $\frac{1-s_0}{2}$.

Remark B.2.1. If the index and reference tests have a tendency to wrongly agree only for $y = 0$, then the function $\bar{m}^0$ defining moment inequalities that represent the corresponding identified set for $\theta \in \bigcup_{(s_1,s_0)\in\mathcal{S}}[0,1] \times [0, \frac{1+s_0}{2}] \times \{(s_1, s_0)\}$ would be:

$$\text{(B.7)} \quad \bar{m}^0(W_i,\theta) = \begin{pmatrix} \bar{m}_1^0(W_i,\theta) \\ \bar{m}_2^0(W_i,\theta) \\ \bar{m}_3^0(W_i,\theta) \\ \bar{m}_4^0(W_i,\theta) \\ \bar{m}_5^0(W_i,\theta) \\ \bar{m}_6^0(W_i,\theta) \\ \bar{m}_7^0(W_i,\theta) \end{pmatrix} = \begin{pmatrix} (-\theta_0 + s_0)\left(1 - \frac{r_i-1+s_0}{s_1-1+s_0}\right) + (r_i - 1)t_i \\ (-\theta_0 + 1 - s_0)\left(1 - \frac{r_i-1+s_0}{s_1-1+s_0}\right) - t_i r_i \\ (-\theta_0 + 1)\left(1 - \frac{r_i-1+s_0}{s_1-1+s_0}\right) - t_i \\ \theta_0\left(1 - \frac{r_i-1+s_0}{s_1-1+s_0}\right) + (t_i - 1) \\ (\theta_0 - s_0)\left(1 - \frac{r_i-1+s_0}{s_1-1+s_0}\right) - r_i(1 - t_i) \\ \left(\theta_0 + \frac{-1+s_0}{2}\right)\left(1 - \frac{r_i-1+s_0}{s_1-1+s_0}\right) - (1 - t_i)(1 - r_i) \\ (\theta_0 - 1)(1 - \frac{r_i-1+s_0}{s_1-1+s_0}) - \theta_1 \frac{r_i-1+s_0}{s_1-1+s_0} + t_i \end{pmatrix}.$$

The proof is analogous to that of Proposition 2.4.1.

Finally, the same steps yield a moment function that defines the identified set when the tests have a tendency to wrongly agree for both $y = 1$ and $y = 0$. As in the case where the tendency exists only for $y = 1$, the appropriate parameter space is $\theta \in \bigcup_{(s_1,s_0)\in\mathcal{S}}[0, \frac{1+s_1}{2}] \times [0, \frac{1+s_0}{2}] \times \{(s_1, s_0)\}$.

Proposition B.2.2. *Assume that the index and reference tests have a tendency to wrongly agree for $y = 1$ and $y = 0$. Let the moment function $\bar{\bar{m}}$ be equal to $\bar{m}^1$ in (2.18)*

in all components except $\bar{\bar{m}}_4(W_i,\theta)$*, and* $\bar{\bar{m}}_6(W_i,\theta)$*:*

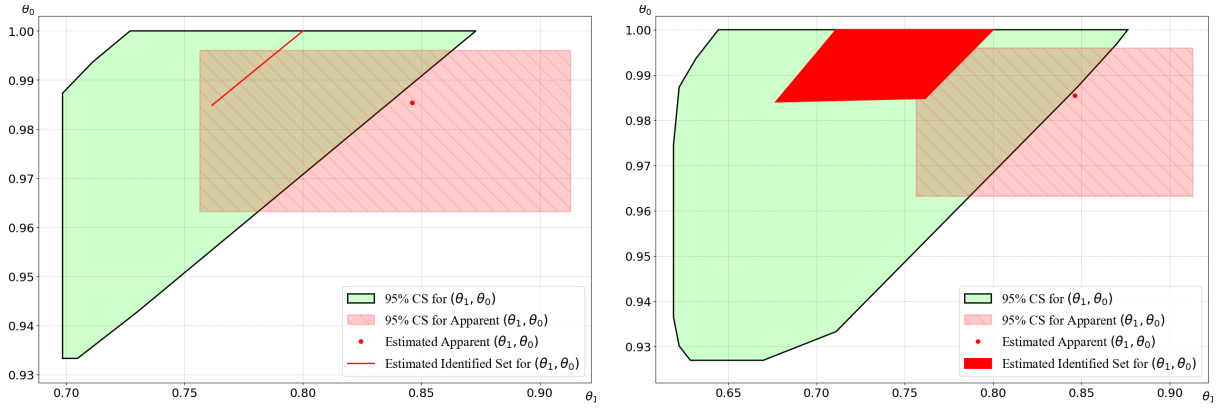
$$
\begin{aligned}
\bar{\bar{m}}_4(W_i,\theta) &= \theta_1\frac{r_i-1+s_0}{s_1-1+s_0} - t_i + \frac{1}{2}\Big(r_i - s_1\frac{r_i-1+s_0}{s_1-1+s_0}\Big) \\
\bar{\bar{m}}_6(W_i,\theta) &= \Big(\theta_1 + \frac{-1+s_1}{2}\Big)\frac{r_i-1+s_0}{s_1-1+s_0} - t_i r_i + \frac{1}{2}\Big(r_i - s_1\frac{r_i-1+s_0}{s_1-1+s_0}\Big)
\end{aligned} \tag{B.8}
$$

Joint identified set $\boldsymbol{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}}\Big(\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1,s_0)\times\{(s_1,s_0)\}\Big)$ *for* $\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1,s_0)$ *defined in Proposition 2.2.2 is represented by the moment function* $\bar{\bar{m}}$*. For each* $\theta \in \bigcup_{(s_1,s_0)\in\mathcal{S}}[0,\frac{1+s_1}{2}]\times[0,\frac{1+s_0}{2}]\times\{(s_1,s_0)\}$ *such that* $E_P\big(\bar{\bar{m}}_j(W_i,\theta)\big)\leq 0$ *for* $j=1,\ldots,6$ *and* $E_P\big(\bar{\bar{m}}_7(W_i,\theta)\big)=0$*, it must be that* $\theta\in\boldsymbol{\Theta}(P)$*. Conversely, if* $\theta\in\boldsymbol{\Theta}(P)$*, then* $E_P\big(\bar{\bar{m}}_j(W_i,\theta)\big)\leq 0$ *for* $j=1,\ldots,6$ *and* $E_P\big(\bar{\bar{m}}_7(W_i,\theta)\big)=0$*.*

Section B.3. Sensitivity Analysis

The majority of estimates obtained from the 34 data sets used by Arevalo-Rodriguez et al. (2020) indicate that s_1 may be even lower than 90%. To explore the implications of that possibility, I perform a sensitivity analysis. For exposition purposes, I assume $s_1 \in [0.8, 0.9]$, so $\mathcal{S} = [0.8, 0.9]\times\{1\}$. Values $s_1 < 0.8$ yield the same conclusion. Estimates of the identified set as well as the corresponding 95% confidence sets are found using data from each of the three samples, and presented together with findings from Section 2.5 in Figures B.1, B.2 and B.3 to facilitate comparison. Panel (b) of each Figure depicts the results under the alternative assumption. The solid red region represents the estimated identified set for (θ_1,θ_0). It is no longer a line. In all figures both the confidence and the estimated identified set become larger, but remain informative. Tables B.1 and B.2 shows estimates of the projected bounds. Bounds for specificity are unchanged, while those for sensitivity expand only downwards. Assumed values $s_1 < 0.8$ accentuate this

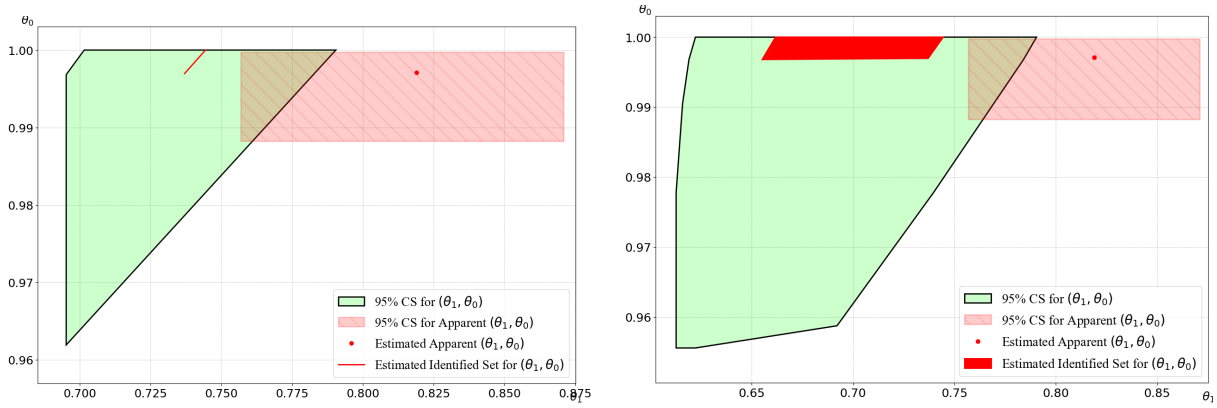
effect. The tendency of "apparent" sensitivity to overestimate true sensitivity increases as s_1 is reduced. On the other hand, allowing for $s_1 > 0.9$ enlarges the estimated upper bounds on sensitivity, but for values of $s_1 < 1$ it never surpasses "apparent" sensitivity. Hence, the finding that "apparent" sensitivity overestimates true sensitivity is robust to different assumed values of s_1.



(a) $s_1 = 0.9$ (b) $s_1 \in [0.8, 0.9]$

Figure B.1. Estimates, and 95% confidence sets for "apparent" measures and points in the identified set for (θ_1, θ_0) in the EUA study.

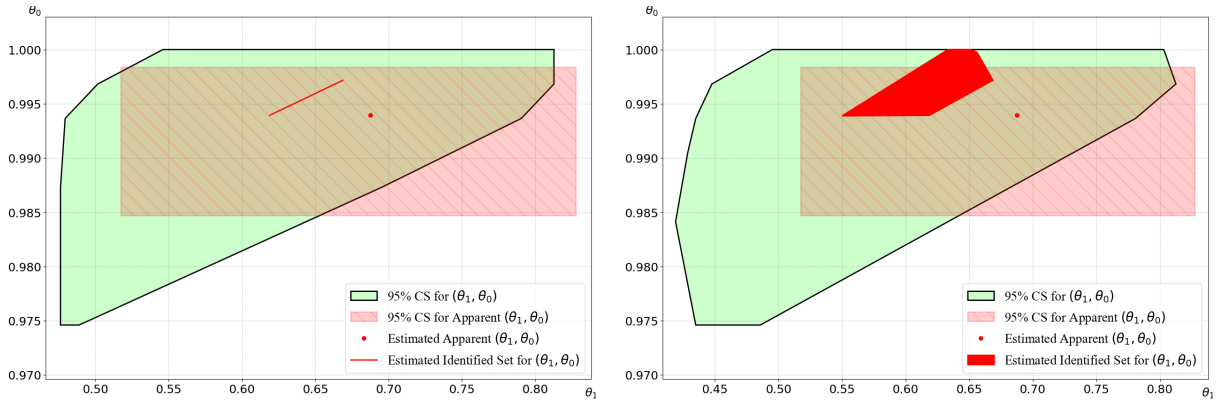
In panel (a) $\mathcal{S} = \{(0.9, 1)\}$, and $\mathcal{S} = [0.8, 0.9] \times \{1\}$ in panel (b).



(a) $s_1 = 0.9$ (b) $s_1 \in [0.8, 0.9]$

Figure B.2. Estimates, and 95% confidence sets for "apparent" measures and points in the identified set for (θ_1, θ_0) in the symptomatic population of Shah et al. (2021).

In panel (a) $\mathcal{S} = \{(0.9, 1)\}$, and $\mathcal{S} = [0.8, 0.9] \times \{1\}$ in panel (b).



(a) $s_1 = 0.9$ (b) $s_1 \in [0.8, 0.9]$

Figure B.3. Estimates, and 95% confidence sets for "apparent" measures and points in the identified set for (θ_1, θ_0) in the asymptomatic population of Shah et al. (2021).

In panel (a) $\mathcal{S} = \{(0.9, 1)\}$, and $\mathcal{S} = [0.8, 0.9] \times \{1\}$ in panel (b).

Table B.1. θ_1 Sensitivity Analysis Estimates.

Data	Appar.	$s_1 = 0.9$	$s_1 \in [0.8, 0.9]$
EUA Sx	0.846	$[0.761, 0.800]$	$[0.677, 0.800]$
Shah et al. Sx	0.819	$[0.737, 0.744]$	$[0.655, 0.744]$
Shah et al. ASx	0.688	$[0.619, 0.669]$	$[0.550, 0.669]$

Table B.2. θ_0 Sensitivity Analysis Estimates.

Data	Appar.	$s_1 = 0.9$	$s_1 \in [0.8, 0.9]$
EUA Sx	0.985	$[0.985, 1.000]$	$[0.984, 1.000]$
Shah et al. Sx	0.997	$[0.997, 1.000]$	$[0.997, 1.000]$
Shah et al. ASx	0.994	$[0.994, 0.997]$	$[0.994, 0.997]$

Note: Apparent estimated values and estimated projected bounds for (θ_1, θ_0) for different $\mathcal{S}$. Sx denotes the symptomatic, and ASx the asymptomatic individuals.

Section B.4. Auxiliary Results

Lemma B.4.1. *For a fixed* (s_1, s_0) *and any* $(j, k, l) \in \{0, 1\}^3$ *it holds that:*

$$P(t = j, r = k) - P_{s_1,s_0}(r = k, y = 1-l) = P_{s_1,s_0}(r = k, y = l) - P(t = 1-j, r = k). \tag{B.9}$$

Proof. Suppressing the subscript in P_{s_1,s_0} for clarity:

$$
\begin{aligned}
&P(t=j,r=k) - P(r=k,y=1-l) = \\
&= P(t=j,r=k,y=l) + P(t=j,r=k,y=1-l) \\
&- P(t=j,r=k,y=1-l) - P(t=1-j,r=k,y=1-l) \\
&= P(t=j,r=k,y=l) - P(t=1-j,r=k,y=1-l) \\
&= P(t=j,r=k,y=l) + P(t=1-j,r=k,y=l) \\
&- P(t=1-j,r=k,y=l) - P(t=1-j,r=k,y=1-l) \\
&= P(r=k,y=l) - P(t=1-j,r=k).
\end{aligned} \tag{B.10}
$$

□

Lemma B.4.2. *Let $P(t,r)$ and (s_1,s_0) be known, and $\mathcal{H}_{\theta_j}(s_1,s_0) = [\theta_j^L, \theta_j^U]$ as in (2.7). Define:*

$$
\begin{aligned}
\hat{\mathcal{H}}_{\theta_j}(s_1,s_0) = &\Big[max\Big(0, P(t=j) - P_{s_1,s_0}(y=1-j)\Big), \\
&min\Big(P(t=j), P_{s_1,s_0}(y=j)\Big)\Big] \frac{1}{P_{s_1,s_0}(y=j)}
\end{aligned} \tag{B.11}
$$

Then $\mathcal{H}_{\theta_j}(s_1,s_0) \subseteq \hat{\mathcal{H}}_{\theta_j}(s_1,s_0)$.

Proof. By Lemma B.4.1, the lower bound in (B.11) is equivalent to $max\Big(0, P(t=j,r=j) - P_{s_1,s_0}(r=j,y=1-j) + P_{s_1,s_0}(r=1-j,y=j) - P(t=1-j,r=1-j)\Big)\frac{1}{P_{s_1,s_0}(y=j)} \leq \theta_j^L$ since the maximum of a sum of functions is at most the sum of individual maxima. Similarly, the upper bound is $min\Big(P(t=j,r=j) + P(t=j,r=1-j), P_{s_1,s_0}(r=j,y=j) + P_{s_1,s_0}(r=1-j,y=j)\Big)\frac{1}{P_{s_1,s_0}(y=j)} \geq \theta_j^U$. □

Lemma B.4.3. *For any $P(t,r)$ and (s_1,s_0) such that $\theta_j^L > s_j$ it must be that $\theta_{1-j}^L < s_{1-j}$.*

Proof. We prove the claim that $\theta_1^L > s_1$ implies $\theta_0^L < s_0$. Symmetrically, one can show that $\theta_0^L > s_0$ implies $\theta_1^L < s_1$.

Suppose that $\theta_1^L > s_1$. This is equivalent to $\theta_1^L P_{s_1,s_0}(y=1) > P_{s_1,s_0}(r=1,y=1)$ by Assumption 2.2.3. Taking any lower bound θ_j^L from Propositions 2.2.1 and 2.2.2 yields:

$$\begin{aligned}
\theta_1^L P_{s_1,s_0}(y=1) &= max\Big(0, P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0)\Big) \\
&\quad + max\Big(0, P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0)\Big) \\
\theta_0^L P_{s_1,s_0}(y=0) &= max\Big(0, P(t=0,r=0) - P_{s_1,s_0}(r=0,y=1)\Big) \\
&\quad + max\Big(0, P_{s_1,s_0}(r=1,y=0) - P(t=1,r=1)\Big).
\end{aligned} \tag{B.12}$$

Since $\theta_1^L > s_1$, it must also be that $\theta_1^L > 0$ by Assumption 2. By (B.12) then $P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0) > 0$ or $P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0) > 0$.

We show that $P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0) > 0$ must hold when $\theta_1^L > s_1$. By way of contradiction suppose that $P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0) \leq 0$. Since $P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0) > 0$ or $P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0) > 0$, it must be $P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0) > 0$. Then:

$$\begin{aligned}
\theta_1 P_{s_1,s_0}(y=1) = P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0) > P_{s_1,s_0}(r=1,y=1) \\
\iff P(t=1,r=1) > P(r=1)
\end{aligned}$$

which is a contradiction, so $P_{s_1,s_0}(r=0,y=1)-P(t=0,r=0)>0$. To complete the proof, we consider two cases.

Suppose first that $P(t=1,r=1)-P_{s_1,s_0}(r=1,y=0)>0$ and $P_{s_1,s_0}(r=0,y=1)-P(t=0,r=0)>0$. Then, it is immediate from (B.12) that $\theta_0^L=0$. That $\theta_0^L<s_0$ is then direct from Assumption 2.2.2 since $s_1+s_0>1$ and $(s_1,s_0)\in[0,1]^2$.

Finally, suppose that $P(t=1,r=1)-P_{s_1,s_0}(r=1,y=0)\leq 0$ and $P_{s_1,s_0}(r=0,y=1)-P(t=0,r=0)>0$. By way of contradiction, suppose that $\theta_0^L\geq s_0$ From (B.12):

$$\begin{aligned}\theta_0^L P_{s_1,s_0}(y=0)=P_{s_1,s_0}(r=1,y=0)-P(t=1,r=1)\geq P_{s_1,s_0}(r=0,y=0)\\ \iff P(t=0,r=1)\geq P_{s_1,s_0}(r=1,y=1)+P_{s_1,s_0}(r=0,y=0)\end{aligned}\tag{B.13}$$

where the second line follows by Lemma B.4.1. Similarly:

$$\begin{aligned}\theta_1^L P_{s_1,s_0}(y=1)=P_{s_1,s_0}(r=0,y=1)-P(t=0,r=0)>P_{s_1,s_0}(r=1,y=1)\\ \iff P(t=1,r=0)>P_{s_1,s_0}(r=1,y=1)+P_{s_1,s_0}(r=0,y=0).\end{aligned}\tag{B.14}$$

Now using $P(r=j)\geq P(t=1-j,r=j)$, (B.13) and (B.14):

$$\begin{aligned}P(r=0)>&s_1P_{s_1,s_0}(y=1)+s_0P_{s_1,s_0}(y=0)\\ P(r=1)\geq&s_1P_{s_1,s_0}(y=1)+s_0P_{s_1,s_0}(y=0)\end{aligned}\tag{B.15}$$

Recall that Assumptions 2.2.2 and 2.2.3 imply that $P(r=1)\in(1-s_0,s_1)$, so $P(r=0)\in(1-s_1,s_0)$. This together with (B.15) implies $s_j>s_1P_{s_1,s_0}(y=1)+s_0P_{s_1,s_0}(y=0)$ for $j=0,1$ which is equivalent to $s_1>s_0$ and $s_0>s_1$ by Assumption 2.2.3. This yields a contradiction, showing that $\theta_0^L<s_0$. □

Lemma B.4.4. *Suppose that t and r have a tendency to wrongly agree for some $y = j$. Then:*

- $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0) \subset \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0) \Leftrightarrow P(t = j, r = 1 - j) > \frac{P_{s_1,s_0}(r=1-j,y=j)}{2} > 0$;
- $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0) \subset \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ *implies* $s_j < 1$.

Proof. Focus on the case where the tests have the tendency to wrongly agree only for $y = 1$.

From (2.6), $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0) \subset \mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ is equivalent to $\bar{\mathcal{H}}_{\theta_1}(s_1, s_0) \subset \mathcal{H}_{\theta_1}(s_1, s_0)$, which are intervals. Next, since lower bounds of the two intervals are θ_1^L, this is further equivalent to $\bar{\theta}_1^U < \theta_1^U$. First, assume $P(t = 1, r = 0) > \frac{P_{s_1,s_0}(r=0,y=1)}{2} > 0$. Then:

$$\begin{aligned}
\theta_1^U - \bar{\theta}_1^U &= min\left(P(t = 1, r = 0), P_{s_1,s_0}(r = 0, y = 1)\right) \\
&\quad - min\left(P(t = 1, r = 0), \frac{P_{s_1,s_0}(r = 0, y = 1)}{2}\right) \\
&= min\left(P(t = 1, r = 0) - \frac{P_{s_1,s_0}(r = 0, y = 1)}{2}, \frac{P_{s_1,s_0}(r = 0, y = 1)}{2}\right) > 0.
\end{aligned}$$

Next suppose $\bar{\theta}_1^U < \theta_1^U$. This is equivalent to:

$$min\left(P(t = 1, r = 0), P_{s_1,s_0}(r = 0, y = 1)\right) > min\left(P(t = 1, r = 0), \frac{P_{s_1,s_0}(r = 0, y = 1)}{2}\right).$$

By way of contradiction suppose that $\frac{P_{s_1,s_0}(r=0,y=1)}{2} = 0$ or $P(t = 1, r = 0) \leq \frac{P_{s_1,s_0}(r=0,y=1)}{2}$. If $\frac{P_{s_1,s_0}(r=0,y=1)}{2} = 0$ then $\bar{\theta}_1^U = \theta_1^U$. If $P(t = 1, r = 0) \leq \frac{P_{s_1,s_0}(r=0,y=1)}{2}$, then again $\bar{\theta}_1^U = \theta_1^U$ proving the first claim for $y = 1$. The argument for $y = 0$ is symmetric.

For the second claim, observe that $P(t = j, r = 1 - j) > \frac{P_{s_1,s_0}(r=1-j,y=j)}{2} > 0$ implies $P_{s_1,s_0}(r = 1 - j, y = j) = (1 - s_j)P_{s_1,s_0}(y = j) > 0$ or that $s_j < 1$ by Assumption 2.2.3. □

Section B.5. Proofs

PROOF OF PROPOSITION 2.2.1. Alternative proofs can be constructed using Theorem 3.10 from Joe (1997) or Artstein's inequalities (Beresteanu, Molchanov, and Molinari (2012)).[1] Here, I offer a direct proof that follows through a series of claims. Intermediate results will be used to prove other propositions. First we derive bounds on $P_{s_1,s_0}(t = j, r = k, y = l)$ for $(j, k, l) \in \{0, 1\}^3$. We then show that the pair of bounds on $P_{s_1,s_0}(t = j, r = j, y = j)$ and $P_{s_1,s_0}(t = j, r = 1 - j, y = j)$ for a fixed j are sharp and that any two points in these bounds are attainable simultaneously. The sharp bound on θ_j follows by summing the individual bounds and dividing by $P_{(s_1,s_0)}(y = 1)$. Finally, the sharp joint identified set for (θ_1, θ_0) is immediate by the law of total probability.

Claim B.5.1. *Bounds on* $P_{s_1,s_0}(t = j, r = k, y = l)$ *for any* $(j, k, l) \in \{0, 1\}^3$ *are:*

$$P_{s_1,s_0}(t = j, r = k, y = l) \in \Big[max\Big(0, P_{s_1,s_0}(r = k, y = l) - P(t = 1 - j, r = k)\Big), \\ min\Big(P(t = j, r = k), P_{s_1,s_0}(r = k, y = l)\Big)\Big]. \tag{B.16}$$

Proof. Probability $P_{s_1,s_0}(t = j, r = k, y = l)$ for any $(j, k, l) \in \{0, 1\}^3$ is the probability of the intersection of events $P_{s_1,s_0}(\{t = j, r = k\} \cap \{r = k, y = l\})$. An upper bound on $P_{s_1,s_0}(t = j, r = k, y = l)$ is then:

(B.17)

$$P_{s_1,s_0}(\{t = j, r = k\} \cap \{r = k, y = l\}) \leq min\Big(P(t = j, r = k), P_{s_1,s_0}(r = k, y = l)\Big).$$

1. I am grateful to an anonymous referee and Gabriel Ziegler for bringing this to my attention.

The upper bound (B.17) holds for any $(j,k,l) \in \{0,1\}^3$. Using the upper bound on $P_{s_1,s_0}(t=j,r=k,y=1-l)$, the lower bound on $P_{s_1,s_0}(t=j,r=k,y=l)$ is:

$$
\begin{aligned}
& P_{s_1,s_0}(t=j,r=k,y=l) \\
& = P(t=j,r=k) - P_{s_1,s_0}(t=j,r=k,y=1-l) \\
& \geq P(t=j,r=k) - min\Big(P(t=j,r=k), P_{s_1,s_0}(r=k,y=1-l)\Big) \qquad \text{(B.18)} \\
& = max\Big(0, P(t=j,r=k) - P_{s_1,s_0}(r=k,y=1-l)\Big) \\
& = max\Big(0, P_{s_1,s_0}(r=k,y=l) - P(t=1-j,r=k)\Big)
\end{aligned}
$$

where the final line of (B.18) follows from Lemma B.4.1. □

Claim B.5.2. *Bounds (B.17) on $P_{s_1,s_0}(t=j,r=j,y=j)$ and $P_{s_1,s_0}(t=j,r=1-j,y=j)$ are sharp. Bounds are independent in the sense that any pair of points within the two bounds is attainable.*

Proof. Write all eight joint and observable probabilities as a matrix equation:

(B.19)

$$\underbrace{\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} P_{s_1,s_0}(t=1,r=1,y=1) \\ P_{s_1,s_0}(t=1,r=1,y=0) \\ P_{s_1,s_0}(t=0,r=1,y=1) \\ P_{s_1,s_0}(t=0,r=1,y=0) \\ P_{s_1,s_0}(t=1,r=0,y=1) \\ P_{s_1,s_0}(t=1,r=0,y=0) \\ P_{s_1,s_0}(t=0,r=0,y=1) \\ P_{s_1,s_0}(t=0,r=0,y=0) \end{pmatrix} = \begin{pmatrix} P(t=1,r=1) \\ P_{s_1,s_0}(r=1,y=1) \\ P(t=0,r=1) \\ P_{s_1,s_0}(r=1,y=0) \\ P(t=1,r=0) \\ P_{s_1,s_0}(r=0,y=1) \\ P(t=0,r=0) \\ P_{s_1,s_0}(r=0,y=0) \end{pmatrix}.$$

Matrix **A** has rank 6. The bottom four rows cannot be represented as a linear combination using any of the top four rows. The bottom four rows are only mutually linearly dependent. Similarly, the top four rows are only mutually linearly dependent. Therefore, the value of $P_{s_1,s_0}(t=j,r=1-k,y=l)$ does not affect the values of $P_{s_1,s_0}(t=j,r=k,y=l)$ for $(j,l) \in \{0,1\}^2$ within their respective bounds.

There exist two separate systems of equations, one for each value of r. Focus on one system for an arbitrary $r = k$:

(B.20)
$$\underbrace{\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}}_{\mathbf{A}'} \begin{pmatrix} P_{s_1,s_0}(t=j, r=k, y=j) \\ P_{s_1,s_0}(t=j, r=k, y=1-j) \\ P_{s_1,s_0}(t=1-j, r=k, y=j) \\ P_{s_1,s_0}(t=1-j, r=k, y=1-j) \end{pmatrix} = \begin{pmatrix} P(t=j, r=k) \\ P_{s_1,s_0}(r=k, y=j) \\ P(t=1-j, r=k) \\ P_{s_1,s_0}(r=k, y=1-j) \end{pmatrix}.$$

Matrix $\mathbf{A}'$ has rank 3. I first show that both the upper and lower bounds on any of the joint probabilities $P_{s_1,s_0}(t=j, r=k, y=l)$ in (B.20) are attainable for $(j,l) \in \{0,1\}^2$.[2] Then, I demonstrate that any value in the interior of the bounds is attainable, proving that bounds on $P_{s_1,s_0}(t=j, r=k, y=l)$ are sharp. Focus on $P_{s_1,s_0}(t=j, r=k, y=j)$. Assume that it is equal to its upper bound, $P_{s_1,s_0}(t=j, r=k, y=j) = min\Big(P(t=j, r=k), P_{s_1,s_0}(r=k, y=j)\Big)$. Let first $P(t=j, r=k) < P_{s_1,s_0}(r=k, y=j)$. From Lemma B.4.1, $P_{s_1,s_0}(r=k, y=1-j) < P(t=1-j, r=k)$. Then from (B.20):

(B.21)
$$\begin{aligned} &P_{s_1,s_0}(t=j, r=k, y=j) = P(t=j, r=k) \\ &P_{s_1,s_0}(t=j, r=k, y=1-j) = 0 \\ &P_{s_1,s_0}(t=1-j, r=k, y=j) = P_{s_1,s_0}(r=k, y=j) - P(t=j, r=k) \\ &P_{s_1,s_0}(t=1-j, r=k, y=1-j) = P_{s_1,s_0}(r=k, y=1-j). \end{aligned}$$

2. Attainable (or equivalently feasible) is meant in the sense that it is consistent with the observed distribution $P(t,r)$ and assumed values for (s_1, s_0).

By assumption, $P_{s_1,s_0}(t=j,r=k,y=j)$ is equal to its upper bound. Consequently, $P_{s_1,s_0}(t=j,r=k,y=1-j)$ is equal to $0 = max(0, P_{s_1,s_0}(r=k,y=1-j) - P(t=1-j,r=k))$ which is its lower bound. Similarly, $P_{s_1,s_0}(t=1-j,r=k,y=j) = P_{s_1,s_0}(r=k,y=j) - P(t=j,r=k) = max(0, P_{s_1,s_0}(r=k,y=j) - P(t=j,r=k))$, which is its lower bound. Finally, $P_{s_1,s_0}(t=1-j,r=k,y=1-j) = P_{s_1,s_0}(r=k,y=1-j) = min(P(t=1-j,r=k), P_{s_1,s_0}(r=k,y=1-j))$, representing the upper bound. All four probabilities achieve their corresponding upper and lower bounds.

Let now $P(t=j,r=k) \geq P_{s_1,s_0}(r=k,y=j)$, or equivalently $P_{s_1,s_0}(r=k,y=1-j) \geq P(t=1-j,r=k)$. The system then is:

$$
\begin{aligned}
&P_{s_1,s_0}(t=j,r=k,y=j) = P_{s_1,s_0}(r=k,y=j)\\
&P_{s_1,s_0}(t=j,r=k,y=1-j) = P(t=j,r=k) - P_{s_1,s_0}(r=k,y=j)\\
&P_{s_1,s_0}(t=1-j,r=k,y=j) = 0\\
&P_{s_1,s_0}(t=1-j,r=k,y=1-j) = P(t=1-j,r=k).
\end{aligned}
\tag{B.22}
$$

As before, $P_{s_1,s_0}(t=j,r=k,y=j)$ and $P_{s_1,s_0}(t=1-j,r=k,y=1-j)$ are equal to their respective upper bounds. $P_{s_1,s_0}(t=j,r=k,y=1-j)$ and $P_{s_1,s_0}(t=1-j,r=k,y=j)$ attain the lower bounds. That $P_{s_1,s_0}(t=j,r=k,y=j)$ and $P_{s_1,s_0}(t=1-j,r=k,y=1-j)$ attain lower bounds when $P_{s_1,s_0}(t=j,r=k,y=1-j)$ and $P_{s_1,s_0}(t=1-j,r=k,y=j)$ are equal to their upper bounds can be shown symmetrically. Thus, for an arbitrary $r=k$, all probabilities can be equal to their upper and lower bounds.

From (B.20), reducing any probability that is on the upper bound will lead to an increase in the probabilities at lower bounds and a decrease in the remaining probability

at the upper bound. Any value in the interior of the bounds must be feasible. Therefore, the bounds (B.16) must be sharp for $P(t=j, r=k, y=l)$ and any $(j,l) \in \{0,1\}^2$. This is true for an arbitrary $r=k$, hence the bounds are sharp for any $P(t=j, r=k, y=l)$ such that $(j,k,l) \in \{0,1\}^3$.

Finally, from (B.19), the value which $P_{s_1,s_0}(t=j, r=j, y=j)$ takes does not influence the value of $P_{s_1,s_0}(t=j, r=1-j, y=j)$. Any pair of values coming from the Cartesian product of the bounds on the two probabilities is feasible. □

By Claim B.5.2, the sharp bounds on $P_{s_1,s_0}(t=j, y=1) = P_{s_1,s_0}(t=j, r=j, y=j) + P_{s_1,s_0}(t=j, r=1-j, y=j)$ are a sum of the sharp bounds on individual probabilities. Hence, the sharp bounds on θ_j are:

$$
\begin{aligned}
\theta_j \geq & \frac{max\Big(0, P(t=j, r=j) - P_{s_1,s_0}(r=j, y=1-j)\Big)}{P_{s_1,s_0}(y=j)} \\
& + \frac{max\Big(0, P_{s_1,s_0}(r=1-j, y=j) - P(t=1-j, r=1-j)\Big)}{P_{s_1,s_0}(y=j)} \\
\theta_j \leq & \frac{min\Big(P(t=j, r=1-j), P_{s_1,s_0}(r=1-j, y=j)\Big)}{P_{s_1,s_0}(y=j)} \\
& + \frac{min\Big(P(t=j, r=j), P_{s_1,s_0}(r=j, y=j)\Big)}{P_{s_1,s_0}(y=j)}
\end{aligned}
\tag{B.23}
$$

Claim B.5.3. *The sharp joint identified set for (θ_1, θ_0) is:*

$$
\mathcal{H}_{(\theta_1,\theta_0)}(s_1,s_0) = \left\{ (t_1,t_0) : t_0 = t_1 \frac{P_{s_1,s_0}(y=1)}{P_{s_1,s_0}(y=0)} + 1 - \frac{P(t=1)}{P_{s_1,s_0}(y=0)}, t_1 \in \mathcal{H}_{\theta_j}(s_1,s_0) \right\}.
$$

Proof.

$$P(t=1) = P_{s_1,s_0}(t=1,y=1) + P_{s_1,s_0}(t=1,y=0) = \\ = \theta_1 P_{s_1,s_0}(y=1) + P_{s_1,s_0}(y=0) - \theta_0 P_{s_1,s_0}(y=0). \tag{B.24}$$

Set $j = 1$ without loss of generality. For any value $t_1 \in \mathcal{H}_{\theta_1}(s_1, s_0)$, it must be that $t_0 P_{s_1,s_0}(y=0) = t_1 P_{s_1,s_0}(y=1) + P_{s_1,s_0}(y=0) - P(t=1)$. Since $\mathcal{H}_{\theta_1}(s_1, s_0)$ is sharp, $\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0)$ is a sharp joint identification region for (θ_1, θ_0). □

□

Proof of Proposition 2.2.2. First, I prove a lemma used below. The proof then follows through a series of claims.

Lemma B.5.1. *The index test has a tendency to wrongly agree with the reference test for $y = j$ for a given (s_1, s_0), if and only if $P_{s_1,s_0}(t = 1-j, r = 1-j, y = j) \geq \frac{P_{s_1,s_0}(r=1-j,y=j)}{2}$.*

Proof. It holds that $P_{s_1,s_0}(t = 1-j, r = 1-j, y = j) + P_{s_1,s_0}(t = j, r = 1-j, y = j) = P(r = 1-j, y = j)$. For sufficiency, note that $2P_{s_1,s_0}(t = 1-j, r = 1-j, y = j) = P_{s_1,s_0}(r = 1-j, y = j) - P_{s_1,s_0}(t = j, r = 1-j, y = j) + P_{s_1,s_0}(t = 1-j, r = 1-j, y = j) \geq P_{s_1,s_0}(r = 1-j, y = j)$, since by assumption $P_{s_1,s_0}(t = 1-j, r = 1-j, y = j) \geq P_{s_1,s_0}(t = j, r = 1-j, y = j)$. Necessity is immediate. □

Claim B.5.4. *Assume that the tests have a tendency to wrongly agree only for $y = j$. The sharp identified set for (θ_1, θ_0) is $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1, s_0)$.*

Proof. From Lemma B.5.1, $P_{s_1,s_0}(t=1-j, r=1-j, y=j) \geq \frac{P_{s_1,s_0}(r=1-j,y=j)}{2}$. Then, $P_{s_1,s_0}(t=j, r=1-j, y=j) \leq \frac{P_{s_1,s_0}(r=1-j,y=j)}{2} \leq P_{s_1,s_0}(r=1-j, y=j)$. Using this and following the steps taken to obtain (B.17):

$$P_{s_1,s_0}(t=j, r=1-j, y=j) \leq min\Big(P(t=j, r=1-j), \frac{P_{s_1,s_0}(r=1-j, y=j)}{2}\Big). \tag{B.25}$$

The lower bound on $P_{s_1,s_0}(t=j, r=1-j, y=j)$ is derived from the upper bound on $P_{s_1,s_0}(t=1-j, r=1-j, y=j)$ which is unaffected by the assumption. Substituting the upper bound into the system (B.20) yields the lower bound $P_{s_1,s_0}(t=j, r=1-j, y=j) \geq \max\Big(0, P_{s_1,s_0}(r=1-j, y=j) - P(t=1-j, r=1-j)\Big)$, as in (B.18).

For the bounds defined by (B.18) and (B.25) on $P_{s_1,s_0}(t=j, r=1-j, y=j)$ to be sharp, all values contained between them must be feasible for a given population distribution. The lower bound is identical as in Proposition 2.2.1. The upper bound in (B.25) is at most as large as the upper bound (B.17) in Proposition 2.2.1. Thus, all points in the bounds on $P_{s_1,s_0}(t=j, r=1-j, y=j)$ are attainable by the same argument as in Claim B.5.2 in the proof of Proposition 2.2.1. Hence, the bounds defined by (B.18) and (B.25) are sharp. Sharp bounds on probabilities $P_{s_1,s_0}(t=k, r=j, y=l)$ from (B.16) are unaffected by the assumption for $(k,l) \in \{0,1\}^2$ as they form an independent system of equations from (B.19). Using the reasoning in Claims B.5.2, and B.5.3 of Proposition 2.2.1, $\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1,s_0)$ is a sharp identification region for (θ_1,θ_0).

□

Claim B.5.5. *Assume that the tests have a tendency to wrongly agree for $y=0$ and $y=1$. The sharp identified set for (θ_1,θ_0) is $\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1,s_0)$.*

Proof. By Lemma B.5.1, $P_{s_1,s_0}(t = 1-j, r = 1-j, y = j) \geq \frac{P_{s_1,s_0}(r=1-j,y=j)}{2}$ for $j \in \{0,1\}$. The sharp upper bound on $P_{s_1,s_0}(t = j, r = 1-j, y = j)$ is again as in (B.25). The sharp upper bound on $P_{s_1,s_0}(t = j, r = j, y = j)$ is no longer equivalent to (B.17). Analogously to the steps used to derive (B.25):

$$\begin{aligned} &P_{s_1,s_0}(t = j, r = j, y = j) \\ &\leq min\Big(P(t = j, r = j) - \frac{P_{s_1,s_0}(r = j, y = 1-j)}{2}, P_{s_1,s_0}(r = j, y = j)\Big), \end{aligned} \tag{B.26}$$

where the first value in the minimum is derived using Lemma B.5.1 and:

$$\begin{aligned} P_{s_1,s_0}(t = j, r = j, y = j) &= P(t = j, r = j) - P_{s_1,s_0}(t = j, r = j, y = 1-j) \\ &\leq P(t = j, r = j) - \frac{P_{s_1,s_0}(r = j, y = 1-j)}{2}. \end{aligned} \tag{B.27}$$

Remark B.5.1. Only the upper bounds on $P_{s_1,s_0}(t = j, r = 1-j, y = j)$ and $P_{s_1,s_0}(t = j, r = j, y = j)$ are changed by the assumption that tests have a tendency to wrongly agree for $y \in \{0,1\}$. The lower bounds remain as in (B.18).

To see this, observe from (B.19) that the bounds on $P_{s_1,s_0}(t = j, r = 1-j, y = j)$ and $P_{s_1,s_0}(t = j, r = j, y = j)$ belong to separate systems of equations and will not affect each other. The bounds on $P_{s_1,s_0}(t = j, r = 1-j, y = j)$ hold as in the Claim B.5.4. The bounds on $P_{s_1,s_0}(t = j, r = j, y = j)$ are derived using $P_{s_1,s_0}(t = j, r = j, y = 1-j)$ which is affected only from below by the assumption. From (B.20) it can be seen that substituting $P_{s_1,s_0}(t = j, r = j, y = 1-j)$ with its upper bound $\min\Big(P(t = j, r = j), P_{s_1,s_0}(r = j, y = 1-j)\Big)$ yields an identical lower bound for $P_{s_1,s_0}(t = j, r = j, y = j)$ as in (B.18).

Bounds (B.18) and (B.25) on $P_{s_1,s_0}(t = j, r = 1 - j, y = j)$ were shown to be sharp in the previous claim. Using the same argument, bounds (B.18) and (B.26) on $P_{s_1,s_0}(t = j, r = j, y = j)$ are also sharp. Any pair of points in the bounds for the two probabilities is feasible. Hence, $\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1, s_0)$ is the sharp identified set for (θ_1, θ_0).

□

□

Proof of Proposition 2.3.1. The proof proceeds in two steps. First we show that the bounds are valid. For sharpness, we consider an arbitrary point in the bounds. We then construct a distribution consistent with the assumptions that generates the point, and that is observationally equivalent to the data. The data alone identify only $Q(t = 1)$.

We study two distinct cases.

Case 1: $\forall(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : \quad \theta_1 \neq 1 - \theta_0$

The bounds on $Q(y = 1)$ can be imposed as:

$$
\begin{aligned}
&Q(y = 1) \\
&= \frac{Q(t = 1) + \tau_0 - 1}{\tau_1 + \tau_0 - 1} \\
&\in \left[\min_{(\theta_1,\theta_0)\in\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)} \frac{Q(t = 1) + \theta_0 - 1}{\theta_1 + \theta_0 - 1}, \max_{(\theta_1,\theta_0)\in\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)} \frac{Q(t = 1) + \theta_0 - 1}{\theta_1 + \theta_0 - 1} \right] \cap [0, 1]
\end{aligned}
$$

where the first line follows by (2.12), and the second by Assumption 2.3.1. The intersection with $[0, 1]$ is added by definition $Q(y = 1)$ and the fact that $\frac{Q(t=1)+\theta_0-1}{\theta_1+\theta_0-1} \notin [0, 1]$ if and only if (θ_1, θ_0) are such that $Q(t = 1) \notin [\min(\theta_1, 1 - \theta_0), \max(\theta_1, 1 - \theta_0)]$, which is possible. The expression $\frac{Q(t=1)+\theta_0-1}{\theta_1+\theta_0-1}$ is increasing in θ_0 and decreasing in θ_1. We will

show that it attains extrema at the end-points of the line segment $\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)$. By Propositions 2.2.1 and 2.2.2, for any $(\theta_1,\theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)$:

$$\theta_0 = 1 - a + \theta_1 b \tag{B.28}$$

for known constants a and $b \in (0,\infty)$. Given that $\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0) \subset [0,1]^2$ is a line segment, it is a connected set. By assumption it does not contain (θ_1,θ_0) such that $\theta_1 + \theta_0 = 1$, therefore it does not intersect the negatively-sloped diagonal of the unit rectangle. Thus, all $(\theta_1,\theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)$ are such that either $\theta_1 + \theta_0 > 1$ or $\theta_1 + \theta_0 < 1$. By (B.28), for all θ_1 in the identified set we have either $\theta_1(b+1) > a$ or $\theta_1(b+1) < a$, so $\theta_1(b+1) \neq a$. For any $(\theta_1,\theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)$ we can then write:

$$\frac{Q(t=1) + \theta_0 - 1}{\theta_1 + \theta_0 - 1} = \frac{Q(t=1) + \theta_1 b - a}{\theta_1(b+1) - a}. \tag{B.29}$$

First derivative of (B.29) with respect to θ_1 is $\frac{a-(b+1)Q(t=1)}{(a-(b+1)\theta_1)^2}$ which has the same sign for all θ_1 in the identified set. If (and only if) $a = (b+1)Q(t=1)$, the expression is a constant function of θ_1.[3] When $a > (b+1)Q(t=1)$, it is minimized at θ_1^L and maximized at θ_1^U. Conversely, if $a < (b+1)Q(t=1)$, the expression is minimized at θ_1^U and maximized at θ_1^L. By (B.28) and $b > 0$, θ_1^L and θ_1^U correspond to θ_0^L and θ_0^U in $\mathcal{G}_{(\theta_1,\theta_0)}(s_1,s_0)$, respectively, showing that Π_{s_1,s_0} are valid bounds for $Q(y=1)$.

To show sharpness, pick an arbitrary point $\pi \in \Pi_{s_1,s_0}$. We demonstrate that one can construct a distribution $Q(t,y)$ such that it is consistent with: *i*) the observed data $Q(t=1)$; *ii*) $Q(y=1) = \pi$; *iii*) the assumptions $(Q(t=1|y=1), Q(t=0|y=0)) =$

3. Note that this is equivalent to prevalence being point identified. Expressions for a and b in (2.6) reveal that this happens if only if $P(t=1)$ in the performance population equals $Q(t=1)$ in the screened population.

$(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$. All marginals of the distribution $Q(t, y)$ are completely determined by observational data and π. To complete the proof, we only need to appropriately specify the dependence structure (θ_1, θ_0) such that it is feasible, i.e. in the identified set for the parameters.

If Assumption 2.3.1 holds, $(Q(t = 1|y = 1), Q(t = 0|y = 0)) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$, so that $\Pi_{s_1,s_0} \neq \emptyset$. Denote then by $\Pi_{s_1,s_0} = [\pi^L, \pi^U] \cap [0, 1]$. Consider the case where $\pi^L = \frac{Q(t=1)+\theta_0^L-1}{\theta_1^L+\theta_0^L-1} \leq \frac{Q(t=1)+\theta_0^U-1}{\theta_1^U+\theta_0^U-1} = \pi^U$. The converse case $\frac{Q(t=1)+\theta_0^L-1}{\theta_1^L+\theta_0^L-1} > \frac{Q(t=1)+\theta_0^U-1}{\theta_1^U+\theta_0^U-1}$ follows a symmetric argument. Let $\theta_1^\beta = \beta\theta_1^L + (1-\beta)\theta_1^U$ for any $\beta \in [0, 1]$. Define:

(B.30)

$$\begin{aligned}\pi^\beta &= \frac{Q(t=1) + \theta_1^\beta b - a}{\theta_1^\beta(b+1) - a} = \frac{\beta\left(Q(t=1) + \theta_1^L b - a\right) + (1-\beta)\left(Q(t=1) + \theta_1^U b - a\right)}{\beta\left(\theta_1^L(b+1) - a\right) + (1-\beta)\left(\theta_1^U(b+1) - a\right)} \\ &= \pi^L \frac{\beta\left(\theta_1^L(b+1) - a\right)}{\beta\left(\theta_1^L(b+1) - a\right) + (1-\beta)\left(\theta_1^U(b+1) - a\right)} \\ &+ \pi^U \frac{(1-\beta)\left(\theta_1^U(b+1) - a\right)}{\beta\left(\theta_1^L(b+1) - a\right) + (1-\beta)\left(\theta_1^U(b+1) - a\right)}\end{aligned}$$

where the second line follows by (B.29) and definition of π^L and π^U. Since $\pi \in [\pi^L, \pi^U]$, then $\exists\alpha \in [0, 1]$ $\pi = \alpha\pi^L + (1-\alpha)\pi^U$. We can define:

$$\beta = \frac{\alpha\left(\theta_1^U(b+1) - a\right)}{\alpha\left(\theta_1^U(b+1) - a\right) + (1-\alpha)\left(\theta_1^L(b+1) - a\right)} \in [0, 1]. \tag{B.31}$$

For β in (B.31), we have $\pi^\beta = \pi$. Then let $(\theta_1^\beta, \theta_0^\beta) = \beta(\theta_1^L, \theta_0^L) + (1-\beta)(\theta_1^U, \theta_0^U)$. Since $(\theta_1^\beta, \theta_0^\beta)$ is a linear combination of endpoints of a line segment, it must also be an element of the line segment. Hence $(\theta_1^\beta, \theta_0^\beta) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0)$, proving that for any $\pi \in \Pi_{s_1,s_0}$ we can construct $Q(t, y)$ such that it is consistent with the observed data and assumptions, with $Q(y = 1) = \pi$. Hence Π_{s_1,s_0} is sharp.

Case 2: $\exists(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}(s_1, s_0) : \theta_1 + \theta_0 = 1$

Fix $(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)} : \theta_1 + \theta_0 = 1$. Then $Q(t = 1|y = 1) = Q(t = 1|y = 0)$, so $t \perp\!\!\!\perp y$ is consistent with Assumption 2.3.1. Hence $Q(y = 1) \in [0, 1]$ is consistent with any observed $Q(t = 1)$. Sharpness is also immediate since for an arbitrary point $\pi \in [0, 1]$, we can fix $(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)} : \theta_1 + \theta_0 = 1$ and define $Q(t, y)$ with $Q(t, y = 1) = Q(t)\pi$ and $Q(t, y = 0) = Q(t)(1 - \pi)$ for any $Q(t)$. Thus, there exists a distribution $Q(t, y)$ which is consistent with $Q(t = 1)$, $(\theta_1, \theta_0) \in \mathcal{G}_{(\theta_1,\theta_0)}$ and $Q(y = 1) = \pi$.

□

Proof of Proposition B.2.1. The proof proceeds by finding $E_P\big(m_j(W_i, \theta)\big)$ for $j = 1, 2 \ldots, 7$ and demonstrating that the resulting system is equivalent to the bounds defined in Proposition 2.2.1 extended to $\mathbf{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \Big(\mathcal{H}_{(\theta_1,\theta_0)}(s_1, s_0) \times \{(s_1, s_0)\}\Big)$.

Suppose that $E_P\big(m_j(W_i,\theta)\big) \leq 0$ for $j = 1, 2 \ldots, 6$ and $E_P\big(m_7(W_i,\theta)\big) = 0$. From (B.5):

(B.32)

$$\begin{aligned}
E_P\big(m_1(W_i,\theta)\big) &= -\theta_1 P_{s_1,s_0}(y=1) + P_{s_1,s_0}(r=1,y=1) - P(r=1) + P(t=1,r=1) \\
&= P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0) - \theta_1 P_{s_1,s_0}(y=1) \leq 0 \\
E_P\big(m_2(W_i,\theta)\big) &= (-\theta_1 + 1 - s_1) P_{s_1,s_0}(y=1) - P(t=0,r=0) \\
&= P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0) - \theta_1 P_{s_1,s_0}(y=1) \leq 0 \\
E_P\big(m_3(W_i,\theta)\big) &= (-\theta_1 + 1) P_{s_1,s_0}(y=1) + P(t=1) - 1 \\
&= P_{s_1,s_0}(y=1)s_1 - P(r=1) + P_{s_1,s_0}(y=1)(1-s_1) + P(t=1) \\
&\quad - P(r=0) - \theta_1 P_{s_1,s_0}(y=1) \\
&= P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0) + P_{s_1,s_0}(r=0,y=1) \\
&\quad - P(t=0,r=0) - \theta_1 P_{s_1,s_0}(y=1) \leq 0.
\end{aligned}$$

Note further that if $\theta_1 \in [0,1]$, which is true by definition, the three inequalities above yield the lower bound from Proposition 2.2.1 for $\theta_1 \in \mathcal{H}_{\theta_1}(s_1,s_0)$ given an arbitrary $(s_1,s_0) \in \mathcal{S}$:

(B.33)

$$\begin{aligned}
\theta_1 P_{s_1,s_0}(y=1) \geq\ & max\Big(0, P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0)\Big) \\
& + max\Big(0, P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0)\Big).
\end{aligned}$$

This is equivalent to the lower bound for the element θ_1 of $(\theta_1, \theta_0, s_1, s_0) \in \boldsymbol{\Theta}(P)$. Consider next:

$$
\begin{aligned}
E_P\big(m_4(W_i,\theta)\big) &= \theta_1 P_{s_1,s_0}(y=1) - P(t=1) \\
&= \theta_1 P_{s_1,s_0}(y=1) - P(t=1,r=0) - P(t=1,r=1) \leq 0 \\
E_P\big(m_5(W_i,\theta)\big) &= (\theta_1 - s_1) P_{s_1,s_0}(y=1) - P(t=1,r=0) \\
&= \theta_1 P_{s_1,s_0}(y=1) - P(t=1,r=0) - P_{s_1,s_0}(r=1,y=1) \leq 0 \\
E_P\big(m_6(W_i,\theta)\big) &= (\theta_1 - 1 + s_1) P_{s_1,s_0}(y=1) - P(t=1,r=1) \\
&= \theta_1 P_{s_1,s_0}(y=1) - P_{s_1,s_0}(r=0,y=1) - P(t=1,r=1) \leq 0
\end{aligned}
\tag{B.34}
$$

Similarly, the upper bound from Proposition 2.2.1 is obtained for the element θ_1 of $(\theta_1, \theta_0, s_1, s_0) \in \boldsymbol{\Theta}(P)$:

$$
\begin{aligned}
\theta_1 P_{s_1,s_0}(y=1) \leq\ & min\Big(P(t=1,r=0), P_{s_1,s_0}(r=0,y=1)\Big) \\
& + min\Big(P(t=1,r=1), P_{s_1,s_0}(r=1,y=1)\Big).
\end{aligned}
\tag{B.35}
$$

Taking the expected value of the final component of the moment function yields:

$$
E_P\big(m_7(W_i,\theta)\big) = (\theta_0 - 1)(1 - P_{s_1,s_0}(y=1)) - \theta_1 P_{s_1,s_0}(y=1) + P(t=1) = 0 \tag{B.36}
$$

It is then is true that $\theta_0 P_{s_1,s_0}(y=0) = P_{s_1,s_0}(y=0) + \theta_1 P_{s_1,s_0}(y=1) - P(t=1)$. This is the linear relationship between (θ_1, θ_0) in the identified set from Proposition 2.2.1. Going in the other direction, it is immediate that if the two bounds and the linear relationship hold so that $\theta \in \boldsymbol{\Theta}(P)$, then $E_P\big(m_j(W_i,\theta)\big) \leq 0$ for $j = 1, 2 \ldots, 6$ and

$E_P\big(m_7(W_i,\theta)\big) = 0$, demonstrating that the expected values of moment functions represent the joint identification region $\theta \in \boldsymbol{\Theta}(P)$. □

Proof of Proposition B.2.2. The proof is analogous to the proof of Proposition B.2.1. From the definition of $\bar{\bar{\mathcal{H}}}_{\theta_1}(s_1, s_0)$ for $j = 1$ in Proposition 2.2.2:

(B.37)

$$\begin{aligned}
\theta_1 P_{s_1,s_0}(y=1) \geq\ & max\Big(0, P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0)\Big) \\
& + max\Big(0, P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0)\Big) \\
\theta_1 P_{s_1,s_0}(y=1) \leq\ & min\Big(P(t=1,r=0), \frac{P_{s_1,s_0}(r=0,y=1)}{2}\Big) \\
& + min\Big(P(t=1,r=1) - \frac{P_{s_1,s_0}(r=1,y=0)}{2}, P_{s_1,s_0}(r=1,y=1)\Big).
\end{aligned}$$

Suppose that $E_P\big(\bar{\bar{m}}_j(W_i,\theta)\big) \leq 0$ for $j = 1, 2 \ldots, 6$ and $E_P\big(\bar{\bar{m}}_7(W_i,\theta)\big) = 0$. From (B.8):

$$\begin{aligned}
E_P\big(\bar{\bar{m}}_4(W_i,\theta)\big) &= \theta_1 P_{s_1,s_0}(y=1) - P(t=1) + \frac{1}{2}\Big(P(r=1) - s_1 P_{s_1,s_0}(y=1)\Big) \\
&= \theta_1 P_{s_1,s_0}(y=1) - P(t=1,r=0) \\
&\quad - P(t=1,r=1) + \frac{P_{s_1,s_0}(r=1,y=0)}{2} \leq 0
\end{aligned}$$

(B.38)

$$\begin{aligned}
E\big(\bar{\bar{m}}_6(W_i,\theta)\big) &= \Big(\theta_1 + \frac{-1+s_1}{2}\Big) P_{s_1,s_0}(y=1) \\
&\quad - P(t=1,r=1) + \frac{1}{2}\Big(P(r=1) - s_1 P_{s_1,s_0}(y=1)\Big) \\
&= \theta_1 P_{s_1,s_0}(y=1) - \frac{P_{s_1,s_0}(r=0,y=1)}{2} \\
&\quad - P(t=1,r=1) + \frac{P_{s_1,s_0}(r=1,y=0)}{2} \leq 0
\end{aligned}$$

Note that $E_P\big(\bar{\bar{m}}_j(W_i,\theta)\big) = E_P\big(m_j(W_i,\theta)\big)$ for $j = 1, 2, 3, 5, 7$. Then, by (B.32), (B.34), (B.36), and (B.38) $E_P\big(\bar{\bar{m}}_j(W_i,\theta)\big) \leq 0$ for $j = 1, \ldots, 6$ and $E_P\big(\bar{\bar{m}}_7(W_i,\theta)\big) = 0$ represent

the joint identification $\mathbf{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \left(\bar{\bar{\mathcal{H}}}_{(\theta_1,\theta_0)}(s_1,s_0) \times \{(s_1,s_0)\}\right)$ by the same argument as in the proof of Proposition B.2.1. □

PROOF OF PROPOSITION 2.4.1. From the definition of $\bar{\mathcal{H}}_{\theta_1}(s_1,s_0)$ for $y=1$ in Proposition 2.2.2:

$$
\begin{aligned}
\theta_1 P_{s_1,s_0}(y=1) \geq\ & max\Big(0, P(t=1,r=1) - P_{s_1,s_0}(r=1,y=0)\Big) \\
& + max\Big(0, P_{s_1,s_0}(r=0,y=1) - P(t=0,r=0)\Big) \\
\theta_1 P_{s_1,s_0}(y=1) \leq\ & min\Big(P(t=1,r=0), \frac{P_{s_1,s_0}(r=0,y=1)}{2}\Big) \\
& + min\Big(P(t=1,r=1), P_{s_1,s_0}(r=1,y=1)\Big).
\end{aligned} \tag{B.39}
$$

Suppose that $E_P\big(\bar{m}_j^1(W_i,\theta)\big) \leq 0$ for $j = 1, 2\ldots, 6$ and $E_P\big(\bar{m}_7^1(W_i,\theta)\big) = 0$. From (2.18):

$$
\begin{aligned}
E_P\big(\bar{m}_6^1(W_i,\theta)\big) &= \Big(\theta_1 + \frac{-1+s_1}{2}\Big) P_{s_1,s_0}(y=1) - P(t=1,r=1) \\
&= \theta_1 P_{s_1,s_0}(y=1) - \frac{P_{s_1,s_0}(r=0,y=1)}{2} - P(t=1,r=1) \leq 0
\end{aligned} \tag{B.40}
$$

Note that $E_P\big(\bar{m}_j^1(W_i,\theta)\big) = E_P\big(m_j(W_i,\theta)\big)$ for $j = 1,2,3,4,5,7$. Then, by (B.32), (B.34), (B.36), and (B.40), $E_P\big(\bar{m}_j^1(W_i,\theta)\big) \leq 0$ for $j = 1,\ldots,7$ and $E_P\big(\bar{m}_7^1(W_i,\theta)\big) = 0$ represent the joint identification region $\mathbf{\Theta}(P) = \bigcup_{(s_1,s_0)\in\mathcal{S}} \left(\bar{\mathcal{H}}_{(\theta_1,\theta_0)}(s_1,s_0) \times \{(s_1,s_0)\}\right)$ by the same argument as in the proof of Proposition B.2.1. □

PROOF OF THEOREM 2.4.1. I first show that $Var_P(m_j(W_i,\theta)) > \frac{1}{M_j^2} > 0$, for any $j \in 1,\ldots,7$ in (B.5) under the assumptions, where M_j do not depend on P and θ. I then demonstrate the same for components (2.18), (B.7), and (B.8) that are not identical. Finally, I show that $m_j(W_i,\theta)$ are bounded irrespective of P and θ, and use that to prove that the second claim is true.

Let $\rho_P(X,Y) = \frac{Cov_P(X,Y)}{\sqrt{Var_P(X)Var_P(Y)}}$ for some binary random vector (X,Y) with distribution $P \in \mathbf{P}$. The following Lemma will be used to bound the variances from below.

Lemma B.5.2. *Suppose that Assumption 2.4.1 holds. Then for any $P \in \mathbf{P}$, the following are true:*

(1) $\max_{P\in\mathbf{P}} \rho_P(r_i,t_i)^2 = (1-4\varepsilon)^2 < 1$;

(2) $\max_{P\in\mathbf{P}} \rho_P(r_i,(1-t_i))^2 = (1-4\varepsilon)^2 < 1$;

(3) $\max_{P\in\mathbf{P}} \rho_P(r_i,r_it_i)^2 = h(\varepsilon)$;

(4) $\max_{P\in\mathbf{P}} \rho_P(r_i,(1-r_i)t_i)^2 = h(\varepsilon)$

(5) $\max_{P\in\mathbf{P}} \rho_P(r_i,r_i(1-t_i))^2 = h(\varepsilon)$

(6) $\max_{P\in\mathbf{P}} \rho_P(r_i,(1-r_i)(1-t_i))^2 = h(\varepsilon)$

where $h(\varepsilon) = \mathbb{1}\{\varepsilon \in [0.2,0.25]\}\frac{2-6\varepsilon}{3-6\varepsilon} + \mathbb{1}\{\varepsilon \in (0,0.2)\}\left(1-\frac{(1-\varepsilon)^2}{(1+\varepsilon)^2}\right) \in (0,1)$.

Proof. Denote $P(t_i = j, r_i = k) = P_{jk}$. Assumption 2.4.1 states that for $(j,k) \in \{0,1\}^2$, $P_{jk} \geq \varepsilon > 0$, and implies that $\varepsilon \leq \frac{1}{4}$.

Statements 1 and 2

Parameter $\rho_P(r_i,t_i)^2$ is the largest when either $P_{01} = P_{10} = \varepsilon$ or $P_{11} = P_{00} = \varepsilon$. I prove the statement for $P_{01} = P_{10} = \varepsilon$, and the argument for the $P_{11} = P_{00} = \varepsilon$ is symmetric. The maximal $\rho_P(r_i,t_i)^2$ must then be for $P_{11} + P_{00} = 1 - 2\varepsilon$ and $P(t_i = 1) = P(r_i = 1)$.

Next, let $P_{11} = \alpha(1-2\varepsilon)$, $P_{00} = (1-\alpha)(1-2\varepsilon)$ for some $\alpha \in [\frac{\varepsilon}{1-2\varepsilon}, \frac{1-3\varepsilon}{1-2\varepsilon}]$, and $P(t_i = 1) = P(r_i = 1) = \alpha(1-2\varepsilon)+\varepsilon$. By plugging in the relevant probabilities, $\rho_P(r_i,t_i)$

becomes a function of α:

$$\begin{aligned}\rho_\alpha(r_i,t_i) &= \frac{P_{11}-P(t_i=1)P(r_i=1)}{\sqrt{P(t_i=1)(1-P(t_i=1))P(r_i=1)(1-P(r_i=1))}} = \\ &= \frac{\alpha(1-2\varepsilon)-(\alpha(1-2\varepsilon)+\varepsilon)^2}{(\alpha(1-2\varepsilon)+\varepsilon)\,(1-\alpha(1-2\varepsilon)-\varepsilon)}.\end{aligned} \tag{B.41}$$

Since we are considering the case $P_{01} = P_{10} = \varepsilon$, the correlation is positive. By maximizing $\rho_\alpha(r_i,t_i)$ with respect to α, we obtain the upper bound on $\rho_P(r_i,t_i)^2$. The second order condition confirms that this is a concave optimization problem. The first order condition yields the maximizing $\alpha^* = \frac{1}{2}$.

For any $\varepsilon \leq \frac{1}{4}$, it is true that $\alpha^* \in [\frac{\varepsilon}{1-2\varepsilon}, \frac{1-3\varepsilon}{1-2\varepsilon}]$. To conclude the proof of statement 1, plug in α^* into (B.41) to find $\max_{P\in\mathbf{P}} \rho_P(r_i,t_i) = \rho_{\alpha^*}(r_i,t_i) = (1-4\varepsilon)$.

By replacing $\tilde{t}_i = 1-t_i$ in Statement 1, it follows directly that $\max_{P\in\mathbf{P}} \rho_P(r_i,(1-t_i)) = \max_{P\in\mathbf{P}} \rho_P(r_i,\tilde{t}_i) = \rho_{\alpha^*}(r_i,\tilde{t}_i) = (1-4\varepsilon)$.

Statement 3

From the definition of $\rho_P(r_i, r_i t_i)$:

(B.42)

$$\begin{aligned}\rho_P(r_i,r_it_i) &= \frac{Cov_P(r_i,t_ir_i)}{\sqrt{Var_P(r_i)Var_P(t_ir_i)}} = \frac{E_P(t_ir_i)(1-E_P(r_i))}{\sqrt{E_P(r_i)(1-E_P(r_i))E_P(t_ir_i)(1-E_P(t_ir_i))}} \\ &= \sqrt{\frac{E_P(t_ir_i)(1-E_P(r_i))}{E_P(r_i)(1-E_P(t_ir_i))}} = \sqrt{\frac{P_{11}(1-P(r_i=1))}{P(r_i=1)(1-P_{11})}} \\ &= \sqrt{\frac{P_{11}(1-P_{11}-P_{01})}{(P_{11}+P_{01})(1-P_{11})}}.\end{aligned}$$

Notice that $\rho_P(r_i, r_i t_i)$ decreases in P_{01}, so at the maximum, $P_{01} = \varepsilon$. Therefore, we only need to maximize $\rho_P(r_i, r_i t_i)^2$ with respect to feasible P_{11}. The maximization problem is:

$$\max_{P \in \mathbf{P}} \rho_P(r_i, r_i t_i) = \max_{P_{11} \in [\varepsilon, 1-3\varepsilon]} \sqrt{\frac{P_{11}(1 - P_{11} - \varepsilon)}{(P_{11} + \varepsilon)(1 - P_{11})}}. \tag{B.43}$$

The objective function is concave. The first order condition implies that for an interior maximum, the maximizing P_{11} is $\frac{1-\varepsilon}{2}$. If $\varepsilon \in [0.2, 0.25]$, the constraint $P_{11} \leq 1 - 3\varepsilon$ will bind. Therefore, the value of the parameter at the maximum is $P_{11}^* = \min\left\{\frac{1-\varepsilon}{2}, 1 - 3\varepsilon\right\}$. The maximum of the objective function obtained by plugging in P_{11}^* into (B.42) is:

$$\begin{aligned} \max_{P \in \mathbf{P}} \rho_P(r_i, r_i t_i)^2 = {} & \mathbb{1}\{\varepsilon \in [0.2, 0.25]\} \left(1 - \frac{1}{3 - 6\varepsilon}\right) \\ & + \mathbb{1}\{\varepsilon \in (0, 0.2)\} \left(1 - \frac{(1-\varepsilon)^2}{(1+\varepsilon)^2}\right) \in (0, 1) \end{aligned} \tag{B.44}$$

Statements 4, 5, and 6

Following the definition of $\rho_P(r_i, (1 - r_i)t_i)$:

$$\begin{aligned} \rho_P(r_i, (1-r_i)t_i) &= \frac{Cov_P(r_i, (1-r_i)t_i)}{\sqrt{Var_P(r_i) Var_P((1-r_i)t_i)}} \\ &= \frac{-E_P(r_i) E_P((1-r_i)t_i)}{\sqrt{E_P(r_i)(1 - E_P(r_i)) E_P((1-r_i)t_i)(1 - E_P((1-r_i)t_i))}} \\ &= -\sqrt{\frac{E_P(r_i) E_P((1-r_i)t_i)}{(1 - E_P(r_i))(1 - E_P((1-r_i)t_i))}} \\ &= -\sqrt{\frac{P(r_i = 1) P_{10}}{(1 - P(r_i = 1))(1 - P_{10})}}. \end{aligned} \tag{B.45}$$

The square of the correlation is increasing in both $P(r_i = 1) = P_{11} + P_{01}$ and P_{10}. Consequently, at the maximum, together they will be at the upper bound, meaning that

$P_{11} + P_{01} + P_{10} = 1 - \varepsilon$, or equivalently, that $P(r_i = 1) = 1 - \varepsilon - P_{10}$. We can then rewrite the problem as:

$$\max_{P \in \mathbf{P}} \rho_P(r_i, (1 - r_i)t_i)^2 = \max_{P_{10} \in [\varepsilon, 1-3\varepsilon]} \frac{(1 - \varepsilon - P_{10})P_{10}}{(\varepsilon + P_{10})(1 - P_{10})}. \tag{B.46}$$

In this form, the problem is identical to the one in (B.43). Following the same steps:

$$\begin{aligned} \max_{P \in \mathbf{P}} \rho_P(r_i, (1 - r_i)t_i)^2 = \mathbb{1}\{\varepsilon \in [0.2, 0.25]\} \left(1 - \frac{1}{3 - 6\varepsilon}\right) \\ + \mathbb{1}\{\varepsilon \in (0, 0.2)\} \left(1 - \frac{(1 - \varepsilon)^2}{(1 + \varepsilon)^2}\right) < 1. \end{aligned} \tag{B.47}$$

Analogously to the proof of Statement 3, for $\rho(r_i, (1 - t_i)r_i)^2$ in Statement 5, the optimization problem can be represented as:

$$\max_{P \in \mathbf{P}} \rho_P(r_i, r_i(1 - t_i))^2 = \max_{P_{01} \in [\varepsilon, 1-3\varepsilon]} \frac{(1 - \varepsilon - P_{01})P_{01}}{(\varepsilon + P_{01})(1 - P_{01})}. \tag{B.48}$$

Following the steps in the proof of Statement 4 $\rho(r_i, (1 - r_i)(1 - t_i))^2$ in Statement 6, the optimization problem will be:

$$\max_{P \in \mathbf{P}} \rho_P(r_i, (1 - r_i)t_i)^2 = \max_{P_{10} \in [\varepsilon, 1-3\varepsilon]} \frac{(1 - \varepsilon - P_{00})P_{00}}{(\varepsilon + P_{00}P)(1 - P_{00})}. \tag{B.49}$$

Consequently, from the solutions to (B.43) and (B.45), (B.48) and (B.49) will yield the same upper bounds on their corresponding squares of correlations:

$$\begin{aligned}\max_{P\in\mathbf{P}} \rho_P(r_i, r_i(1-t_i))^2 &= \max_{P\in\mathbf{P}} \rho_P(r_i, (1-r_i)t_i)^2 \\ &= \mathbb{1}\{\varepsilon \in [0.2, 0.25]\}\left(1 - \frac{1}{3-6\varepsilon}\right) \\ &+ \mathbb{1}\{\varepsilon \in (0, 0.2)\}\left(1 - \frac{(1-\varepsilon)^2}{(1+\varepsilon)^2}\right).\end{aligned} \tag{B.50}$$

□

Claim B.5.6. *For any $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$ it holds that $Var_P(m_j(W_i, \theta)) > 0$ for all $m_j(W_i, \theta)$ in (B.5).*

Proof. Consider first a component of m pertaining to the upper bound of θ_1. The variance $Var_P(m_4(W_i, \theta))$ for some θ and P is defined as:

(B.51)

$$\begin{aligned}Var_P(m_4(W_i, \theta)) &= Var_P\left(\theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - t_i\right) \\ &= \left(\frac{\theta_1}{s_1 - 1 + s_0}\right)^2 Var_P(r_i) + Var_P(t_i) - 2\frac{\theta_1}{s_1 - 1 + s_0} Cov_P(r_i, t_i).\end{aligned}$$

Fix any $(s_1, s_0) \in \mathcal{S}$. As discussed in Section 2.2.4, Assumptions 2A and 2.2.3 imply $P(r = 1) \in (1 - s_0, s_1)$ so $Var_P(r_i) > 0$. The value θ_1^* where $Var_P(m_4(W_i, \theta))$ is globally minimized given s_1 and s_0 from the first order condition is:

$$\frac{\partial Var_P(m_4(W_i, \theta))}{\partial \theta_1} : \quad \theta_1^* = (s_1 - 1 + s_0)\frac{Cov_P(r_i, t_i)}{Var_P(r_i)}. \tag{B.52}$$

The second order condition shows that this indeed is a minimization problem. Let $\theta^* = (\theta_1^*, \theta_0, s_1, s_0)$, where I suppress the dependence $\theta_1^*(s_1, s_0)$ for clarity. The minimum variance for any $(s_1, s_0) \in \mathcal{S}$ is then:

$$\begin{aligned} Var_P(m_4(W_i, \theta^*)) &= \frac{(Cov_P(r_i, t_i))^2}{Var_P(r_i)} + Var_P(t_i) - 2\frac{(Cov_P(r_i, t_i))^2}{Var_P(r_i)} \\ &= Var_P(t_i)\left(1 - \rho_P(r_i, t_i)^2)\right). \end{aligned} \tag{B.53}$$

For any θ it follows:

$$\begin{aligned} Var_P(m_4(W_i, \theta)) &\geq Var_P(m_4(W_i, \theta^*)) = Var_P(t_i)\left(1 - \rho_P(r_i, t_i)^2)\right) \\ &\geq 2\varepsilon(1 - 2\varepsilon)\left(1 - (1 - 4\varepsilon)^2\right) = \frac{1}{M_4^2} > 0 \end{aligned} \tag{B.54}$$

where the first inequality follows from the definition of θ^*. Focus on the second inequality. We wish to find the lower bound on the variance $\frac{1}{M_4^2}$ over all possible $P \in \mathbf{P}$. One such bound is equal to the expression at the smallest value of $Var_P(t_i)$ and the largest value of $\rho_P(r_i, t_i)^2$. The second is given by Lemma B.5.2, and the first follows directly from Assumption 2.4.1 which implies that $P(t_i = 1) \in [2\varepsilon, 1 - 2\varepsilon]$, so $Var_P(t_i) \geq 2\varepsilon(1 - 2\varepsilon)$.[4] Therefore, $Var_P(m_4(W_i, \theta)) \geq \frac{1}{M_4^2} > 0$ for all $P \in \mathbf{P}$ and $\theta \in [0, 1]^2 \times \mathcal{S}$.

4. As long as $\varepsilon < 0.25$, the inequality is strict, since the largest value of $\rho_P(r_i, t_i)^2$ warrants that $P(t_i = 1, r_i = 1) = \frac{1-2\varepsilon}{2}$ while the smallest $Var_P(t_i)$ requires $P(t_i = 1, r_i = 1) = \varepsilon$ or $P(t_i = 1, r_i = 1) = 1 - 3\varepsilon$.

Following the same steps for the remaining components pertaining to the upper bound, the smallest variances for any $P \in \mathbf{P}$ and θ are:

$$\begin{aligned} Var_P(m_5(W_i, \theta^*)) &= Var_P(t_i(1-r_i)) \left(1 - \rho_P(r_i, t_i(1-r_i))^2\right) \\ &\geq \varepsilon(1-\varepsilon)\left(1 - h(\varepsilon)\right) = \frac{1}{M_5^2} > 0 \\ Var_P(m_6(W_i, \theta^*)) &= Var_P(t_i r_i) \left(1 - \rho_P(r_i, t_i r_i)^2\right) \\ &\geq \varepsilon(1-\varepsilon)\left(1 - h(\varepsilon)\right) = \frac{1}{M_6^2} > 0 \end{aligned} \tag{B.55}$$

where the inequalities follow from the definition of θ^*, the fact that $Var_P(t_i(1-r_i)) \geq \varepsilon(1-\varepsilon)$ and $Var_P(t_i(1-r_i)) \geq \varepsilon(1-\varepsilon)$, and Lemma B.5.2.

Next observe the components pertaining to the lower bound. First for $Var_P(m_1(W_i, \theta))$ for any θ and P:

$$\begin{aligned} Var_P(m_1(W_i, \theta)) &= Var_P\left((-\theta_1 + s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + (t_i - 1)r_i\right) \\ &= \left(\frac{s_1 - \theta_1}{s_1 - 1 + s_0}\right)^2 Var_P(r_i) + Var_P((t_i - 1)r_i) \\ &\quad - 2\frac{s_1 - \theta_1}{s_1 - 1 + s_0} Cov_P((1 - t_i)r_i, r_i) \end{aligned} \tag{B.56}$$

Fix an arbitrary s_1 and s_0. The value θ_1^* where $Var_P(m_1(W_i, \theta))$ is globally minimized given s_1 and s_0 from the first order condition is:

$$\frac{\partial Var_P(m_1(W_i, \theta))}{\partial \theta_1} : \quad \theta_1^* = (s_1 - 1 + s_0)\frac{Cov_P((1 - t_i)r_i, r_i)}{Var_P(r_i)} + s_1. \tag{B.57}$$

The second order condition shows that this indeed is a minimization problem. The minimum variance $Var_P(m_1(W_i, \theta^*))$ for an arbitrary $(s_1, s_0) \in \mathcal{S}$ is:

$$\begin{aligned} Var_P(m_1(W_i, \theta^*)) &= \left(-\frac{Cov_P((1-t_i)r_i, r_i))}{Var_P(r_i)}\right)^2 Var_P(r_i) + Var_P((1-t_i)r_i) \\ &\quad - 2\left(\frac{Cov_P((1-t_i)r_i, r_i)}{Var_P(r_i)}\right) Cov_P((1-t_i)r_i, r_i) \\ &= Var_P((1-t_i)r_i)\left(1 - \frac{(Cov_P((1-t_i)r_i, r_i))^2}{Var_P(r_i)Var_P((1-t_i)r_i)}\right) \\ &= Var_P((1-t_i)r_i)\left(1 - \rho_P((1-t_i)r_i, r_i)^2\right) \\ &\geq \varepsilon(1-\varepsilon)(1-h(\varepsilon)) = \frac{1}{M_1^2} > 0 \end{aligned} \tag{B.58}$$

where the first inequality follows from the definition of θ^*. And the second follows from Lemma B.5.2 and $Var_P(t_i r_i) \geq \varepsilon(1-\varepsilon)$. Therefore $Var_P(m_1(W_i, \theta)) \geq \frac{1}{M_1^2} > 0$ for all $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$.

Again, following the same steps for the remaining components pertaining to the lower bound, the smallest variances for any $P \in \mathbf{P}$ and θ are:

$$\begin{aligned} Var_P(m_2(W_i, \theta^*)) &= Var_P((1-t_i)(1-r_i))\left(1 - \rho_P(r_i, (1-t_i)(1-r_i))^2\right) \\ &\geq \varepsilon(1-\varepsilon)\left(1 - h(\varepsilon)\right) = \frac{1}{M_2^2} > 0 \\ Var_P(m_3(W_i, \theta^*)) &= Var_P(1-t_i)\left(1 - \rho_P(r_i, (1-t_i))^2\right) \\ &\geq 2\varepsilon(1-2\varepsilon)\left(1 - (1-4\varepsilon)^2\right) = \frac{1}{M_3^2} > 0 \end{aligned} \tag{B.59}$$

Finally, consider the component pertaining to the moment equality $Var(m_7(W_i,\theta))$. It is defined as:

(B.60)
$$\begin{aligned} Var_P(m_7(W_i,\theta)) &= Var_P\left((1-\theta_0)\left(1-\frac{r_i-1+s_0}{s_1-1+s_0}\right)+\theta_1\frac{r_i-1+s_0}{s_1-1+s_0}-t_i\right) \\ &= Var_P\left((\theta_0+\theta_1-1)\frac{r_i-1+s_0}{s_1-1+s_0}-t_i\right) \\ &= Var_P\left(\bar{\theta}\frac{r_i-1+s_0}{s_1-1+s_0}-t_i\right) \\ &= \left(\frac{\bar{\theta}}{s_1-1+s_0}\right)^2 Var_P(r_i)+Var_P(t_i)-2\frac{\bar{\theta}}{s_1-1+s_0}Cov_P(r_i,t_i) \end{aligned}$$

for $\theta_1+\theta_0-1=\bar{\theta}$. Notice how the function (B.60) resembles (B.51). Following the same steps as for finding $\frac{1}{M_4^2}$, we obtain that $Var(m_7(W_i,\theta)) \geq 2\varepsilon(1-2\varepsilon)\left(1-(1-4\varepsilon)^2\right) = \frac{1}{M_7^2} > 0$ for all $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$. □

Claim B.5.7. *For any $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$ it holds that $Var_P(m_j(W_i,\theta)) > 0$ for all $m_j(W_i,\theta)$ in (2.18), (B.7), and (B.8).*

Proof. Functions $\bar{m}$ and m are such that $\bar{m}_j^1(W_i,\theta) \neq m_j(W_i,\theta)$ only if $j=6$. Thus for all other components, the proof follows from Claim B.5.6, so $Var_P(\bar{m}_j^1(W_i,\theta)) \geq \frac{1}{M_j^2} > 0$ for $j \neq 6$.

The variance $Var_P(\bar{m}_6^1(W_i,\theta))$ for some θ and P is:

(B.61)
$$\begin{aligned} Var_P(\bar{m}_6^1(W_i,\theta)) &= Var_P\left(\left(\theta_1+\frac{-1+s_1}{2}\right)\frac{r_i-1+s_0}{s_1-1+s_0}-t_ir_i\right) \\ &= \left(\frac{\theta_1+\frac{-1+s_1}{2}}{s_1-1+s_0}\right)^2 Var_P(r_i)+Var_P(r_it_i)-2\frac{\theta_1+\frac{-1+s_1}{2}}{s_1-1+s_0}Cov_P(r_i,r_it_i). \end{aligned}$$

Fix any $(s_1, s_0) \in \mathcal{S}$. The value θ_1^* where $Var_P(m_6(W_i, \theta))$ is globally minimized given s_1 and s_0 from the first order condition is:

$$\frac{\partial Var_P(\bar{m}_6^1(W_i,\theta))}{\partial \theta_1} : \quad \theta_1^* = (s_1 - 1 + s_0)\frac{Cov_P(r_i, r_i t_i)}{Var_P(r_i)} + \frac{1 - s_1}{2}. \tag{B.62}$$

The second order condition shows that this indeed is a minimization problem. Following the same steps as before, for any $\theta \in [0,1]^2 \times \mathcal{S}$:

$$\begin{aligned} Var_P(Cov_P(r_i, r_i t_i)(W_i, \theta)) &\geq Var_P(Cov_P(r_i, r_i t_i)(W_i, \theta^*)) \\ &= Var_P(r_i t_i)\left(1 - \rho_P(r_i, r_i t_i)^2)\right) \\ &\geq \varepsilon(1-\varepsilon)\left(1 - h(\varepsilon)\right) = \frac{1}{M_6^2} > 0 \end{aligned} \tag{B.63}$$

The case for $\bar{m}_j^0(W_i, \theta)$ is symmetric, $Var_P(\bar{m}_j^0(W_i, \theta)) \geq \frac{1}{M_j^2} > 0$ for all $j = 1, \ldots, 7$, $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$.

Likewise, for $\bar{\bar{m}}$ note that $\bar{\bar{m}}_j(W_i, \theta) = \bar{m}_j(W_i, \theta)$ except for $j \in \{4, 6\}$. From (B.8):

(B.64)

$$\begin{aligned} V_P(\bar{\bar{m}}_4(W_i,\theta)) &= V_P\left(\theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - t_i + \frac{1}{2}\Big(r_i - s_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0}\Big)\right) \\ &= V_P\left(\left(\frac{\theta_1 - \frac{s_1}{2}}{s_1 - 1 + s_0} + \frac{1}{2}\right) r_i - t_i\right) \\ V_P(\bar{\bar{m}}_6(W_i,\theta)) &= V_P\left(\Big(\theta_1 + \frac{-1 + s_1}{2}\Big)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - t_i r_i + \frac{1}{2}\Big(r_i - s_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0}\Big)\right) \\ &= V_P\left(\left(\frac{\theta_1 - \frac{1}{2}}{s_1 - 1 + s_0} + \frac{1}{2}\right) r_i - t_i r_i\right). \end{aligned}$$

As above, for any $\theta \in [0,1]^2 \times \mathcal{S}$:

$$
\begin{aligned}
V_P(\bar{\bar{m}}_4(W_i,\theta)) \geq V_P(\bar{\bar{m}}_4(W_i,\theta^*)) &= V_P(t_i)\left(1-\rho_P(r_i,t_i)^2\right) \\
&\geq 2\varepsilon(1-2\varepsilon)\left(1-(1-4\varepsilon)^2\right) = \frac{1}{M_4^2} > 0 \\
V_P(\bar{\bar{m}}_6(W_i,\theta)) \geq V_P(\bar{\bar{m}}_6(W_i,\theta^*) &= V_P(t_i r_i)\left(1-\rho_P(r_i,t_i r_i)^2\right) \\
&\geq \varepsilon(1-\varepsilon)\left(1-h(\varepsilon)\right) = \frac{1}{M_6^2} > 0.
\end{aligned} \tag{B.65}
$$

It is true that $Var_P(\bar{\bar{m}}_j^0(W_i,\theta)) \geq \frac{1}{M_j^2} > 0$ for all $j = 1,\ldots,7$, $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$. □

Claim B.5.8. *For any $P \in \mathbf{P}$, $\theta \in [0,1]^2 \times \mathcal{S}$, and $m_j(W_i,\theta)$ in (B.5), (2.18), (B.7), and (B.8):*

(B.66)

$$
\limsup_{\lambda \longrightarrow \infty} \sup_{P \in \mathbf{P}} \sup_{\theta \in \mathbf{\Theta}(P)} E_P\left[\left(\frac{m_j(W_i,\theta)-\mu_j(\theta,P)}{\sigma_j(\theta,P)}\right)^2 \mathbb{1}\left\{\left|\frac{m_j(W_i,\theta)-\mu_j(\theta,P)}{\sigma_j(\theta,P)}\right| > \lambda\right\}\right] = 0.
$$

We have shown above that for any $\sigma_j(\theta,P)$ corresponding to components $m_j(W_i,\theta)$ in (B.5), (2.18), (B.7), and (B.8), there exists a finite constant $M_j > 0$ such that $\sigma_j(\theta,P) \geq \frac{1}{M_j} > 0$ for all $P \in \mathbf{P}$ and $\theta \in [0,1]^2 \times \mathcal{S}$.

Then for any j, $P \in \mathbf{P}$, $\theta \in [0,1]^2 \times \mathcal{S}$ and λ:

$$
E_P\left[M_j^2\left(m_j(W_i,\theta)-\mu_j(\theta,P)\right)^2 \mathbb{1}\left\{|m_j(W_i,\theta)-\mu_j(\theta,P)| > \frac{\lambda}{M_j}\right\}\right] \tag{B.67}
$$

$$
\geq E_P\left[\left(\frac{m_j(W_i,\theta)-\mu_j(\theta,P)}{\sigma_j(\theta,P)}\right)^2 \mathbb{1}\left\{\left|\frac{m_j(W_i,\theta)-\mu_j(\theta,P)}{\sigma_j(\theta,P)}\right| > \lambda\right\}\right] \geq 0. \tag{B.68}
$$

As $W_i = (t_i, r_i) \in \{0,1\}^2$, and $\mathcal{S}$ is a compact set such that $\forall (s_1, s_0) \in \mathcal{S} : s_1 > 1 - s_0$, $|m_j(W_i, \theta)| \leq B_j(\theta) \leq B_j^* < \infty$ for each j, where $B_j^* = \max_{\theta \in [0,1]^2 \times \mathcal{S}} B_j(\theta)$. That implies that $|\mu_j(\theta, P)| \leq B_j^* < \infty$, and $(m_j(W_i, \theta) - \mu_j(\theta, P))^2 \leq 4B^{*2}_j$. Consequently:

$$4M_j^2 B^{*2}{}_j \mathbb{1}\left\{2B^*{}_j > \frac{\lambda}{M_j}\right\} \tag{B.69}$$

$$\geq E_P\left[\left(\frac{m_j(W_i, \theta) - \mu_j(\theta, P)}{\sigma_j(\theta, P)}\right)^2 \mathbb{1}\left\{\left|\frac{m_j(W_i, \theta) - \mu_j(\theta, P)}{\sigma_j(\theta, P)}\right| > \lambda\right\}\right] \geq 0. \tag{B.70}$$

Finally, since neither B_j^* nor M_j depend on P or θ, it follows that (B.66) holds, concluding the proof.

□

APPENDIX C

Appendix to Chapter 3

Section C.1. Ambiguous Pre-Test Probability

Throughout the main analysis, we focus on the case where the clinician assesses a unique pre-test probability π. In this section, we extend our main identification result to cases, where such an accurate assessment is not possible, so that they are only able to assess that the pre-test probability lies within a specified set Π.

With a set of pre-test probabilities and a generic identified st Θ for the test accuracy, we can readily extend the definitions of identified post-test probabilities from Equation 3.6 to

$$V_j(\Theta;\Pi) := \left\{ v_j(\theta;\pi) : \theta \in \Theta, \pi \in \Pi \right\} \text{ for } j = 0, 1. \tag{C.1}$$

Furthermore, the generalization of the notion of dilation is immediate, where we only impose some mathematical structure on Π to avoid unnecessary technical complications.

Definition C.1.1. Let Θ be the identified set for θ. The index test induces dilation for the set of pre-test probabilities Π if

$$\Pi \subsetneq V_1(\Theta;\Pi) \text{ and } \Pi \subsetneq V_0(\Theta;\Pi). \tag{C.2}$$

An index test is called *dilation-inducing* if it induces dilation for every closed interval $\Pi \subseteq (0,1)$.

With this generalization, Proposition 3.2.1 remains true if we also assume that Θ is compact, which is generally the case under standard assumptions.

Proposition C.1.1. *Let Θ be a compact and connected identified set for the performance measure of the index test. The index test is dilation-inducing if and only if there exist $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict.*

Proof. Following the arguments in the proof of Proposition 3.2.1, note that $\overline{v}_j(\pi) := \max_{\theta \in \Theta} v_j(\theta; \pi)$ and $\underline{v}_j(\pi) := \min_{\theta \in \Theta} v_j(\theta; \pi)$ are well-defined for $j = 0, 1$ and $v_j(\Theta, \pi) = \left[\underline{v}_j(\pi), \overline{v}_j(\pi)\right]$. Note that for $j = 0, 1$ $\overline{v}_j(\cdot)$ is a strictly increasing function. Similarly, for closed interval Π, we write $\Pi = [\underline{\pi}, \overline{\pi}]$.

For sufficiency, assume that there are $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 > 1$ and fix an arbitrary compact $\Pi \subseteq (0,1)$. Again, following the arguments in the proof of Proposition 3.2.1, we have $\{\pi\} \subsetneq \left[\underline{v}_j(\pi), \overline{v}_j(\pi)\right]$ for $j = 0, 1$. Taking the union across Π on both sides, we obtain that $\Pi \subseteq V_j(\Theta, \Pi)$ for $j = 0, 1$. That the inclusion remains strict follows from $\overline{v}_j(\cdot)$ being strictly increasing. The argument for $\theta_0 + \theta_1 < 1$ and $\theta'_0 + \theta'_1 \geq 1$ is symmetric.

For necessity, fix an arbitrary closed interval $\Pi \subseteq (0,1)$ and suppose that $\Pi \subsetneq V_1(\Theta; \Pi)$ and $\Pi \subsetneq V_0(\Theta; \Pi)$. Then, in particular, $\overline{\pi} \leq \overline{v}_i(\overline{\pi})$ and $\underline{\pi} \geq \min_{\pi \in \Pi} \underline{v}_i(\pi)$ with at least one of them strict. First, consider the case that $\overline{\pi} < \overline{v}_1(\overline{\pi})$. Thus, there exists $\theta \in \Theta$ with $\theta_0 + \theta_1 > 1$. Letting $\pi^* \in \arg\max_{\pi \in \Pi} \underline{v}_1(\pi)$, then by the second condition we know that

$\pi^* \geq \underline{\pi} \geq \underline{v}_i(\pi^*)$ and therefore there exists $\theta' \in \Theta$ such that $\theta'_0 + \theta'_1 \leq 1$. The other case is similar. □

Finally, with this result, the simplification of Theorem 3.2.1 remains true even if dilation is defined in the more general way of Definition C.1.1.

Section C.2. Nomograms

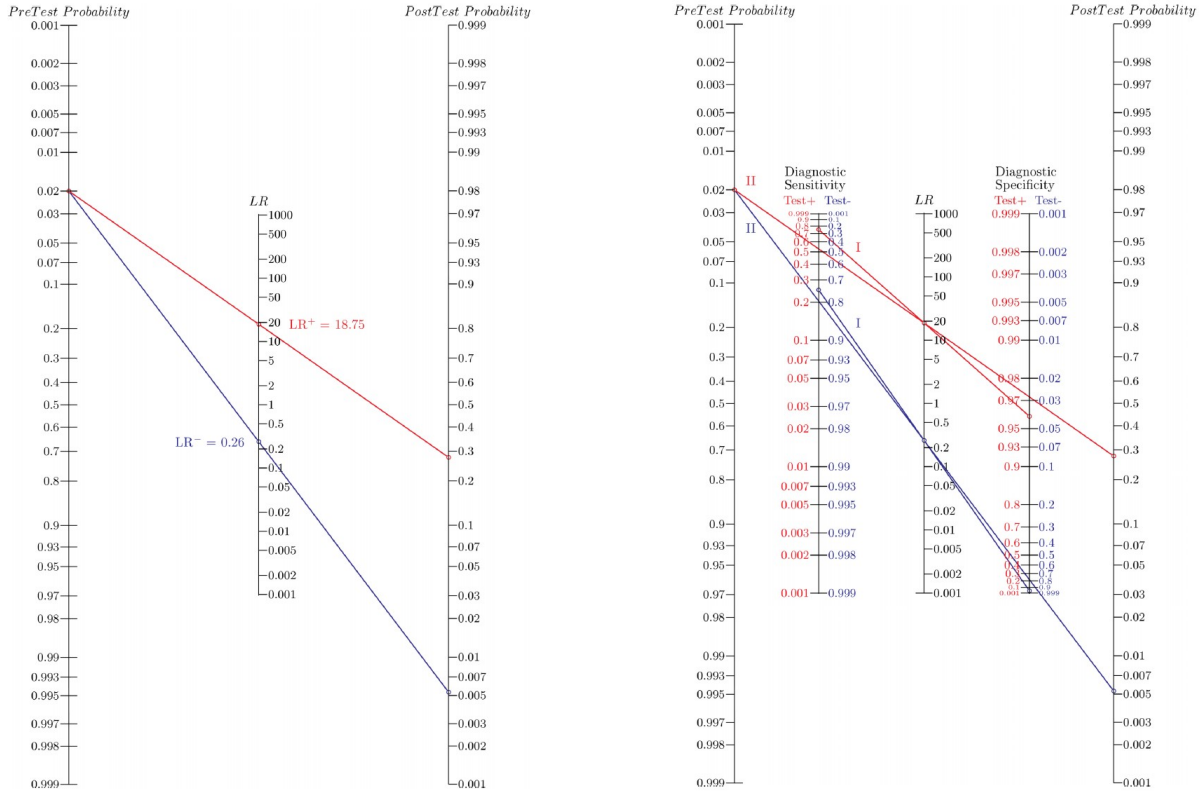
In practice, clinicians frequently use physical or online Fagan's nomograms, originally introduced by Fagan (1975), to more easily calculate the post-test probability at the bedside. The original nomogram takes π and the positive and negative likelihood ratios LR^+ and LR^- obtained from θ as an inputs. These are defined as:

$$LR^+ = \frac{\theta_1}{1 - \theta_0} \tag{C.3}$$

$$LR^- = \frac{1 - \theta_1}{\theta_0} \tag{C.4}$$

The clinician draws a straight line connecting their assessed π and the appropriate likelihood ratio to obtain the corresponding post-test probability. This is shown in Figure C.1a.

Since θ is more commonly reported than the likelihood ratios in diagnostic test performance studies, so-called two-step Fagan's nomograms were developed to facilitate post-test probability calculation without having to calculate the ratios from θ. The clinician first draws one line to connect appropriate θ_0 and θ_1 values which yields the corresponding likelihood ratio. Then a second line is drawn through that ratio to yield the corresponding post-test prrobability. This is shown in Figure C.1b.



(a) Fagan's nomogram.

(b) Two-step Fagan's nomogram.

Figure C.1. Fagan's and two-step Fagan's nomogram calculation example from Caraguel and Vanderstichel (2013).

An MRI screening test for breast cancer with $\theta = (0.96, 0.75)$ and $\pi = 0.02$ is represented. In Figure C.1a, LR^+ and LR^- are calculated and used as inputs. Red and blue lines then yield $PPP = 0.28$ and $NPP = 0.0006$, respectively. In Figure C.1b, likelihood ratios first follow graphically from θ. $LR^+ = 18.75$ (red line, I) and $LR^- = 0.26$ (blue line I). Then, $PPP = 0.28$ (red line, II) and $NPP = 0.0006$ (blue line, II) are found.

Subsection C.2.1. Nomograms and Ambiguous Parameters

Clinicians typically incorporate ambiguity about π and/or θ by finding appropriate bounds on the post-test probability. This may be done directly via the Bayes' rule as in Manski (2021). Equivalently, it may be done graphically, as shown via examples given by Srinivasan,

Westover, and Bianchi (2012) in Figure C.2. Panel A shows ambiguity in π only. Bounds on the post-test probability are formed by drawing straight lines through the appropriate likelihood ratio from the lowest and highest pre-test probability values. Panel B shows ambiguity in θ and hence likelihood ratios only. Here the bounds are formed by drawing lines from the single assessed π value through the boundary points of the bounds on the likelihood ratio. Panel C shows ambiguity in both, found analogously.

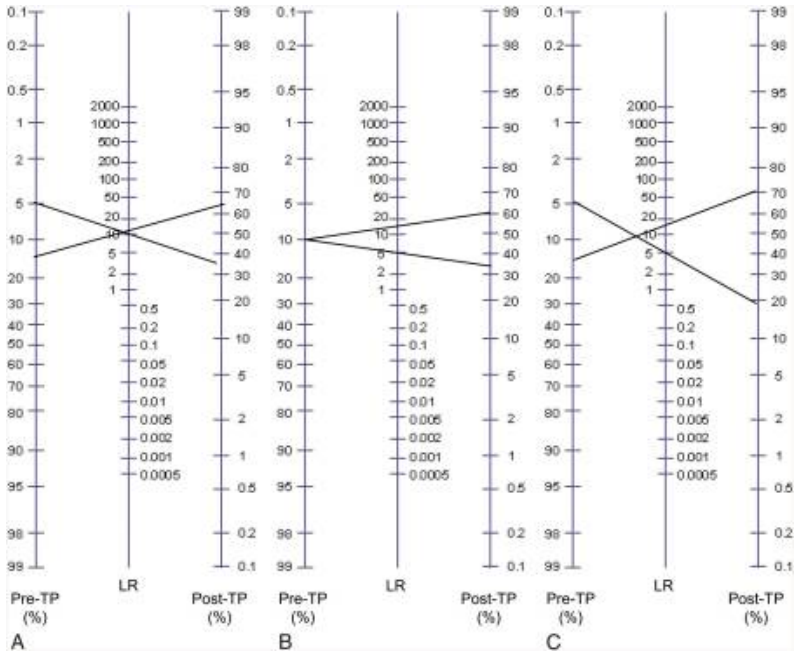


Figure C.2. Fagan's nomograms with ambiguous π and/or θ. Examples from Srinivasan, Westover, and Bianchi (2012).

The Post-TP axis shows the bounds on the post-test probabilities obtained via the nomogram. Panels A, B and C depict ambiguity in π only, θ only, and both π and θ, respectively.

Section C.3. A Test Based on Goodman (1965)

An alternative test for the parameters of a multinomial distribution can be formulated using simultaneous confidence intervals, following the approach of Goodman (1965). This test is useful for constructing confidence intervals that ensure coverage of the true parameters with at least asymptotic level α. The details of this test, which we implement in the simulations presented in Section 3.3.3, are outlined here.

$P(t, r)$ is a categorical distribution with four categories, which can be viewed as a multinomial distribution with a single draw. This allows us to apply the multinomial framework to construct confidence intervals that simultaneously cover the true parameters $P(t = j, r = k)$ for all $(j, k) \in 0, 1^2$ with at least asymptotic level α. The statistical imprecision in the estimates of θ arises only from the uncertainty in estimating $P(t = j, r = k)$.

Recall that from Theorem 3.2.1, we test whether there exists $\theta \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$. To do this, we form bounds on the sum $\theta_1 + \theta_0$ using equations (3.3), (3.4), and (3.5), which gives:

$$\theta_1 + \theta_0 \in \left[1 + \frac{\theta_1^L(s)P_s(y = 1) - P(t = 1)}{P_s(y = 0)}, 1 + \frac{\theta_1^U(s)P_s(y = 1) - P(t = 1)}{P(y = 0)}\right].$$

Now, if $\mathcal{C}_n$ denotes the (closed) confidence set for the parameters $P(t = j, r = k)$ for all $(j, k) \in 0, 1^2$ given an i.i.d sample of size n, we can derive the confidence interval $CI^n_{\theta_1+\theta_0}$

for $\theta_1 + \theta_0$ as

$$CI^n_{\theta_1+\theta_0} = \left[\min_{P(t=j,r=k)\in\mathcal{C}_n} 1 + \frac{\theta_1^L(s)P_s(y=1) - P(t=1)}{P(y=0)}, \right.$$
$$\left. \max_{P(t=j,r=k)\in\mathcal{C}_n} 1 + \frac{\theta_1^U(s)P_s(y=1) - P(t=1)}{P(y=0)} \right].$$

Thus, $CI^n_{\theta_1+\theta_0}$ provides a tool to test our null hypothesis, namely, that there exists a pair $(\theta_1, \theta_0) \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$, which is equivalent to determining whether the index test is dilation-inducing. Using the reasoning outlined in Molinari (2008, Section 2.3), we can show that $\lim_{n\to\infty} P(\theta_1 + \theta_0 \in CI^n_{\theta_1+\theta_0}) \geq 1 - \alpha$, ensuring asymptotic coverage.

Table 3.7 demonstrates that the test exhibits significantly lower power for any given sample size n and design compared to our preferred approach based on subvector inference. This is expected, given the known conservativeness of projection-based inference. Our simulations also indicate that the test based on Goodman (1965) is substantially less computationally demanding than the inference procedure based on Bugni, Canay, and Shi (2017) while achieving adequate coverage in the study. However, while the test based on Goodman (1965) achieves asymptotic coverage, our preferred inference procedure offers uniform asymptotic coverage and possible gains in power in return for the added resource burden.

Section C.4. Loan Riskiness Prediction

Our method can be applied more broadly to evaluate whether general classifiers may induce dilation due to imperfect references measuring their misclassification rates. Here, we apply our proposed statistical test to the context of loan approval decisions. Specifically, we

examine binary classification models used to assess the risk associated with loan applicants. The real-time binary classification model proposed by Abakarim, Lahby, and Attioui 2018 offers a suitable case study for evaluating whether such a machine learning model is informative or dilation-inducing.

In their framework, the machine learning algorithm classifies loan applications as either risky or not. In our framework, this classification serves as the index test, assigning $t = 1$ if the application is classified as "good risk," indicating that it should be approved. Conversely, a classification as "risky" corresponds to $t = 0$, which Abakarim, Lahby, and Attioui refer to as "bad risk." To evaluate their proposed algorithm, they use the commonly referenced German Credit dataset, a publicly available dataset that contains a binary classification of whether a credit application is considered "good" or"bad" (Hofmann 1994). While the dataset is based on actual historical accounts, it is known to contain errors. These errors could result in incorrect classifications within the data set (Groemping 2019).

For our purposes, this dataset can be treated as a reference test, where $r = 1$ indicates a "good" application predicting whether the borrower will not default $y = 1$. However, there are concerns whether r reveals the true y. Aforementioned data issues may lead to erroneous classifications. Moreover, even in the absence of data issues, it is unclear why r would be perfectly predictive of whether individuals will default or not. They may do so despite their previous credit history. Because of these imperfections, the data from Abakarim, Lahby, and Attioui 2018, reproduced in Table C.1 provides a valuable basis for applying our proposed test.

As explained, the assumption of an imperfect reference test seems reasonable in this application, but it remains unclear what level of accuracy should be assumed. Therefore,

Table C.1. Data from Abakarim, Lahby, and Attioui 2018.

	$r=0$	$r=1$	
$t=0$	203	43	246
$t=1$	97	657	754
	300	700	$n=1000$

$t=1$ and $r=1$ denote a "good risk" application according to the machine learning algorithm and the data set, respectively.

we proceed with two exploratory cases: (1) $s = (0.9, 0.9)$ and (2) $s = (0.95, 0.95)$. These values are chosen to reflect a relatively high but not perfect performance in the first case, and even greater accuracy in the second. Additionally, we depict the dilator set in Figure C.3 as a sensitivity analysis.

For the test with the null hypothesis that the machine learning algorithm is dilation-inducing, we proceed with a nominal significance level of $\alpha = 5\%$. In the first case, with $s = (0.9, 0.9)$, we obtain a test statistic of $T_n = 35.96$ and a critical value of $\hat{c}_n^{1-\alpha} = 5.25$, leading to a rejection of the null hypothesis. In the second case, where the reference test assumes higher accuracy ($s = (0.95, 0.95)$), we similarly reject the null hypothesis. Therefore, considering the aforementioned caveats regarding the assumptions about the reference test's accuracy, we conclude that, for these exploratory cases, the machine learning approach is informative in predicting loan riskiness in the sense of not inducing dilation.

Figure C.3 illustrates the resulting dilator set for this application. Specifically, $\mathcal{D}_P$ is represented by the shaded green area, meaning that if (and only if) the reference test's performance measure falls within this area. For instance, $s = (0.7, 0.8)$ would result in dilation, whereas $s = (0.9, 0.9)$ would not, as concluded previously.

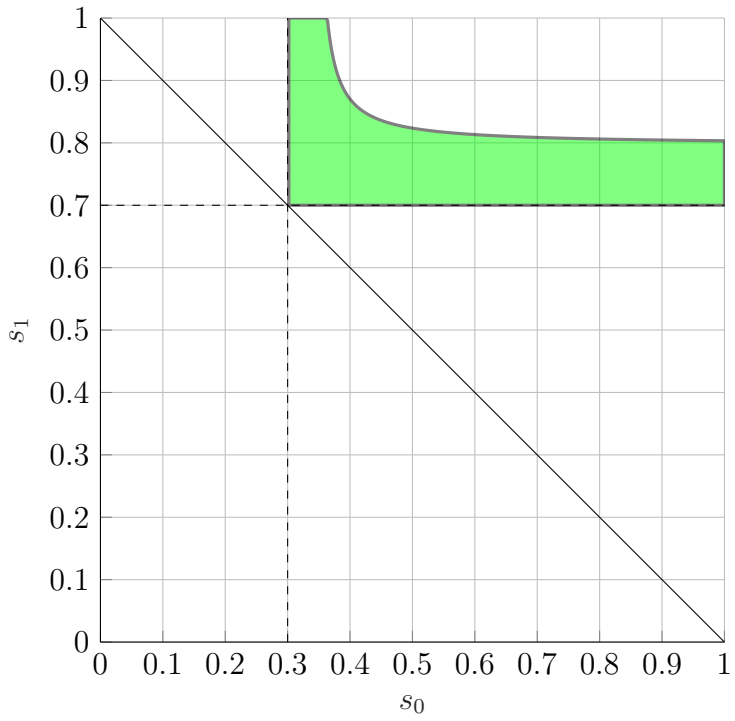


Figure C.3. Dilator set $\mathcal{D}_P$ for the algorithm from Abakarim, Lahby, and Attioui (2018).

The dashed lines correspond to $s_0 = P(r = 0)$ and $s_1 = P(r = 1)$.

Section C.5. Proofs for Section 3.2

Proof of Proposition 3.2.1. First we prove that t being dilation-inducing is equivalent to $\exists \theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict. We begin with a preliminary observation: $\pi \leq v_1(\theta; \pi)$ and $\pi \geq v_0(\theta; \pi)$ for any pre-test probability $\pi \in (0, 1)$ if and only if $\theta_0 + \theta_1 \geq 1$ holds. By a similar argument, all inequalities can be reversed, and the statement remains true. The same holds for all strict inequalities.

For sufficiency, assume that there are $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 > 1$ and fix an arbitrary pre-test probability $\pi \in (0, 1)$. By the preliminary observation and

the existence of θ and θ' we know that $v_1(\theta;\pi) \leq \pi < v_1(\theta';\pi)$ and $v_0(\theta';\pi) < \pi \leq v_0(\theta;\pi)$. Since $v_1(\cdot;\pi) : \Theta \to [0,1]$ and $v_0(\cdot;\pi) : \Theta \to [0,1]$ are both continuous in θ and Θ is a connected set, $V_1(\Theta;\pi)$ and $V_0(\Theta;\pi)$ are connected sets in $[0,1]$ and therefore non-trivial intervals. Thus, $\{\pi\} \subsetneq V_1(\Theta;\pi)$ and $\{\pi\} \subsetneq V_0(\Theta;\pi)$. The argument for $\theta_0 + \theta_1 < 1$ and $\theta'_0 + \theta'_1 \geq 1$ is symmetric.

For necessity, fix an arbitrary $\pi \in (0,1)$ and suppose that $\{\pi\} \subsetneq V_1(\Theta;\pi)$ and $\{\pi\} \subsetneq V_0(\Theta;\pi)$. Since $\{\pi\} \subsetneq V_1(\Theta;\pi)$, there must exist $\theta, \theta' \in \Theta$ such that $\pi \leq v_1(\theta;\pi)$ and $\pi \geq v_1(\theta';\pi)$ where at least one inequality is strict (and, also $\pi \geq v_0(\theta;\pi)$ and $\pi \leq v_0(\theta';\pi)$ where at least one inequality is strict; this follows either from $\{\pi\} \subsetneq V_0(\Theta;\pi)$, or the law of total probability). Therefore, t is dilation-inducing if and only if $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict.

Now we prove that this is equivalent to existence of $\theta, \theta' \in \Theta$ such that $\theta_1 + \theta_0 = 1$ and $\theta'_1 + \theta'_0 \neq 1$. For sufficiency, fix any $\theta, \theta' \in \Theta$ such that $\theta_1 + \theta_0 = 1$ and $\theta'_1 + \theta'_0 \neq 1$. Then either $\theta_1 + \theta_0 \leq 1$ and $\theta'_1 + \theta'_0 > 1$, or $\theta_1 + \theta_0 \geq 1$ and $\theta'_1 + \theta'_0 < 1$.

For necessity, fix any $\theta, \theta' \in \Theta$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 > 1$. Let $f : \Theta \to \mathbb{R}$ be a continuous function defined by $f(\theta) := \theta_1 + \theta_0 - 1$. Note that (i) $f(\Theta) \subseteq \mathbb{R}$ is an interval, because Θ is connected and continuity preserves connectedness; (ii) $f(\theta') \geq 0$, and (iii) $f(\theta) \leq 0$. Thus, there exists $\theta'' \in \Theta$ such that $f(\theta'') = 0$, i.e. $\theta''_1 + \theta''_0 = 1$, by the multidimensional intermediate value theorem (Munkres (2000, Theorem 24.3)). Since $\theta'_1 + \theta'_0 > 1$, $\theta'_1 + \theta'_0 \neq 1$. The argument for $\theta_0 + \theta_1 < 1$ and $\theta'_0 + \theta'_1 \geq 1$ is symmetric.

□

Proof of Lemma 3.2.1. Recall that under Assumptions 3.2.1 and 3.2.2, $P_s(y = 1) = \frac{P(r=1)+s_0-1}{s_1+s_0-1} \in (0,1)$.

By expanding, one can easily show $P(t = j, r = k) - P_s(r = k, y = l) = P_s(r = k, y = 1 - l) - P(t = 1 - j, r = k)$ for any $(j, k, l) \in \{0, 1\}^3$. Thus, the following additional expressions are true:

$$P(t = 1, r = 0) - s_0 P_s(y = 0) = (1 - s_1) P_s(y = 1) - P(t = 0, r = 0),$$

$$P(t = 1, r = 1) - (1 - s_0) P_s(y = 0) = s_1 P_s(y = 1) - P(t = 0, r = 1),$$

$$P(t = 1, r = 1) - s_1 P_s(y = 1) = (1 - s_0) P_s(y = 0) - P(t = 0, r = 1).$$

Using these expressions together with definitions in Equation 3.4 and Equation 3.5, it is immediate that $\theta_1^U(s) \geq \theta_1^L(s)$ so $[\theta_1^L(s), \theta_1^U(s)]$ is a proper interval and therefore non-empty.

For the second part, first note that if $s = (1, 1)$ then $\Theta_P(s)$ is a singleton, no matter whether Assumption 3.2.3 holds. Thus, it remains to show that if Assumption 3.2.3 holds, we also have that $\Theta_P(s)$ is not a singleton for $s \neq (1, 1)$. We will establish this by contraposition: Supposing there exists $s \in \mathcal{S} \setminus \{(1, 1)\}$ such that $|\Theta_P(s)| \leq 1$, we will show that then there exists $(j, k) \in \{0, 1\}^2$ such that $P(t = j, r = k) = 0$.

Since we already established non-emptiness, $|\Theta_P(s)| \leq 1$ is equivalent to $|\Theta_P(s)| = 1$ which holds if and only if $\theta_1^L(s) = \theta_1^U(s)$. To complete the proof, we show that $\theta_1^L(s) = \theta_1^U(s)$ implies that $P(t = j, r = k) = 0$ for some $(j, k) \in \{0, 1\}^2$. For this, there are 4 cases to consider in terms of $\theta_1^L(s)$.

We consider the first case in which $\theta_1^L(s) = 0$. Then $\theta_1^U(s) = 0$ only if $P(t = 1, r = 0) = 0$ and $P(t = 1, r = 1) = 0$, i.e. $P(t = 1) = 0$.

Consider next $\theta_1^L(s) = P(t=1, r=0) - s_0 P_s(y=0)$. Let first $P(t=1, r=0) \leq (1-s_1)P_s(y=1)$. Then for $\theta_1^U(s) = \theta_1^L(s)$ to hold, it would have to be that $\min\{P(t=1, r=1) + s_0 P_s(y=0), s_1 P_s(y=1) + s_0 P_s(y=0))\} = 0$ which is not possible by Assumption 3.2.2. Next suppose $P(t=1, r=0) > (1-s_1)P_s(y=1)$. Observe that $\theta_1^L(s) = (1-s_1)P_s(y=1) - P(t=0, r=0)$. It must be $-P(t=0, r=0) = \min\{P(t=1, r=1), s_1 P_s(y=1))\}$, implying that $P(t=0, r=0) = 0$.

Next, suppose $\theta_1^L(s) = P(t=1, r=1) - (1-s_0)P_s(y=0)$ Let $P(t=1, r=1) \leq s_1 P_s(y=1)$. Then $\theta_1^U(s) = \theta_1^L(s)$ only if $\min\{P(t=1, r=0), (1-s_1)P_s(y=1)\} = -(1-s_0)P_s(y=0)$ which contradicts Assumption 3.2.2. Next, notice $\theta_1^L(s) = s_1 P_s(y=1) - P(t=0, r=1)$ and let $P(t=1, r=1) > s_1 P_s(y=1)$. It must be that $-P(t=0, r=1) = \min\{P(t=1, r=0), (1-s_1)P_s(y=1)\}$ so $P(t=0, r=1) = 0$.

Finally, let $\theta_1^L(s) = P(t=1, r=0) + P(t=1, r=1) - P_s(y=0)$. Suppose $P(t=1, r=0) \leq (1-s_1)P_s(y=1)$. Then

$$\min\{P(t=1, r=1) + P_s(y=0), s_1 + (1-s_1)P_s(y=0)\} = P(t=1, r=1).$$

By the law of total probability and Assumption 3.2.2, we have $s_1 \geq P(r=1)$, and therefore $s_1 \geq P(t=1, r=1)$. Hence, the case contradicts Assumption 3.2.2. The ultimate case is when $P(t=1, r=0) > (1-s_1)P_s(y=1)$. We can rewrite $\theta_1^L(s) = P_s(y=1) - P(t=0, r=0) - P(t=0, r=1)$. It must be that $-P(t=0) = \min\{P(t=1, r=1) - s_1 P_s(y=1), 0\} = \min\{(1-s_0)P_s(y=0) - P(t=0, r=1), 0\}$. Hence, $\min\{(1-s_0)P_s(y=0) + P(t=0, r=0), P(t=0)\} = 0$ which implies that $P(t=0, r=0) = P(t=0, r=1) = 0$.

Therefore, if $|\Theta_P(s)| \leq 1$ there exists $(j,k) \in \{0,1\}^2$ such that $P(t=j, r=k) = 0$, concluding the proof.

□

PROOF OF THEOREM 3.2.1. By Proposition 3.2.1 and, in particular Corollary 3.2.1, we need to show that there exist $\theta, \theta' \in \Theta_P(s)$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict, if and only if $s \neq (1,1)$ and there exists $\theta'' \in \Theta_P(s)$ such that $\theta''_1 + \theta''_0 = 1$.

For necessity, first note that $s \neq (1,1)$ must hold, because if not then $\Theta(s)$ would be a singleton by Lemma 3.2.1, contradicting the existence of θ and θ', such that $\theta \neq \theta''$. Second, suppose there exist $\theta, \theta' \in \Theta_P(s)$ such that $\theta_1 + \theta_0 \geq 1$ and $\theta'_1 + \theta'_0 \leq 1$ where at least one inequality is strict. Again, by Lemma 3.2.1, $\Theta_P(s)$ is a non-singleton, non-empty set. Furthermore, $\Theta_P(s)$ is a line segment by definition. It is then immediate, that there exists $\theta'' \in \Theta : \theta''_1 + \theta''_0 = 1$.

For sufficiency, suppose that $s \neq (1,1)$ and that there exists $\theta'' \in \Theta_P(s)$ with $\theta''_1 + \theta''_0 = 1$. By Lemma 3.2.1 there exists $\theta \in \Theta_P(s)$ such that $\theta \neq \theta''$. $\Theta_P(s)$ is a line segment with positive and finite slope, so $\theta_1 + \theta_0 \neq 1$. Thus, there exists $\theta \in \Theta_P(s)$ such that either $\theta_1 + \theta_0 > 1$ or $\theta_1 + \theta_0 < 1$. Then setting $\theta' = \theta'' \in \Theta_P(s)$, we have $\theta'_1 + \theta'_0 = 1$ demonstrating sufficiency. □

Section C.6. Proofs for Section 3.3

PROOF OF LEMMA 3.3.1. Recall that $\mathcal{P}$ is a bounded subset of a finite dimensional Euclidean space. Let us denote the set of distributions considered under Assumption 3.2.3' with $\mathcal{P}'$, which is directly seen to be closed because of the weak inequalities. Let $\mathcal{P}''$

denote the set of distributions satisfying Assumption 3.2.2', which is also closed because of the weak inequalities. Therefore, both sets are compact. Now, we are interested in $\mathcal{P} = \mathcal{P}' \cap \mathcal{P}''$, which is compact as the intersection of two compact sets. □

Subsection C.6.1. Proof of Theorem 3.3.1

We will prove our Theorem 3.3.1, by means of an application of Bugni, Canay, and Shi (2017, Theorem 4.1). Thus, we need to verify their Assumption A.1–A.3 and that our space of considered distribution satifies $\mathcal{P}$ satifies their Definition 4.2. To do this and without further explicitly stating it, we assume throughout this section that (i) Assumption 3.3.1 holds for all $P \in \mathcal{P}$, (ii) Assumption 3.2.1 holds, (iii) $\mathcal{P}$ satisfies Assumption 3.2.2' (and *a fortiori* Assumption 3.2.2), and (iii) $\mathcal{P}$ satisfies Assumption 3.2.3' (and *a fortiori* Assumption 3.2.3).

The following Lemmata verify Bugni, Canay, and Shi (2017, Definition 4.2).

Lemma C.6.1. *For all $j = 1, \ldots, 7$, there exists $M_j \in (0, \infty)$ such that for all $(P, \theta) \in \mathcal{P} \times [0,1]^2$,*

$$\sigma^2_{P,j}(\theta; s) := \mathbb{V}_P[m_j(W_i, \theta; s)] \geq \frac{1}{M_j}$$

holds.

Proof. This is a special case of Claim 6 in Obradović (2024, p.29). □

Lemma C.6.2. *There exist $\underline{\sigma}, \overline{\sigma} \in (0, \infty)$ with $\underline{\sigma} \leq \overline{\sigma}$ such that for all $j = 1, \ldots, 7$ and all $(P, \theta) \in \mathcal{P} \times [0,1]^2$, we have $\sigma^2_{P,j}(\theta; s) \in [\underline{\sigma}^2, \overline{\sigma}^2]$.*

Proof. Consider any $j = 1, \ldots, 7$. The lower bound follows immediately from Lemma C.6.1 by setting $\underline{\sigma} = \min_j M_j^{-1}$. For the upper bound, note that for any $\theta \in [0,1]^2$, $m_j(\cdot, \theta; s)$ is bounded and hence $\mathbb{E}_P[m_j(W_i, \theta; s)] < \infty$. Then, $\big(m_j(\cdot, \theta; s) - \mathbb{E}_P[m_j(W_i, \theta; s)]\big)^2$ is bounded. Since P is a categorical distribution supported on $\{0,1\}^2$, the expression is also integrable, so $\mathbb{E}\big[m_j(\cdot, \theta; s) - \mathbb{E}_P[m_j(W_i, \theta; s)]\big]^2 < \infty$. Furthermore, because $\mathcal{P}$ is compact (Lemma 3.3.1), there exists a uniform upper bound which is also finite. □

Lemma C.6.3. *For all $j = 1, \ldots, 7$, $\left\{ \frac{m_j(\cdot, \theta; s)}{\sigma^2_{P,j}(\theta; s)} : \{0,1\}^2 \to \mathbb{R} \right\}$ is a measurable class of functions indexed by $\theta \in [0,1]^2$.*

Proof. The lemma follows directly from the definition of the m_j's together with Lemma C.6.2. □

Henceforth, define $\sigma_{P,j}(\theta; s) = \sqrt{\sigma^2_{P,j}(\theta; s)}$.

Lemma C.6.4. *There exists a constant $a > 0$ such that for all $j = 1, \ldots, 7$, we have*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\sup_{\theta \in [0,1]^2} \left| \frac{m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)} \right|^{2+a} \right] < \infty \tag{C.5}$$

Proof. We will prove the stronger statement that the inequality holds for any $a > 0$. For this, consider an arbitrary $j = 1, \ldots, 7$ and an arbitrary constant $a > 0$. Now, for any $P \in \mathcal{P}$ and $W \in \{0,1\}^2$, using Lemma C.6.1 we have

$$\sup_{\theta \in [0,1]^2} \left| \frac{m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)} \right|^{2+a} \leq \sup_{\theta \in [0,1]^2} |M_j m_j(W, \theta; s)|^{2+a},$$

where $M_j < \infty$ does not depend on P. Furthermore, since M_j is clearly continuous in θ, we can replace the sup with a max and therefore

$$\begin{aligned}\sup_{\theta\in[0,1]^2}\left|\frac{m_j(W,\theta;s)}{\sigma_{P,j}(\theta;s)}\right|^{2+a} &\leq \max_{(W',\theta,s)\in\{0,1\}^2\times[0,1]^2}\left|M_j m_j(W',\theta;s)\right|^{2+a},\\ &= \left|M_j m_j(W_j^*,\theta_j^*;s)\right|^{2+a} < \infty,\end{aligned}$$

where (W_j^*,θ_j^*) is an element of the arg max. Since the maximizer depends on j only, we conclude that

$$\begin{aligned}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\sup_{s\in[0,1]^2}\left|\frac{m_j(W,\theta;s)}{\sigma_{P,j}(\theta;s)}\right|^{2+a}\right] &\leq \sup_{P\in\mathcal{P}}\mathbb{E}_P\left[\left|M_j m_j(W_j^*,\theta_j^*;s)\right|^{2+a}\right]\\ &= \left|M_j m_j(W_j^*,\theta_j^*;s)\right|^{2+a} < \infty.\end{aligned}$$

□

For $i,j = 1,\dots 7$, $P\in\mathcal{P}$, and $\theta,\theta'\in[0,1]^2$ define

$$\begin{aligned}&\Omega_P(\theta,\theta')_{i,j} :=\\ &\mathbb{E}_P\left[\left(\frac{m_i(W,\theta;s)-\mathbb{E}_P[m_i(W,\theta;s)]}{\sigma_{P,i}(\theta;s)}\right)\left(\frac{m_j(W,\theta';s)-\mathbb{E}_P[m_j(W,\theta';s)]}{\sigma_{P,j}(\theta';s)}\right)\right]\end{aligned}$$

and let $\Omega_P(\theta,\theta')$ denote the 7×7 matrix with row $i = 1,\dots,7$ and column $j = 1,\dots,7$ given by $\Omega_P(\theta,\theta')_{i,j}$.

Lemma C.6.5.

$$\lim_{\delta\downarrow 0}\sup_{\|(\theta,\theta')-(t,t')\|<\delta}\sup_{P\in\mathcal{P}}\|\Omega_P(\theta,\theta')-\Omega_P(t,t')\| = 0$$

Proof. First note that for any given $\theta \in [0,1]^2$, $\sigma_{P,i}(\theta; s)$ continuous in P (recalling that $\mathcal{P}$ is endowed with the Euclidean topology). Then Lemma C.6.2 implies that

$$\left(\frac{m_i(\cdot,\theta,s) - \mathbb{E}_P[m_i(W,\theta,s)]}{\sigma_{P,i}(\theta,s)}\right)\left(\frac{m_j(\cdot,\theta',s') - \mathbb{E}_P[m_j(W,\theta',s')]}{\sigma_{P,j}(\theta',s')}\right)$$

is a bounded function for all $j = 1,\ldots,7$, all $\theta,\theta' \in [0,1]^2$, and all $P \in \mathcal{P}$. Thus, $\Omega_P(\theta,\theta')$ as a function of P is obtained from finitely many continuity-preserving operations on continuous functions and therefore continuous itself in P. Joint-continuity in (P,θ,θ') follows then directly from the definition. By Berge's theorem (which is applicable due to Lemma 3.3.1; see Aliprantis and Border 2006, Theorem 17.31)

$$D(\theta,\theta',t,t') := \sup_{P\in\mathcal{P}} \|\Omega_P(\theta,\theta') - \Omega_P(t,t')\|$$

is continuous. Then

$$\hat{D}(\delta) := \sup_{\|(\theta,\theta')-(t,t')\|\leq\delta} D(\theta,\theta',t,t')$$

is continuous too by Berge's theorem. Thus, $\lim_{\delta\downarrow 0}\hat{D}(\delta) = 0$. The conclusion follows from a squeeze argument, because

$$\hat{D}(\delta) \geq \sup_{\|(\theta,\theta')-(t,t')\|<\delta} D(\theta,\theta',t,t') \geq 0.$$

□

Next, we consider the following class of function index by $\theta \in [0,1]^2$:

$$\mathcal{F} = \left\{ v(\theta) = \big(v_j(\theta)\big)_{j=1}^7 : \{0,1\}^2 \to \mathbb{R}^7 \,\middle|\, v_j(\theta)(W) = \frac{m_j(W,\theta;s) - \mathbb{E}_P m_j(\cdot,\theta;s)}{\sigma_{P,j}(\theta;s)} \right\}$$

and for a given random sample $(W_i)_{i=1}^n$, $j = 1, \ldots, 7$, and $\theta \in [0,1]^2$ define

$$v_{n,j}(\theta) := \frac{1}{\sqrt{n}\,\sigma_{P,j}(\theta; s)} \sum_{i=1}^{n} \Big(m_j(W_i, \theta; s) - \mathbb{E}_P[m_j(\cdot, \theta; s)] \Big)$$

and $v_n(\theta) := (v_{n,j}(\theta))_{j=1}^7$ as the corresponding empirical process.

Furthermore, let ρ_P denote the coordinate-wise intrinsic variance semimetric given by

$$\rho_P(\theta, \theta') := \left\| \left(\sqrt{\mathbb{V}_P \left[\frac{m_j(\cdot, \theta; s)}{\sigma^2_{P,j}(\theta; s)} - \frac{m_j(\cdot, \theta'; s)}{\sigma^2_{P,j}(\theta'; s)} \right]} \right)_{j=1}^{7} \right\|.$$

Lemma C.6.6 (Donsker class)**.** *The class $\mathcal{F}$ is $\mathcal{P}$-uniform Donsker.*

Proof. For each $j = 1, \ldots, 7$, observe that $v_j(\theta)$ is a function of θ (for a given W) expressed as the ratio of a linear function in the numerator and the square root of a polynomial in the denominator. The denominator is strictly positive everywhere, as guaranteed by Lemma C.6.1 and Assumption 3.2.1. The function is defined on the compact set $[0,1]^2$, and therefore, it is Lipschitz continuous. This holds uniformly for all W, since $W \in \{0,1\}^2$, and for all $j = 1, \ldots, 7$. Furthermore, the Lipschitz constant can be chosen to hold uniformly in P, because of Lemma C.6.2 and Lemma C.6.4. Letting $K < \infty$ denote the corresponding uniform Lipschitz-constant, we trivially have $\mathbb{E}_P[K^r] = K^r < \infty$ for any $r \in \mathbb{R}$ and therefore, following the arguments in Vaart (2000, Example 19.7), we can get an upper bound on the bracketing integral that is independent of $P \in \mathcal{P}$, i.e. the bound holds $\mathcal{P}$-uniformly. The conclusion now follows from an application of Vaart (2000, Theorem 19.5). □

Lemma C.6.7 (Pre-Gaussian class)**.** *The class $\mathcal{F}$ is $\mathcal{P}$-uniform pre-Gaussian.*

Proof. First,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\sup_{\theta\in[0,1]^2} \|v(\theta)\| \right] < \infty,$$

holds because the LHS is bounded above by

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\sup_{\theta\in[0,1]^2} \left\| \left(\frac{m_j(W,\theta,s)}{\sigma_{P,j}(\theta,s)} \right)_{j=1}^{7} \right\| + \sup_{\theta\in[0,1]^2} \left\| \left(\frac{\mathbb{E}_P m_j(W,\theta,s)}{\sigma_{P,j}(\theta,s)} \right)_{j=1}^{7} \right\| \right] < \infty,$$

where finiteness follows from Lemma C.6.4.

Secondly,

$$\lim_{\delta\downarrow 0} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\sup_{\rho_P(\theta,\theta')<\delta} \|v(\theta) - v(\theta')\| \right] = 0,$$

holds, because ρ_P is a seminorm it gives rise to a convex constraint set, which is furthermore continuous in P, and therefore similar arguments as in the proof of Lemma C.6.5 show the required continuity properties. This proves the lemma. □

Lemma C.6.8. *The empirical process $v_n(\theta)$ is asymptotically ρ_P-equicontinuous uniformly in $P \in \mathcal{P}$. That is, for any $\varepsilon > 0$,*

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P^*\Big(\sup_{\rho_P(\theta,\theta')<\delta} \|v_n(\theta) - v_n(\theta')\| > \varepsilon \Big) = 0,$$

where P^* denotes the outer probability.

Proof. Note that the considered class $\mathcal{F}$ posses a $\mathcal{P}$-uniform, measurable, square integrable envelope, because all considered functions are uniformly bounded. Then Lemma C.6.6 and Lemma C.6.7 together are equivalent to the class being asymptotically

ρ_P-equicontinuous uniformly in $P \in \mathcal{P}$. (Van Der Vaart and Wellner 1997, Theorem 2.8.2) □

Combined all of the above, show that $\mathcal{P}$ satisfies the properties stated in Bugni, Canay, and Shi 2017, Definition 4.2. It remains to verify their Assumptions A.1–A.3.

We first note that their Assumption A.1 is automatically satisfied. Our GMS function φ as defined in Section 3.3, satisfies the needed properties as explained just after Equation (4.3) and Remark B.1 in Bugni, Canay, and Shi 2017.[1] Second, Assumption A.2 is not needed in our implementation because, as suggested by Bugni, Canay, and Shi (2017, Remark B.2), we adjusted the critical value by a small constant as mentioned in Footnote 13. Third, we verify their Assumption A.3, which requires the introduction of some further notation first.

$$\mathcal{P}_0 := \{P \in \mathcal{P} : \Theta_0 \cap \Theta_P(s) \neq \emptyset\}$$

$$Q_P(\theta; s) := \sum_{j=1}^{6} \left[\min\left\{0, \frac{\mathbb{E}_P m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)}\right\}\right]^2 + \left[\frac{\mathbb{E}_P m_7(W, \theta; s)}{\sigma_{P,7}(\theta; s)}\right]^2$$

$$g_{P,j}(\theta; s) := \frac{\mathbb{E}_P m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)}$$

$$\mathcal{P}_* := \{P \in \mathcal{P} : \Theta_P(s) \neq \emptyset\}$$

Assumption A.3. *The following conditions hold.*

1. More formally, this follows from Bugni, Canay, and Shi (2015, Lemma D.9). See also Remark B.1 *ibidem*. Note that Bugni, Canay, and Shi (2017, Remark B.1) incorrectly refers to Lemma D.*8* of Bugni, Canay, and Shi (2015).

(1) For all $P \in \mathcal{P}_0$ and all $\theta \in \Theta_0$,

$$Q_P(\theta; s) \geq c \min \left\{ \delta^2, \inf_{\tilde{\theta} \in \Theta_0 \cap \Theta_P(s)} ||\theta - \tilde{\theta}||^2 \right\}$$

for some constants c,$\delta > 0$.

(2) Θ_0 is convex.

(3) The functions $g_{P,i}$ are differentiable in θ for any $P \in \mathcal{P}_*$ and the class of functions $\left\{ (\nabla g_{P,j})_{j=1}^7 \mid P \in \mathcal{P}_* \right\}$ is equicontinuous, that is:

$$\lim_{\delta \longrightarrow 0} \sup_{P \in \mathcal{P}_*, (\theta, \theta') : ||\theta - \theta'|| \leq \delta} ||(\nabla g_{P,j})_{j=1}^7(\theta; s) - (\nabla g_{P,j})_{j=1}^7(\theta'; s)|| = 0.$$

Note that (2) in Assumption A.3 holds trivially in our case. We will verify the other two conditions formally in the next two lemmata next.

Lemma C.6.9. *Assumption A.3(1) holds.*

Proof. First, note that for all θ such that $\theta_0 + \theta_1 = 1$, we have $\mathbb{E}_P[m_7(W, \theta; s)] = P(t = 1) - \theta_1$. Second, note that for all $P \in \mathcal{P}_0$, the intersection $\Theta_0 \cap \Theta_P(s)$ consists of a single point, i.e., $\Theta_0 \cap \Theta_P(s) = \{\theta^*\}$, where $\theta^* = (P(t = 0), P(t = 1))$. This is because $\mathbb{E}_P[m_7(W, \theta; s)] = 0$ must hold together with $\theta_0 + \theta_1 = 1$.

Next, observe that for all $\theta \in [0, 1]^2$, we have

$$Q_P(\theta; s) \geq \left[\frac{\mathbb{E}_P[m_7(W, \theta; s)]}{\sigma_{P,7}(\theta; s)} \right]^2.$$

Therefore, it suffices to prove that

$$\left[\frac{\mathbb{E}_P[m_7(W, \theta; s)]}{\sigma_{P,7}(\theta; s)} \right]^2 \geq c \, \|\theta - \theta^*\|^2,$$

for some constant $c > 0$ and all $\theta \in \Theta_0$.

By Lemma C.6.2, we have $\sigma_{P,7}(\theta; s) \leq \overline{\sigma}$ for some positive constant $\overline{\sigma}$. Since $\theta_0 + \theta_1 = 1$ and $P(t = 0) + P(t = 1) = 1$, we get the squared Euclidean distance between θ and θ^* as

$$\|\theta - \theta^*\|^2 = (\theta_0 - P(t = 0))^2 + (\theta_1 - P(t = 1))^2 = 2(P(t = 1) - \theta_1)^2.$$

Combining these results, we get

$$\left[\frac{\mathbb{E}_P[m_7(W, \theta; s)]}{\sigma_{P,7}(\theta; s)}\right]^2 \geq \left(\frac{P(t = 1) - \theta_1}{\overline{\sigma}}\right)^2 = \frac{1}{2\overline{\sigma}^2}\|\theta - \theta^*\|^2.$$

Thus, setting $c = \dfrac{1}{2\overline{\sigma}^2} > 0$, we have

$$Q_P(\theta; s) \geq c\,\|\theta - \theta^*\|^2.$$

Finally, note that we can take $\delta > 0$ arbitrarily to get

$$Q_P(\theta; s) \geq c \min\left\{\delta^2, \inf_{\theta^* \in \Theta_0 \cap \Theta_P(s)} \|\theta - \theta^*\|^2\right\},$$

holding for all $\theta \in \Theta_0$. □

Lemma C.6.10. *Assumption A.3(3) holds.*

Proof. By the same argument as in the proof of Lemma C.6.6 all $g_{P,j}$ treated as functions of (θ, P) are smooth and locally Lipschitz, which carries over to their derivatives too. As functions of θ these derivatives are defined on compact sets, $[0, 1]^2$, and therefore they are (globally) Lipschitz. Let $K_{P,j}$ denote the Lipschitz constant for a given $P \in \mathcal{P}$ and $j = 1, \ldots, 7$. Now define $K = \max_{P \in \mathcal{P}, j=1,\ldots,7} K_{P,j}$, which is well-defined and finite

because of Lemma 3.3.1 and the smoothness property mentioned before. Now, K is a uniformly valid Lipschitz constant for the whole class $\{(\nabla g_{P,j})_{j=1}^{7} : P \in \mathcal{P}\}$ and therefore the class is equicontinuous. □

Combining all the results obtained in this section allows us to invoke Bugni, Canay, and Shi (2017, Theorem 4.1), which then proves our Theorem 3.3.1.

Section C.7. Proofs for Section 3.5

Lemma C.7.1 (Hiriart-Urruty 1985)**.** *The image of a connected space under a continous correspondence with non-empty and connected values is connected.*

Proof of Lemma 3.5.1. Using the notation and the arguments just before the statement of Lemma 3.5.1 in the main text, $\Theta_P(\cdot)$ is a correspondence mapping a connected set (*cf.* Assumption 3.2.1S) into the unit square (with the usual topology). Furthermore, since $\Theta_P(s)$ is non-empty (*cf.*, Lemma 3.2.1) and a line-segment for every $s \in \mathcal{S}$ (*cf.*, Equation 3.3), the correspondence has connected values. Note that this correspondence is the convex hull of two functions: namely, the boundaries of the line segments ($\theta^L(\cdot)$ and $\theta^H(\cdot)$). These two functions are continuous by definition, and therefore the correspondence is continous too. (Aliprantis and Border 2006, Theorem 17.37(1))

Finally, $\Theta_P(\mathcal{S})$ being connected follows from an application of Lemma C.7.1. □

Proof of Lemma 3.5.2. If $\mathcal{S} = \{(1,1)\}$, then $\Theta_P(\mathcal{S})$ is a singleton by Lemma 3.2.1. If $\mathcal{S} \neq \{(1,1)\}$, then there exists $s \in \mathcal{S}$ such that $s \neq (1,1)$ and, again by Lemma 3.2.1, $\Theta_P(s)$ is not a singleton. Then, $\Theta_P(s) \subseteq \Theta_P(\mathcal{S})$ and $\Theta_P(\mathcal{S})$ is not a singleton either. □

PROOF OF THEOREM 3.5.1. Start with necessity. By Proposition 3.2.1 and, in particular Corollary 3.5.1, we need to show that there exist $\theta, \theta' \in \Theta_P(\mathcal{S})$ such that $\theta_0 + \theta_1 \leq 1$ and $\theta'_0 + \theta'_1 \geq 1$, where at least one inequality is strict, if and only if $\mathcal{S} \neq \{(1,1)\}$ and there exists $\theta'' \in \Theta_P(\mathcal{S})$ such that $\theta''_1 + \theta''_0 = 1$. Suppose there exist $\theta, \theta' \in \Theta_P(\mathcal{S})$ such that $\theta_1 + \theta_0 \geq 1$ and $\theta'_1 + \theta'_0 \leq 1$ where at least one inequality is strict. By Lemma 3.5.1, $\Theta_P(\mathcal{S})$ is a connected set. Now, let $f : \Theta \to \mathbb{R}$ be a continous function given by $f(\theta) = \theta_0 + \theta_1 - 1$ and note that (i) $f(\Theta) \subseteq \mathbb{R}$ is an interval, because Θ is connected, (ii) $f(\theta) \geq 0$, and (iii) $f(\theta) \leq 0$. Thus, there exists $\theta'' \in \Theta$ such that $f(\theta'') = 0$, i.e. $\theta''_1 + \theta''_0 = 1$, by the multidimensional intermediate value theorem (Munkres (2000, Theorem 24.3)). Furthermore, $\mathcal{S} \neq \{(1,1)\}$ holds, because if not[2] $\Theta_P(\mathcal{S})$ would be a singleton (*cf.*, Lemma 3.2.1), contradicting the existence of θ and θ' as they need to be different.

For sufficiency, suppose that $\mathcal{S} \neq \{(1,1)\}$ and there exists $\theta'' \in \Theta_P(\mathcal{S})$ with $\theta''_1 + \theta''_0 = 1$. Fix $s \in \mathcal{S}$ such that $\theta'' \in \Theta_P(s)$ and consider two cases:

(1) If $s \neq (1,1)$, then by Theorem 3.2.1 t is dilation-inducing.

(2) If $s = (1,1)$, then by hypothesis, there exists $s' \in \mathcal{S}$ with $s' \neq s$ and then Lemma 3.2.1 ensures that $\Theta_P(s')$ must contain at least two points. Since $\Theta_P(s')$ is a line segment with positive and finite slope, there must exist $\theta \in \Theta_P(s')$ such that $\theta_1 + \theta_0 \neq 1$. Now set $\theta' = \theta''$ and apply Proposition 3.2.1.

□

PROOF OF PROPOSITION 3.5.1. In the following (1), (2), (3), and (4) refer to the inequalities indicated by the same numbers in the definition of $\mathcal{D}_P^{\llcorner}$.

2. Note that $\mathcal{S}$ is non-empty by Assumption 3.2.1S.

First consider $s \in \mathcal{D}_P$. If $s_0 + s_1 = 1$, we have $s \in S_0$ and we are done. Thus, consider $s_0 + s_1 > 1$ and we will show that all the moment inequalities of $\mathcal{D}_P^{\llcorner}$ are satisfied. First, note that $P_s(y = 1) = \frac{P(r=1)+s_0-1}{s_1+s_0-1} \in [0, 1]$ holds if and only if (3) and (4) hold. To see this note that the lower bound is equivalent to $s_0 \geq 1 - P(r = 1)$, which is (4), and the upper bound is $s_1 \geq P(r = 1)$, which is (3). Second, $\Theta_P(s) \neq \emptyset$ holds if and only if $P_s(y = 1) \in [0, 1]$. To see this, note that if $P_s(y = 1) \notin [0, 1]$ then $\Theta_P(s) = \emptyset$ by convention. Conversely, $P_s(y = 1) \in (0, 1)$ makes the first part of Lemma 3.2.1 applicable giving $\Theta_P(s) \neq \emptyset$. If $P_s(y = 1) \in \{0, 1\}$ then $\Theta_P(s) = [0, 1]^2$ by definition. Then note that $\Theta_P(s) \cap \Theta_0 \neq \emptyset$ means that if $P_s(y = 1) \in (0, 1)$, then Proposition 3.3.1 is applicable and therefore $\mathbb{E}_P[m_j(\cdot, \theta, s)] \geq 0$ for all $j = 1, \ldots, 6$ and $\mathbb{E}_P[m_7(\cdot, \theta, s)] = 0$ holds for $\theta_0 + \theta_1 = 1$. The latter then gives $\theta = \big(P(t = 0), P(t = 1)\big)$. With this now note that $\mathbb{E}_P[m_1(\cdot, \theta, s)] \geq 0$ is equivalent to (1) and $\mathbb{E}_P[m_6(\cdot, \theta, s)] \geq 0$ is equivalent to (2). If $P_s(y = 1) = 0$, i.e. $P(r = 1) = 1 - s_0$, (4) holds with equality. Since $s_1 > 1 - s_0$, (3) holds too. (1) and (2) are, in this case, equivalent to $P(t = 0, r = 1) \geq 0$ and $P(t = 1, r = 1) \geq 0$, respectively, which hold trivially. If $P_s(y = 1) = 1$, i.e. $P(r = 1) = s_1$, (3) holds with equality and since $s_0 > 1 - s_1$, (4) holds too. (1) is, in this case, equivalent to $P(t = 1) - P(r = 1) \geq -P(t = 0, r = 1)$, which is the same as $P(t = 1) \geq P(r = 1) - P(t = 0, r = 1) = P(t = 1, r = 1)$ making it trivially true. Similarly, (2) is in this case equivalent to $P(t = 0) - P(r = 1) \geq -P(t = 1, r = 1)$ which is the same as $P(t = 0) \geq P(r = 1) - P(t = 1, r = 1) = P(t = 0, r = 1)$ showing that the inequality holds trivially.

For the other inclusion, if $s \in S_0$ or s is such that $P_s(y = 1) \in \{0, 1\}$, we are done because $\Theta_P(s) = [0, 1]^2$ by definition. If $s \in \mathcal{D}_P^{\llcorner} \setminus S_0$ such that $P_s(y = 1) \notin \{0, 1\}$, then

by the same argument above, $\Theta_P(s) \neq \emptyset$ and $P_s(y = 1) \in (0, 1)$. We will prove that $\theta = \big(P(t = 0), P(t = 1)\big) \in \Theta_P(s) \cap \Theta_0$, establishing that $s \in \mathcal{D}_P$. Trivially, $\theta \in \Theta_0$. To show that $\theta \in \Theta_P(s)$, we will show that $\mathbb{E}_P[m_j(\cdot, \theta, s)] \geq 0$ for all $j = 1, \ldots, 6$ and $\mathbb{E}_P[m_7(\cdot, \theta, s)] = 0$ hold (see Proposition 3.3.1). (1) and (2) are equivalent to the inequities with $j = 1$ and $j = 6$. The equality for m_7 holds because $\theta_1 = 1 - \theta_0$. The remaining four inequalities will be established next:

(1) (m_2 is implied by (3) and (4)) We want to show that

$$(1 - \theta_1 - s_1)\frac{P(r = 1) - 1 + s_0}{s_1 - 1 + s_0} \leq P(t = 0, r = 0),$$

which is true if $P(t = 0, r = 0) \geq 1 - \theta_1 - s_1$, because $\frac{P(r=1)-1+s_0}{s_1-1+s_0} = P_s(y = 1) \in [0, 1]$. To establish this inequality, recall that $P(t = 0) = 1 - P(t = 1) = 1 - \theta_1$ here and then

$$\begin{aligned}
& s_1 \geq P(r = 1) && \text{(by (3))} \\
\implies\ & s_1 \geq P(t = 0, r = 1) \\
\iff\ & P(t = 0) \geq P(t = 0, r = 1) - s_1 + (1 - \theta_1) && (\pm P(t = 0)) \\
\iff\ & P(t = 0, r = 0) \geq 1 - \theta_1 - s_1.
\end{aligned}$$

(2) (m_3 is essentially equivalent to (3)) We need to show the following inequality:

$$(1 - \theta_1)\frac{P(r = 1) - 1 + s_0}{s_1 - 1 + s_0} \leq P(t = 0),$$

which holdtrivially if $P(t = 0) = 0$, because $1 - \theta_1 = \theta_0 = P(t = 0)$. If $P(t = 0) = 1 - \theta_1 \neq 0$, the the desired inequality holds if and only if

$$\frac{P(r = 1) - 1 + s_0}{s_1 - 1 + s_0} \leq 1,$$

which holds if and only if $P(r = 1) \leq s_1$, which is m_7.

(3) (m_4 is essentially equivalent to (3)) We need to establish the following inequality:

$$\theta_1 \frac{P(r = 1) - 1 + s_0}{s_1 - 1 + s_0} \leq P(t = 1),$$

which holds trivially if $P(t = 1) = 0$, because $P(t = 1) = \theta_1$. If $P(t = 1) = \theta_1 \neq 0$ then the inequality holds holds if and only if

$$\frac{P(r = 1) - 1 + s_0}{s_1 - 1 + s_0} \leq 1,$$

which holds if and only if $P(r = 1) \leq s_1$, which is (3).

(4) (m_5 is implied by (3) and (4)) We want to show that

$$(\theta_1 - s_1) \frac{P(r = 1) - 1 + s_0}{s_1 - 1 + s_0} \leq P(t = 1, r = 0),$$

which is true if $P(t = 1, r = 0) \geq \theta_1 - s_1$, because $\frac{P(r=1)-1+s_0}{s_1-1+s_0} = P_s(y = 1) \in [0, 1]$ by (3) and (4). To establish this inequality, recall that $P(t = 1) = \theta_1$ here and

then

$$
\begin{aligned}
& s_1 \geq P(r=1) && \text{(by (3))} \\
\implies\ & s_1 \geq P(t=1, r=1) && \\
\iff\ & P(t=1) \geq P(t=1, r=1) - s_1 + \theta_1 && (\pm P(t=1)) \\
\iff\ & P(t=1, r=0) \geq \theta_1 - s_1. &&
\end{aligned}
$$

□