Binary Classifiers as Dilations[*]

Filip Obradović[†] Gabriel Ziegler[‡]

November 7, 2024

Abstract

Seidenfeld and Wasserman (1993) define the phenomenon of *dilation*. When a dilation occurs, any additional information *increases* the *uncertainty* about the true state of the world. In this paper, we show that dilation may manifest in real-world scenarios when information is provided by binary classifiers, such as diagnostic tests and predictive algorithms. This can happen when classifier performance measures are partially identified due to an imperfect reference classifier, which are ubiquitous in practice. We characterize when a dilation occurs and develop corresponding inference procedures based on methods for subvector inference in moment inequality models. We apply the approach to diagnostic procedures for COVID-19 detection, using CT chest scans evaluated by radiologists and AI algorithms. We cannot reject the hypothesis that the radiologists' assessments exhibit a dilation, thus showcasing a potential real-world instance of a dilation. We additionally illustrate the broader applicability of our methodology by rejecting the hypothesis that data-mining techniques for predicting the riskiness of credit card applications are non-informative in the sense of a dilation.

Keywords: Ambiguity, partial identification, dilation, binary classifier, diagnostic tests.
JEL Classification: C14, C38, D83, D90, I12, I18.

[*] This paper supersedes Ziegler (2021). We are thankful for comments and questions from Eric Mbakop and audience members at the Pennsylvania Theory Conference 2023, the Durham York Theory Workshop 2023, the Theory Day 2023 at the University of Edinburgh, the NOeG Winter Workshop 2023, and seminar participants at the University of Graz, NYU Shanghai, Royal Holloway University of London, University of Surrey, and University of Udine. Parts of the research have been conducted while Ziegler was visiting the University of Pittsburgh and NYU Shanghai, and he expresses his gratitude for their hospitality. Obradović gratefully acknowledges financial support from the Robert Eisner Memorial Fellowship. All errors remain ours.

[†] Northwestern University; obradovicfilip@u.northwestern.edu.

[‡] The University of Edinburgh; ziegler@ed.ac.uk.

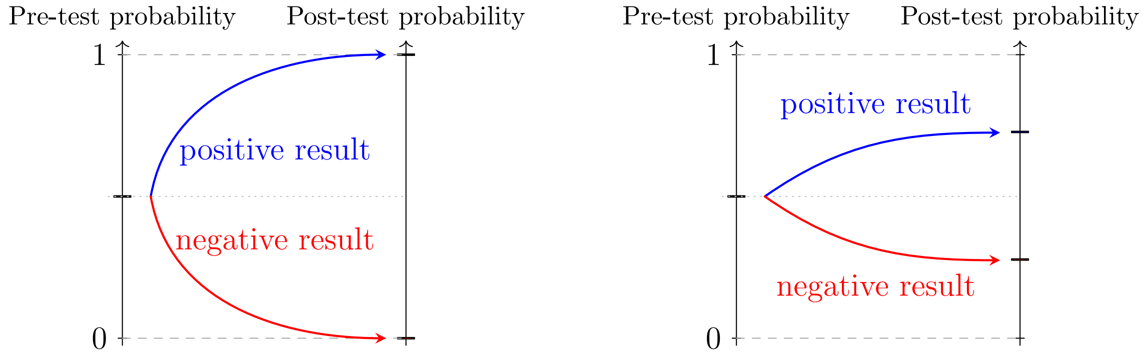
1 Introduction

In economic theory, additional information is typically viewed as valuable—more data should (weakly) help learn the true state of the world. However, this tenet may break down when the additional information is ambiguous. In such cases, a counterintuitive phenomenon called *dilation*, originally formalized by Seidenfeld and Wasserman (1993), can arise. When a dilation occurs, any additional information may only make it more *difficult* to learn the true state of the world. Dilations have traditionally been viewed as an abstract theoretical concept, and their real-world implications have remained elusive.

In this paper, we show that dilation may arise in practice when information is provided by binary classifiers, such as diagnostic tests and predictive models. To do so, we equivalently characterize dilation in common empirical settings involving binary classifiers. Using the characterizations, we develop a statistical test for dilation using methods for subvector inference in moment inequality models. The test indicates that computed tomography (CT) chest scans for detecting COVID-19 infection exhibited dilation in the early stages of the pandemic, when evaluated by radiologists.

To illustrate the phenomenon of dilation, consider a clinician diagnosing a patient based on the results of a diagnostic test. Watson et al. (2020) explain that the doctor first forms a *pre-test* probability of the patient having the disease based on heuristics and expert knowledge. They then observe a test result and form the corresponding *post-test* probability. Ideally, the test is perfect, and the post-test probability is 0 in the case of a negative result or 1 in the case of a positive result. This is depicted in Figure 1a. In practice, the test is almost always imperfect, and is often assumed to have precisely measured false positive and false negative rates. Then the post-test probabilities may not be 0 or 1, but may still be informative, as in Figure 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.

Uniqueness is lost if the misclassification rates of the diagnostic test are ambiguous. Then we say that the test provides *ambiguous* information. In this case, even a unique pre-test probability will result in a set of post-test probabilities for either test result. Thus, the doctor will face greater uncertainty than before, regardless of the test result. Despite this, the test remains informative of the disease status if all post-test probabilities



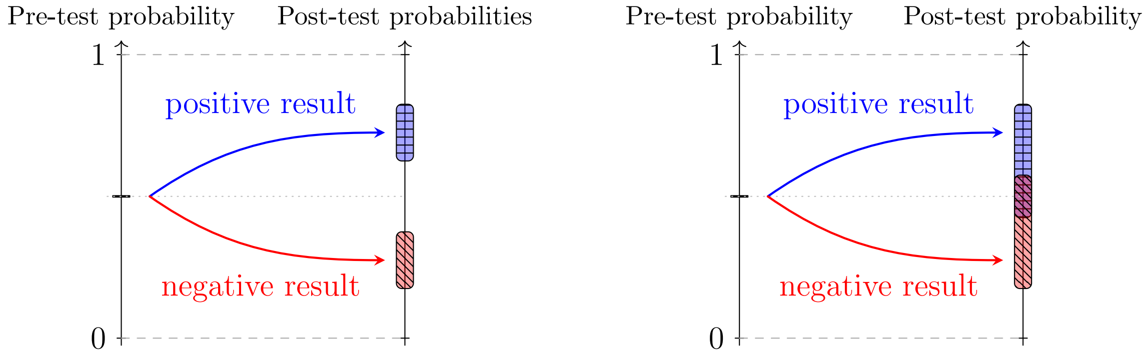
(a) Perfect information

(b) Imperfect, but unambiguous information

Figure 1: Unambiguous information

are shifted with respect to the pre-test one. This is depicted in Figure 2a.

However, this increase in uncertainty may be so severe that the test ceases to be informative. Figure 2b exemplifies such a situation. Here, the pre-test probability is included in the set of post-test probabilities, *regardless of the test result.* We then say that a *dilation* occurs and the test becomes completely uninformative. Since all tests entail some costs, a regulatory body may prefer not to approve such tests.



(a) Ambiguous information

(b) A Dilation

Figure 2: Ambiguous information

Misclassification rates of a binary classifier may be evaluated relative to an imperfect reference. In this case, the rates become partially identified, and the information provided is ambiguous. This is ubiquitous in practice, especially in the context of diagnostic tests. In such cases, classifiers may appear precise with respect to the reference, while inducing a dilation. Since dilations make the classifier completely uninformative, it is desirable to determine whether they occur. To address this, we develop a framework for identifying a dilation relying on results from Obradović (2024). We then use it to develop a statistical test using recent methods for subvector inference in moment inequality models (Bugni et al. (2017)). Our approach yields a formal inference procedure that can be applied to various binary classifiers, including diagnostic tests; predictive models in the contexts of credit approval, fraud detection, or spam filtering; and remote sensing image classification, such as satellite imaging for land cover change.

We investigate dilations in two real-world applications. First, we analyze COVID-19 detection using CT chest scans evaluated by both radiologists and AI algorithms. In this case, we find evidence that the radiologists' assessments correspond to a dilation. Observing the assessment only introduces uncertainty about the true health status, without providing information about the ground truth. This example indicates that dilations may have important policy implications. Second, we apply our methodology to the domain of credit risk assessment, analyzing data-mining techniques used to predict the riskiness of credit card applications. Here, we reject the hypothesis that the predictive models are non-informative in the sense of a dilation, illustrating the broader applicability of our approach.

The remainder of the paper is structured as follows. Subsection 1.1 reviews the related literature. Section 2 introduces the setting, defines a dilation and provides identification results. Section 3 formalizes the statistical test for the occurrence of dilation. Section 4 discusses the applications. Section 5 explores extensions that relax the knowledge assumptions about the reference test and policy implications. All formal proofs are relegated to the appendix.

1.1 Related Literature

Measurement of binary classifier misclassification rates originates in studies of medical diagnostic test performance (Yerushalmy (1947), Binney et al. (2021)). In this context, the concept of *gold standard bias*—the discrepancy between observed and actual test misclassification rates—is well-known. Early research by Gart and Buck (1966), Staquet et al. (1981), and Zhou et al. (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 relationship between the tests affects the gold standard bias. Deneef (1987) found that if tests are conditionally independent, apparent performance underestimates true performance, while positive correlation can inflate apparent accuracy. Boyko et al. (1988) and Valenstein (1990) further investigated how this relationship changes with disease prevalence and the correlation of classification errors, respectively. The main focus of this work is the qualitative direction of the bias. Yet, the practical application of these findings is limited by the challenge of measuring the correlation between test results, as it depends on unobservable factors. Notable exceptions include Thibodeau (1981) and later Emerson et al. (2018). Our analyses directly builds on Obradović (2024) who derives the sharp joint identified set for the true misclassification rates.

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.[1] 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 et al. (1997) systematically analyzed dilations. Our identification results are based on the definitions in the latter two papers. A substantial corpus of theoretical literature on dilations, which is too extensive to comprehensively summarize here, followed this work. Interested readers are encouraged to explore recent contributions by Bradley (2019, in particular, Section 3.1) and Gong and Meng (2021), along with the references therein. Implicitly in many of these approaches, 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 Dominiak et al. (2022) and Lin and Payró (2024). Moving beyond theoretical work, Shishkin and Ortoleva (2023) conducted experimental studies on how individuals value dilations. It is noteworthy that there has been a recent surge of interest in studying ambiguous information in experimental economics, exemplified by Epstein and Halevy (2024), Kellner et al. (2022), Kops and Pasichnichenko (2023), and Liang (2024). Manski (2018) mentions the possibility of a dilation in concrete questions about personalized patient care. In contrast to prior research, to our best knowledge,

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

our study is the first to conduct a statistical analysis of dilations in a real-world context. In our contribution, we propose a statistical test to detect the presence of a dilation. However, we do not take a stance on the decision-making processes individuals employ when encountering a dilation.

Our proposed method of statical inference is based on subvector inference in moment inequality models, as introduced by Bugni et al. (2017). Thus, we contribute to the recent developments exploring issues of partial identification in medical and epidemiological settings such as Bhattacharya et al. (2012), Manski (2020), Toulis (2021), Manski (2021), Stoye (2022), Sacks et al. (2022), and Obradović (2024)

2 Identification of a Dilation

In this section, we first present the setting and expound on the identification of test performance in practically relevant settings. We then define dilation in the context of diagnostic tests, and derive an equivalence result which provides tractable necessary and sufficient conditions for a test to be a dilation. The result forms a basis for the statistical test we propose in Section 3.

2.1 Test Performance Identification

We are concerned with evaluating whether a novel classifier t induces a dilation. As we will show, this is closely associated with performance measures of t, or its misclassification rates. We rely on standard terminology related to binary classifier performance measurement, which originates in the context of diagnostic tests. We will thus refer to t as the *index test*, highlighting that it may pertain to any binary classifier, not necessarily a diagnostic test. Henceforth, any classifier is also referred to as a *test*, and its predicted class as *a test result.*

Let $t = 1$ denote a positive, and $t = 0$ a negative result. Similarly, $y = 1$ denotes the existence of the underlying condition we are classifying, and $y = 0$ the absence of it. Identification of test misclassification rates requires knowledge of y, which is most often unobservable.[2] For this reason, the ground truth is commonly measured by a reference test r, which we refer to as the *reference test.* Let $r = 1$ and $r = 0$ denote positive and

[2] Otherwise, the classifier would be superfluous.

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.

Test performance is predominantly quantified in the form of *sensitivity* and *specificity*, also referred to as performance measures or operating characteristics.[3]

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

$$\text{Specificity: } \theta_0 := P(t = 0|y = 0) \tag{2}$$

The parameters are defined when $P(y = 1) \in (0, 1)$ in the population of interest.

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 consider a setting in which we 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 $s := (s_0, s_1)$.

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 $P(t, r)$, but not $P(t, r, y)$ or more specifically $P(t|y)$. 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 et al., 2009) Moreover, it is known that incorrectly assuming $s_1 = s_0 = 1$ will produce biased estimates of the true θ.

One approach for identifying θ relies on assuming exact knowledge of s in addition to 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) indicates that allowing $t \not\perp\!\!\!\perp r|y$ partially identifies θ when s is

[3] In the machine learning and related literature, these two measures are also called *recall* and *true negative rate*, respectively.

known. 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) provides the sharp identified set under standard assumptions in the test performance study literature. That is, he provides 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 provide the assumptions sufficient to characterize $\Theta_P(s)$ and build upon these results.

Assumption 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, it weakens standard assumptions. The current norm of assuming that the reference test is perfect means that $s = (1, 1)$ and is therefore covered by Assumption 1.[4]

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 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 5. Additionally, there, we will provide an alternative formulation which yields a confidence set for s such that t is a dilation. One can then consider whether it is plausible for performance of the reference test to lie in this set.

Assumption 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). Second, note that $s_1 > 1 - s_0$ is merely a normalization and therefore without loss of generality. To see that it without

[4] Examples in the literature that assume perfect references are too numerous to cite.

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 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 1.*

Although, the population prevalence $P(y = 1)$ is unobservable by itself, Assumptions 1 and 2 jointly point-identify $P(y = 1)$ through to the implied prevalence $P_s(y = 1) := \frac{P(r=1)+s_0-1}{s_1+s_0-1}$, where we indicate the dependence on s explicitly, and then we also have $P_s(y = 0) = 1 - P_s(y = 1)$, of course. If $P_s(y = 1) \notin [0, 1]$ at least one of the two assumptions is refuted. We call Assumption 2 *Bounded Prevalence* because it is then equivalent to assuming the population prevalence satisfies $P(y = 1) \in (0, 1)$, which is an assumption that is implicit in any test performance study identifying sensitivity or specificity. Without this assumption, the performance measures are not properly defined.

Under Assumptions 1 and 2, and given the data distribution $P(t, r)$, Obradović (2024, Proposition 1) defines $\Theta_P(s)$ as

$$\Theta_P(s) := \left\{ (\theta_0, \theta_1) \in [0,1]^2 \;\middle|\; \begin{array}{l} \theta_1 \in [\theta_1^L(s), \theta_1^U(s)] \text{ and} \\ \theta_0 = \dfrac{\theta_1 P_s(y=1) - P(t=1)}{P_s(y=0)} + 1 \end{array} \right\}, \tag{3}$$

where

$$\theta_1^L(s) := \frac{1}{P_s(y=1)} \Big[\max\{0, P(t=1, r=0) - s_0 P_s(y=0)\} + \max\{0, P(t=1, r=1) - (1-s_0) P_s(y=0)\} \Big], \tag{4}$$

and

$$\theta_1^U(s) := \frac{1}{P_s(y=1)} \Big[\min\{P(t=1, r=0), (1-s_1) P_s(y=1)\} + \min\{P(t=1, r=1), s_1 P_s(y=1)\} \Big]. \tag{5}$$

Remark 1. *If $\Theta_P(s)$ is non-empty, then it is either one point or it corresponds to a line segment in $[0, 1]^2$ with positive and finite slope. In particular, note that when $s = (1, 1)$, i.e. the reference test is perfect, then $\theta_1^L(s) = \theta_1^H(s) = P(t = 1|r = 1)$ and therefore*

$\Theta_P(s)$ is a singleton set. That is, under a perfect gold standard, point identification is achieved.

Assumption 3. *(Anything Goes) For any $(j, k) \in \{0, 1\}^2$, $P(t = j, r = k) > 0$.*

To obtain a characterization of a dilation that is conducive to testing by existing subvector inference methods, we maintain Assumption 3. This condition is 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. Such dependence is generally not expected for two tests and it is testable. We emphasize that Assumption 3 is not necessary to identify $\Theta_P(s)$ nor to characterize t as a dilation in terms of $\theta \in \Theta_P(s)$. It is only used to further simplify the characterization. We expound on the details in Subsection 2.2.

2.2 Decisions and Dilations

After determining the performance $\Theta_P(s)$ of the novel test t, it is natural to consider its prospects as a tool for decision-making. In line with our motivating example in the introduction, we discuss the use t in the context of a clinical setting. Analogous discussions follow for classifiers in other settings. The performance of t is learned from the performance study upon observing $P(t, r)$ using s. Let $Q(t, y)$ denote any clinical population distribution such that the test has the same sensitivity and specificity as in the performance study population. Formally, let $Q(t, y)$ be any distribution such that $P(t|y) = Q(t|y)$. Note that it is possible that $Q(y = 1) \neq P(y = 1)$. Suppose also $Q(y = 1) \in (0, 1)$ since using t is not warranted otherwise.

We first explain the importance of post-test probabilities for decision-making and define them. Then we define dilations using post-test probabilities. Finally, we provide a characterization for t to be a dilation in terms of $\Theta_P(s)$.

2.2.1 Post-Test Probabilities

Sensitivity and specificity measure the likelihood of obtaining a particular test result given a specific health condition. However, in risk assessment and (clinical) decision-making, the focus is on determining the probability of actually having or not having a disease based on the test result, expressed as $Q(y = j|t = j)$ for $j = 0, 1$. The probability of having the disease, given a positive test result ($t = 1$), is called the *positive predictive* *value* (PPV), while the probability of being healthy given a negative test result ($t = 0$) is known as the *negative predictive value* (NPV). Equivalently, one can consider the *positive post-test probability* (PPP), $Q(y = 1|t = 1)$, and the *negative post-test probability* (NPP), $Q(y = 1|t = 0)$. For our purposes, we will use PPP and NPP throughout the discussion for ease of exposistion. In the context of medical decision-making, Altman and Bland (1994), and more recently Manski (2021), argue that sensitivity and specificity are less relevant than post-test probabilities for decision-making. However, they note that sensitivity and specificity are commonly extrapolated from test performance studies to find post-test probabilities for members of relevant clinical populations.[5]

Watson et al. (2020) explain that clinicians assess $\pi := Q(y = 1) \in (0, 1)$, also known as the *pre-test probability*, prior to conducting the test. This is done based on local rates of illness, patients' symptoms and signs, likelihood of alternative diagnoses, and history of relevant exposure. PPP and NPP are formed using Bayes' rule, based on knowledge of $\theta \in \Theta_P(s)$ from the performance study and the assessed π. Decisions are made depending on the relevant post-test probability upon observing t. In this paper, we limit the analysis to informativeness of t in terms of post-test probabilities, and we do not discuss the intricacies of decision-making.

Remark 2. *The following results do not depend on the clinician accurately assessing π, as they hold uniformly for all $\pi \in (0, 1)$. However, these results apply only to clinical populations where the (potentially unknown) parameter θ accurately reflects the test's performance,* i.e. *$P(t|y) = Q(t|y)$ is assumed to hold.*

Thus, for a given θ and π, PPP and NPP are

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

respectively.

[5] Mulherin and Miller (2002) and Willis (2008) discuss design 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.

2.2.2 Dilation

Suppose first that θ is point identified. As previously mentioned, post-test probabilities follow directly for a given π from the Bayes' rule. Figure 3a illustrates this updating graphically for $\theta_1 + \theta_0 > 1$. 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.

When 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 as:

$$V_j(\Theta;\pi) := \Big\{ v_j(\theta;\pi) : \theta \in \Theta \Big\} \text{ for } j = 0, 1, \tag{6}$$

which are depicted in Figure 3. The interpretation of the post-test probabilities is unchanged, but they are not known exactly. Indeed, this is an instance of imprecise probability or ambiguity in the test result. However, in the figure the test is still informative in the following sense: 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.

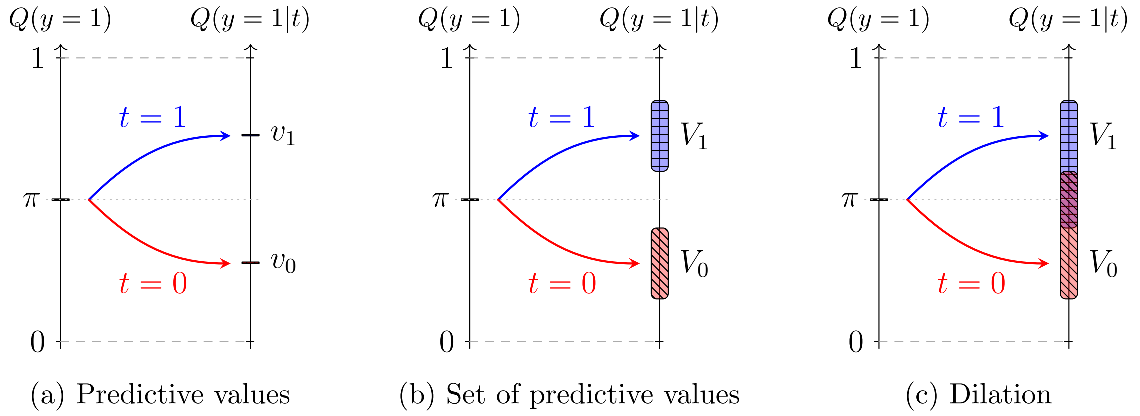
Finally, consider the case in Figure 3c. Here, the pre-test probability is strictly contained within the identified set for the post-test probability, regardless of the observed test result. That is, observing the test result not only introduces ambiguity, but this ambiguity is so pronounced that for neither post-test probability an unambiguous direction of change can be identified. We will call such a test *uninformative*, and this is the main idea behind the phenomenon known as *dilation*.

Definition 1 (Seidenfeld and Wasserman, 1993)**.** *Given the identified set* Θ*, the index test is called* a dilation for pre-test probability π *if*

$$\{\pi\} \subsetneq V_1(\Theta;\pi) \textit{ and } \{\pi\} \subsetneq V_0(\Theta;\pi). \tag{7}$$

An index test is called a dilation *if it is a dilation for every pre-test probability* $\pi \in (0, 1)$.

Thus, we say that t is a *dilation for* π if the pre-test probability π is strictly contained within the identified set of possible post-test probabilities of being diseased, regardless of the test outcome. We refer to a test t as a *dilation* if it is a dilation for any possible,



(a) Predictive values (b) Set of predictive values (c) Dilation

Figure 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 a dilation.

non-trivial pre-test probability π. In other words, a dilation occurs when the test result, rather than narrowing down the likelihood of disease, strictly broadens the range of possible probabilities, leaving more uncertainty than before for any initially assigned pre-test probability.

Remark 3. *A couple of remarks are necessary to clarify our Definition 1 in relation to the original definition by Seidenfeld and Wasserman (1993). On one hand, they accommodate imprecise probabilities in the pre-test probability, allowing for cases where π is contained within a known set. On the other hand, apart from the just-mentioned difference, our definition demands uniformity across all non-trivial π. In contrast, their definition applies to a specific π, i.e., what we refer to as dilation for π.*

2.2.3 Characterizing Dilation

Whether a test t is a dilation critically depends on the identified sets for its performance measures, generically Θ. Our first main result characterizes necessary and sufficient conditions for t to be a dilation.

Proposition 1. *Let Θ be a connected identified set for the performance measure of the index test. The index test is a dilation 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.*

Remark 4. *According to Seidenfeld and Wasserman (1993, Theorem 2.1) and assuming convexity of Θ, existence of dilation implies that $t \perp\!\!\!\perp y$ is consistent with the data and* *assumptions. Proposition 1 nests this result. The convexity requirement would imply path-connectedness of Θ in our setting, and when t is a dilation then there must exist $\theta \in \Theta$ such that $\theta_1 + \theta_0 = 1$ (*cf. *Theorem 1), which is equivalent to $t \perp\!\!\!\perp y$. However, although necessary, existence of such a θ is not sufficient to obtain a dilation, because of the strict set-containment requirement; see also the next remark. Furthermore, the proof itself only requires Θ to be connected, but the argument here requires path-connectedness. Connectedness, however, is required for the result hold.*[6]

Remark 5. *Note that dilation can only occur if Θ is not a singleton. Thus, a necessary condition for a dilation is the presence of ambiguity that arises naturally in our setting from partially identified performance measures. In other words, if the performance measures are point-identified, a dilation cannot occur. The easiest example of this is when $s = (1, 1)$, meaning, the reference test is perfect—–a proper gold standard,* cf. *Remark 1.*

Proposition 1 takes Θ as a primitive object, but for the relevant application this set will be derived from the data and s, namely, as $\Theta_P(s)$ given by Equation 3. Under the maintained assumption, and then combining Remark 1 with Proposition 1 gives the desired result as a corollary.

Corollary 1. *Suppose Assumptions 1 and 2 hold, and let $\Theta_P(s)$ denote the corresponding identified set for the performance measures of the index test. The index test is a dilation 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.*

In light of Remark 5, we also need $\Theta_P(s)$ to be a non-singleton set to ensure that there is actual ambiguity in the performance of the index test. Assumption 3 provides a sufficient condition for such ambiguity in the absence of a perfect reference test. Furthermore, it is suitable for the empirical applications we have in mind, as it only fails if a specific result for $r = k$ makes a particular outcome for $t = j$ impossible almost surely. Such dependence is generally not expected for two imperfect binary classifiers.

Lemma 1. *Suppose $s \in [0, 1]^2$ satisfies Assumption 1 and maintain Assumption 2. Then $\Theta_P(s)$—as defined in Equation 3—is non-empty. Furthermore, if Assumption 3 holds additionally, then $\Theta_P(s)$ is not a singleton set if and only if $s = (s_0, s_1) \neq (1, 1)$.*

[6] For an easy example, consider $\Theta = \{(1, 1), (0, 0)\}$. Then, $V_1(\Theta; \pi) \cap V_0(\Theta; \pi) = \{0, 1\}$, which does not intersect with $(0, 1)$ at all, of course.

Now we are ready to establish the main identification result formally as Theorem 1, in which Assumption 3 allows a simplification of the dilation characterization. Later we will show how this identification result allows formulating a tractable subvector inference problem which can be solved using existing tools.

Theorem 1. *Maintain Assumptions 1, 2 and 3, and let $\Theta_P(s)$ be the resulting identified set as in Equation 3. Then t is a dilation if and only if (1) $s \neq (1,1)$ and (2) there exists $\theta \in \Theta_P(s)$ such that $\theta_1 + \theta_0 = 1$.*

For results of Theorem 1 to be used directly for inference, one must first assume knowledge of s, which might be unsatisfactory for some applications. In Section 5, we extend Theorem 1 to cases where s is only known approximately or not at all.

2.3 Numerical Examples

To illustrate the key points from the previous section, we present three examples. First, we consider the extreme case where the index test is independent of the reference test, with the joint distribution given in Table 1. Second, we examine a case where the index test is weakly correlated with the reference test, as shown in Table 2. 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. In all cases, we set $s = (0.9, 0.9)$, indicating that the reference test performs reasonably well, although it is not perfect. This choice of s satisfies Assumption 1 under the normalization $s_1 > 1 - s_0$. Additionally, Assumption 3 is satisfied in all three cases.

Table 1: Independent joint distribution of index and reference test results.

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

In the first case, where the index test is independent of the reference test, and if the reference test were perfect, the index test would also be independent of the underlying health condition. However, due to the imperfection of the reference test, the performance measure of the index test is only partially identified. The left panel of Figure 4 illustrates

Table 2: 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: Highly correlated joint distribution of index and reference test results.

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

the resulting partially identified sets. Even though the performance of the reference test is known precisely, the set $\Theta_P(s)$ still contains multiple possible performance measures (denoted by θ), represented by the dark, solid line in the figure. Consequently, we lack point identification. Intuitively, this is because, while we know that the index and reference tests are independent, we do not know the exact correlation between the index test and the underlying health condition. For the other two tests, partial identification occurs for the same reason, as shown in the center and right panels of Figure 4.

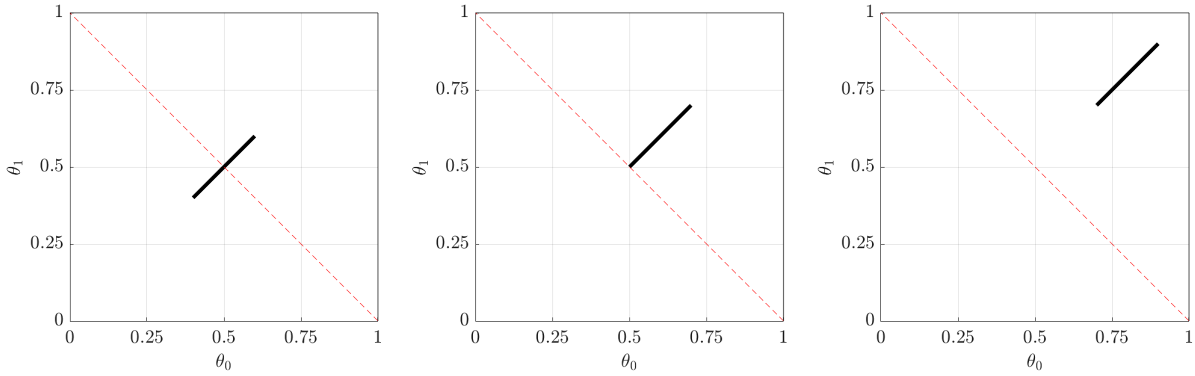


Figure 4: $\Theta_P(s)$ for independent tests, weakly correlated, and highly correlated test, respectively from left to right.

Maybe not surprisingly, in the independent case, the index test is a dilation.[7] This is clearly visible in Figure 4 by applying Theorem 1: the index test is a dilation if and only if the identified set intersects the antidiagonal, represented by the red, dashed line in the figure. This illustrates the simplification provided by Theorem 1. Now for the correlated

[7] It is worth noting, however, that the index test would *not* be a dilation if the reference test were perfect. See Remark 5.

cases in the center and right panel of Figure 4, we observe that the index test remains a dilation 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.

To further verify these observations, we turn to the post-test probabilities. In Figure 5, we consider a specific pre-test probability, $\pi = 0.5$.[8] This figure confirms the earlier insights by directly applying the definition of a dilation: the index test is a dilation in the first two cases (independence and weak correlation), but not in the case of high correlation. Notably, the weak correlation case is only marginally a dilation, because a slight perturbation of the joint distribution towards higher correlation would render it an informative test.

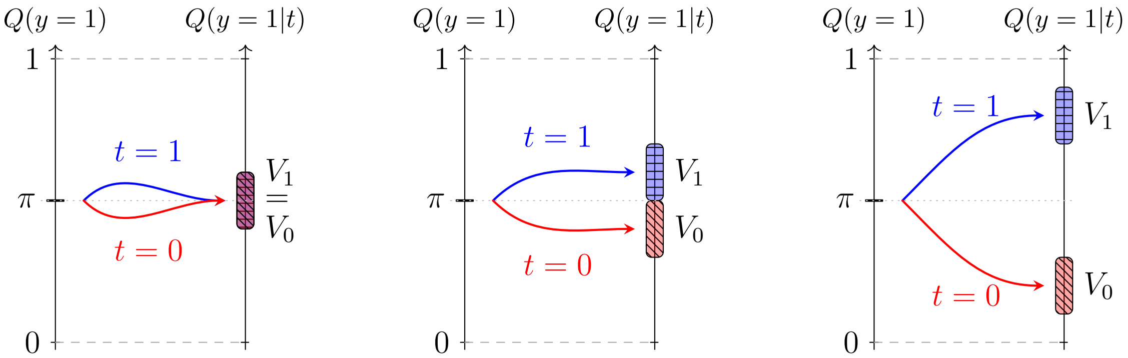


Figure 5: Sets of post-test probabilities for independent tests, weakly correlated, and highly correlated test, respectively from left to right.

3 Estimation and Inference

Theorem 1 in Section 2 fully characterizes when the index test t is a dilation, but this requires knowledge of the data distribution $P(t, r)$. In practice, this distribution is typically unknown, and only sample data from $P(t, r)$ are available. To address this, based on our characterization in Theorem 1 we develop an inference procedure to test whether t is a dilation which is uniformly consistent in level across a broad class of permissible distributions.

Analogous to standard point estimation and inference problems, one can first "estimate" whether t is a dilation. Replacing population parameters with consistent estimators

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

in closed-form expressions yields the consistent plug-in estimator $\hat{\Theta}_P(s)$ of the identified set $\Theta_P(s)$. (Manski and Pepper, 2000 and Tamer, 2010) Then, based on Theorem 1, t is "estimated" to be a dilation if there exist $\theta \in \hat{\Theta}_P(s)$ such that $\theta_1 + \theta_0 = 1$.

To account for sampling variability, we construct a hypothesis test for dilation. Our test is uniformly valid across a wide range of distributions. Uniformity ensures that the actual coverage probability closely matches the prescribed confidence level for all distributions, regardless of sample size. Without uniformity, there is a risk that—for any given sample size—some distributions may produce confidence regions that deviate from the intended coverage, undermining inference reliability. Moreover, uniformity generally leads to better finite sample performance than tests that are not uniformly valid. (Canay and Shaikh, 2017, Section 3.1; Canay et al., 2023, Remark 2.1)

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 identified with an element Δ^3—the three-simplex. Therefore, we identify $\mathcal{P}$ with 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.

As usual, we will assume that we have access to a random sample.

Assumption 4. *(Random Sampling) For every $P \in \mathcal{P}$, 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.*

To address the aforementioned uniformity issue, we need to strengthen some of the assumptions from Section 2 to hold uniformly, too. Obviously, the results in the previous section remain true with these stronger assumptions.

Assumption 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 1.*

Together with Assumption 1, Assumption 2' is equivalent to a uniform bound on the implied prevalence $P_s(y = 1)$. To see this, recall—from the discussion following

Assumption 2—that specific s determines $P_s(y = 1)$ via $P(r = 1)$. Here, we have the corresponding statement that holds uniformly across all $P \in \mathcal{P}$.

Finally, Assumption 3 requires a strengthening to hold uniformly too.

Assumption 3'. *(Uniformly Non-degenerate Data) There exists an $\varepsilon_d \in \left(0, \frac{1}{4}\right)$ such that for every $P \in \mathcal{P}$ and every $(t, r) \in \{0, 1\}^2$, $P(t, r) \geq \varepsilon_d$ holds.*

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

Lemma 2. *If Assumption 1, Assumption 2' and Assumption 3' hold, then $\mathcal{P}$ is compact.*

3.2 The Proposed Test

We are interested in testing whether the index test is uninformative in the sense of being a dilation. Using Theorem 1, we can formulate the hypothesis as follows:

$$H_0 : \theta_0 + \theta_1 = 1 \quad \text{vs.} \quad H_1 : \theta_0 + \theta_1 \neq 1. \tag{8}$$

As mentioned above, 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 $\mathcal{P}$. We accomplish this by leveraging two recent results: following Obradović (2024) we characterize the identified set through moment (in)equalities, and then are able to apply the minimum resampling test from Bugni et al. (2017), which ensures the desired properties.

In the following, we describe how these two components work concretely in our context. Specifically, we will start by showing how the conditions that define the identified set $\Theta_P(s)$ can be recast as moment (in)equalities through the introduction of an appropriate *moment function*, defined as follows:

$$m(W_i, \theta; s) := \begin{pmatrix} m_1(W_i, \theta; s) \\ m_2(W_i, \theta; s) \\ m_3(W_i, \theta; s) \\ m_4(W_i, \theta; s) \\ m_5(W_i, \theta; s) \\ m_6(W_i, \theta; s) \\ m_7(W_i, \theta; s) \end{pmatrix} := \begin{pmatrix} (\theta_1 - s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - (t_i - 1)r_i \\ (\theta_1 - 1 + s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - (r_i - 1)(1 - t_i) \\ (\theta_1 - 1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} - (t_i - 1) \\ -\theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i \\ (-\theta_1 + s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i(1 - r_i) \\ (-\theta_1 + 1 - s_1)\frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i r_i \\ (\theta_0 - 1)\left(1 - \frac{r_i - 1 + s_0}{s_1 - 1 + s_0}\right) - \theta_1 \frac{r_i - 1 + s_0}{s_1 - 1 + s_0} + t_i \end{pmatrix}, \tag{9}$$

where $W_i = (t_i, r_i)$. Intuitively, the first three functions handle the condition $\theta_1 \geq \theta_1^L(s)$ (*c.f.* Equation 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 overall parameter space we are considering. Therefore, only three functions are both necessary and sufficient. Similarly, functions 4 through 6 address $\theta_1 \leq \theta_1^U(s)$ (*c.f.* Equation 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 Equation 3 on θ_1. Finally, the last function establishes the linear relationship between θ_0 and θ_1 (*c.f.* Equation 3 and Remark 1). These functions collectively allow us to represent the identified set $\Theta_P(s)$ through moment (in)equalities, thus enabling the application of results from this literature.

Proposition 2 (Obradović, 2024, Proposition 5)**.** *Suppose s satisfies Assumption 1 and P satisfies Assumption 2, then*

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

With the problem now framed as a moment (in)equality model, we are ready to formally define the proposed test, following Bugni et al. (2017). The test rejects the null hypothesis in Equation 8 when the profiled test statistic, denoted as T_n, is large enough and exceeds a certain critical value, where n represents the sample size of $(W_i)_{i=1}^n$. To define T_n rigorously and clarify how it functions in the test procedure, we must first introduce the necessary additional notation.

For $j = 1, \ldots, 7$, let

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

denote the sample mean and standard varaince of the moment functions, respectively. Furthermore, we need the so-called modified method of moments test statistic

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

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, which——as previously discussed——plays a key role in the test.

To determine whether the test statistic is sufficiently large to reject the null hypothesis, we also need a critical value $\hat{c}_n^{1-\alpha}$, which depends on the significance level $\alpha \in (0, 1)$. The formal definition of $\hat{c}_n^{1-\alpha}$ requires additional notation, and thus we defer the details to Subsection A.1.

With this notation in place, we can now formally establish that our proposed test controls size uniformly over all $P \in \mathcal{P}$, under the assumptions stated in Subsection 3.1.

Theorem 2. *Let Assumptions 1, 2', 3', and 4 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 2 asserts that, under the specified assumptions and as the sample size n tends to infinity, the maximum probability—across all considered distributions that satisfy the null hypothesis——that the test statistic T_n exceeds the critical value $\hat{c}_n^{1-\alpha}$ does not surpass α. This means that the test maintains its nominal significance level asymptotically, ensuring the probability of incorrectly rejecting the null hypothesis remains controlled at α in the limit, regardless of which distribution within the considered class generated the data. Thus, the test controls size uniformly because the error rate is controlled simultaneously for all distributions satisfying the null hypothesis, not just a specific one.

The proof of Theorem 2, provided in Subsection C.1, relies on an application of Theorem 4.1 from Bugni et al. (2017). Our assumptions allow us to verify that their result applies in this setting. Specifically, we explicitly show how the relevant polynomial minorant condition—–which ensures that the test statistic grows sufficiently fast as it moves away from the null hypothesis—–is satisfied. Additionally, we establish the uniform Donsker and pre-Gaussian property directly, whereas Bugni et al. (2017) impose assumptions that imply it.

Remark 6. *As discussed in Section 4 of Bugni et al. (2017), their result extends to more general test functions and critical values. We conjecture, though without formal argument, that our Theorem 2 also generalizes to this broader class.*

3.3 Simluations

In this section, we analyze the finite sample behavior of our method through a simulation study. It is well known that inference in partially identified models often tends to be overly conservative. Therefore, this simulation study aims to shed light on how our proposed test performs in finite samples, particularly in terms of observed significance and power. To provide a meaningful comparison, we evaluate the performance of our method alongside two other established approaches. Given the conservativeness typically encountered in partially identified models, this comparison helps assess how each method balances significance and power in finite samples.

First, we consider a test based on the popular two-step procedure of Romano et al. (2014), which is designed for testing a finite number of moment inequalities and—as we have argued earlier—encompasses our test. Additionally, we evaluate a test based on the approach of Goodman (1965). While the detailed discussion of this test is deferred to Subsection A.2, in brief, it leverages the multinomial nature of our data, and Goodman (1965) provides a method for obtaining simultaneous confidence intervals for the parameters. In our case, this results in confidence intervals for $\theta_0 + \theta_1$.

Throughout the simulation study, we fix the reference test with performance measure $s = (0.9, 0.9)$ and consider five data-generating processes. In addition to the three cases introduced in Subsection 2.3, we introduce two additional designs to explore the power of the tests. The first additional design involves a slight perturbation of the joint distribution in Table 2, which previously resulted in a dilation. This perturbation, presented in Table 4, results in an index test that is no longer a dilation, providing insight into how the tests respond to small deviations from the dilation condition. The second additional design, shown in Table 5, 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. This design allows investigating, again, he power of the test for an intermediate scenario.

Table 4: Perturbation of Table 2.

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

Table 5: Data generating process for intermediate case of 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%	

In terms of sample size, we consider three different scenarios: $n \in \{50, 100, 500\}$. These sample sizes are relatively small but are typical for the applications we have in mind. For each design, we perform $1,000$ Monte Carlo iterations and set the significance level at 5%.

Table 6 presents the results from the simulation study, showing the rejection probabilities for all the considered designs. Across all designs, we observe that all three tests tend to be conservative: they reject the null hypothesis with a probability lower than the nominal 5% significance level, even when the null hypothesis is true. However, our proposed test (denoted as BCS in the table) consistently outperforms the two other tests. For the first two designs, where the null hypothesis is true and the index test is a dilation, our test rejects the null hypothesis more frequently than the alternatives while maintaining significance below the nominal 5% level. This demonstrates that our test has

Table 6: Simluation results: Observed rejection probabilites.

		Design 1	Design 2	Design 3	Design 4	Design 5
DGP		Table 1	Table 2	Table 4	Table 5	Table 3
H_0		true	true	false	false	false
$n = 50$	G	0.9%	0.6%	0.2%	0.7%	4.3%
	RSW	0%	0%	0%	0.1%	4.3%
	BCS	0%	0.8%	3.9%	22%	76%
$n = 100$	G	0%	0%	0%	1%	30%
	RSW	0%	0%	0%	0.1%	19%
	BCS	0%	1%	3.1%	43%	100%
$n = 500$	G	0%	0%	0%	50%	100%
	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 et al. (2014), and Bugni et al. (2017), respectively. The last one is our proposed test.

better performance in controlling the error rate while still being conservative, as expected for partially identified models. In contrast, for the other three designs, where the null hypothesis is false, the power of the tests becomes the relevant measure. Here, our proposed test shows significantly higher power, even with smaller sample sizes. Considering Design 4 and the largest sample size, $n = 500$, our test rejects the null hypothesis nearly 100% of the time, whereas the other tests only reject about half the time at best. Notably, in the borderline case of Design 3—an especially difficult scenario where there is no dilation—our test still manages to reject the null hypothesis occasionally, while the other two tests never reject it. Thus, we conclude that even with relatively small sample sizes, our proposed test shows reasonable power. Although it remains somewhat conservative, it is notably less so than the two other tests.

4 Applications

In this section, we apply our proposed method to real-world data to demonstrate its practical relevance and performance in empirically relevant settings.

4.1 CT chest scans for the detection of COVID-19

Early in the COVID-19 pandemic, some hospitals used CT chest scans, interpreted by radiologists, as a method to test for COVID-19. This diagnostic technique was typically evaluated against a PCR test, which served as the reference. Since PCR tests are not entirely perfect[9], this scenario fits precisely within our framework: the index test is the CT chest scan, the reference test is the PCR test, and the underlying health condition is whether the patient has COVID-19. As a concrete application, we use data from Ai et al. (2020), 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." The data they obtained is reproduced in Table 7. Furthermore, we need to specify the accuracy of the reference test, the PCR test. Following Kanji et al. (2021), we assume $s = (1, 0.9)$.

Table 7: Data from Ai et al. (2020) with $t = 1$ and $r = 1$ denoting a positive CT-chest scan and a positive PCR-test, respectively.

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

Taking the empirical distribution as if it represents the population distribution, Figure 6 displays the corresponding identified set for the accuracy measures of the CT chest scan, $\Theta_P(s)$. Additionally, assuming a pre-test probability of $\pi = 1/3$, Figure 7 illustrates the ambiguity regarding the post-test probabilities. 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. More importantly, both figures suggest that a CT chest scan acts as a dilation, implying that it is uninformative. However, this conclusion depends on treating the empirical distribution as if it were the true population distribution, and therefore does not account for sampling variability in the data.

Applying our proposed test from Subsection 3.2 at a nominal significance level of $\alpha = 5\%$, we obtain a test statistic of $T_n = 1.2518 \times 10^{-18}$ and a critical value of $\hat{c}_n^{1-\alpha} = 1.112$. Therefore, we cannot reject the null hypothesis that the CT chest scan is a dilation. Furthermore, by varying the significance level, we find that the p-value for this null hypothesis is greater than 99%. In light of the simulation insights from Subsection 3.3, which show that the test is somewhat conservative, this result strongly suggests that the CT chest scan is indeed a dilation——the first concrete real-world instance of such a case.

[9] See, for example, Arevalo-Rodriguez et al. (2020).

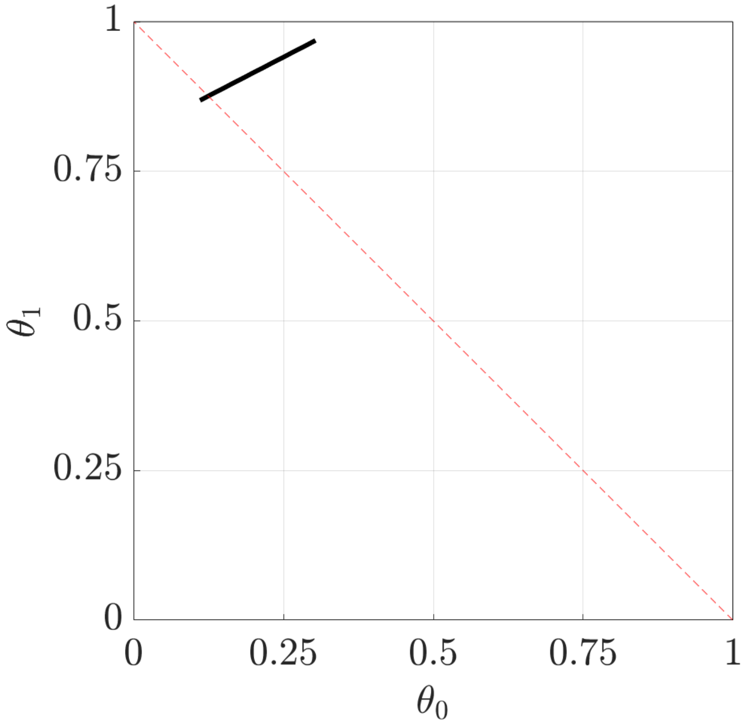


Figure 6: $\Theta_P(s)$ for the empirical distribution of Ai et al. (2020) assuming $s = (1, 0.9)$.

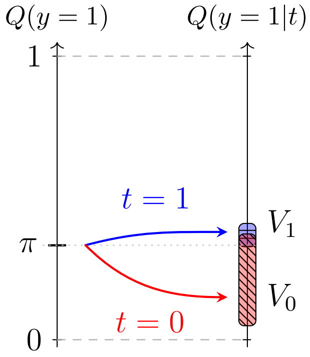


Figure 7: Sets of post-test probabilities for independent for empirical distribution of Ai et al. (2020) assuming $\pi = 1/3$.

Instead of having humans interpreting the CT chest scan, Mei et al. (2020) use AI algorithms to classify the CT chest scan as either positive or negative. Their data is reproduced in Table 8. They also use a PCR test as a reference and therefore we set $s = (1, 0.9)$ here too.

Again, we want to test if the index test, i.e. the CT chest scan interpreted by their AI algorithm, is a dilation at a significance level of $\alpha = 5\%$. Here we get a test statistic of $T_n = 39.6268$ and a critical value of $\hat{c}_n^{1-\alpha} = 4.5822$. Since $T_n > \hat{c}_n^{1-\alpha}$, we reject the null hypothesis that the index test is a dilation. Varying the significance level, we furthermore

Table 8: Data from Mei et al. (2020) with $t = 1$ and $r = 1$ denoting a positive CT-chest scan and a positive PCR-test, respectively.

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

find a p-value of less than 0.001%. In contrast to the previous case, here we conclude that the proposed AI algorithm is informative in the sense of not being a dilation.

4.2 A deep neural network to predict loan approval

In this section, we apply our proposed statistical test to the context of loan approval decisions, a critical area in the financial sector that heavily relies on data-driven methods. Specifically, we examine binary classification models used to assess the risk associated with loan applicants. The real-time binary classification model proposed by Abakarim et al. (2018) offers a suitable case study for evaluating whether such a machine learning model is informative or a dilation.

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 et al. 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. However, two factors raise concerns about whether the reference test perfectly reveals the truth. First, the aforementioned data issues may lead to erroneous classifications. Second, it is unclear whether an objective measure of "riskiness" truly exists in this context—even if it is correct historic data, the recordings of riskiness must be somewhat subjective. Because of these imperfections, the data from Abakarim et al. (2018), reproduced in Table 9 provides a valuable basis for applying our proposed test.

Table 9: Data from Abakarim et al. (2018) with $t = 1$ and $r = 1$ denoting a ”good risk” application according to the machine learning algorithm and the data set, respectively.

	$r = 0$	$r = 1$	
$t = 0$	203	43	246
$t = 1$	97	657	754
	300	700	$n = 1000$

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 to apply our model. Therefore, we will 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. To mitigate the need for precise assumptions about the reference test's accuracy, our extensions in Section 5 offer a more flexible approach that accommodates uncertainty about—or even a complete lack of knowledge of—the quality of the reference test.

For the test with the null hypothesis that the machine learning algorithm is a dilation, we proceed with a nominal significance level of $\alpha = 5\%$, as before. In the first case, with $s = (0.9, 0.9)$, we obtain a test statistic of $T_n = 35.9578$ and a critical value of $\hat{c}_n^{1-\alpha} = 5.2481$, 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 observe an even stronger rejection of the null hypothesis. Therefore, considering the aforementioned caveats regarding the assumptions about the reference test's accuracy, we conclude that, at least for these exploratory cases, the machine learning approach is informative in predicting loan riskiness. This analysis exemplifies the usefulness of our proposed test in evaluating machine learning algorithms in financial decision-making contexts.

5 Extensions and Discussion

5.1 Uncertainty about the reference test's performance

Throughout, we have maintained Assumption 1, which assumes that the performance of the reference test, while allowed to be imperfect, is exactly known. This assumption may introduce challenges, particularly when reliable estimates for the performance of the reference test are difficult to obtain. Real-world data often contains uncertainties or varying estimates, making such strict assumptions difficult to justify in some cases. For example, in Subsection 4.1, to apply our framework effectively, we had to assume that the sensitivity and specificity of the reference test—specifically, the PCR test for COVID-19—were known with precision. However, estimates for the sensitivity of such a PCR test can vary, as illustrated by Alcoba-Florez et al. (2020).

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 1 as follows.

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

Path-connectedness of $\mathcal{S}$ will be used to extend our characterization in Theorem 1. While the preliminary characterization in Proposition 1 only requires connectedness, the simplification provided by path-connectedness is crucial for our inference procedure in the case where $\mathcal{S}$ is not a singleton set. We believe that Assumption 1S is relatively mild for applications, as it 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 a non-trivial assumption, it is clearly a relaxation of Assumption 1. We further 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 (path-)connected, the entire set lies either fully above or fully below the antidiagonal of the unit rectangle. Thus, we impose the same normalization, $s_0 + s_1 > 1$, as before, but now across all $s \in \mathcal{S}$.

Treating $\Theta_P(\cdot)$, as defined in Equation 3, as a correspondence, we can readily extend the sharply identified set of performance measures for the index test—now denoted as $\Theta_P(\mathcal{S})$—by taking it as the image of $\mathcal{S}$ under $\Theta_P(\cdot)$. Moreover, the path-connectedness of $\mathcal{S}$ carries over to $\Theta_P(\mathcal{S})$, as we establish next.

Lemma 3. *If Assumption 1S holds, then* $\Theta_P(\mathcal{S})$ *is a path-connected set.*

Since path-connectedness implies connectedness, Lemma 3 ensures that Proposition 1 remains applicable, and we therefore get a direct generalization of Corollary 1 with a suitable extension of Assumption 2.

Assumption 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 1S.*

Corollary 2. *Suppose Assumptions 1S and 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 a dilation 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.*

Furthermore, we also derive a similar implication of Assumption 3 as in Lemma 1, namely the emergence of ambiguity, indicating non-point-identification in the index test's performance when the reference test is not perfect. In addition, the non-emptiness of $\Theta_P(\mathcal{S})$ carries over too.

Lemma 4. *Suppose $\mathcal{S} \subset [0,1]^2$ satisfies Assumption 1S and maintain Assumption 2S. Then $\Theta_P(\mathcal{S})$—as defined above—is non-empty. Furthermore, if Assumption 3 holds additionally, then $\Theta_P(\mathcal{S})$ 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 a dilation in Theorem 1, which as before allows us to use subvector inference as before.

Theorem 3. *Maintain Assumptions 1S, 2S and 3, and let $\Theta_P(\mathcal{S})$ be the resulting identified set. Then t is a dilation 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 similar to before. In fact, a direct application of Obradović (2024, Proposition 5) gives here too that, under Assumptions 1S and 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 Subsection 3.2 to develop a test for the null hypothesis that the index test is a dilation. Specifically, by modifying the test statistic to

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

the approach of Bugni et al. (2017) becomes applicable once again. However, at this stage, we have not been able to formally extend Theorem 2 which would establish that this extended test controls size uniformly across the wide class of distributions outlined in Subsection 3.1.[10] Nonetheless, our simulations indicate that the size is indeed controlled by this test, leading us to conjecture that a version of Theorem 2 is valid in this setting. Future research may address this gap, providing a more formal extension of Theorem 2 for this case.

5.2 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 Subsection 4.2, even this assumption may be too demanding. In these situations, the researcher might prefer not to make any assumptions about the reference test. Therefore, in this section, we lay out how our approach can be extended to accommodate this lack of knowledge by introducing the concept of a *dilator set*.

Intuitively, we can ask, given the data $P(t,r)$, which reference test, characterized by its performance measure s, would make the index test a dilation. 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 $S_{\geq} := \{(s_0,s_1)\in[0,1]^2 \mid s_0+s_1\geq 1\}$. First, when $s_0+s_1=1$ or $P_s(y=1)\in\{0,1\}$, we define $\Theta_P(s)=[0,1]^2$.[11] Second, for all values of s such that $P_s(y=1)\notin[0,1]$, we

[10] More concretely, we were able to prove all but one condition necessary to apply Theorem 4.1 in Bugni et al. (2017). Subsection C.1 states all these conditions for our main setting. While the relevant polynomial minorant condition, i.e. Assumption A.3(1), is relatively straightforward to establish in our main setting (Lemma A9), it becomes non-trivial in this extension. In particular, we were not able to verify the conditions for s satisfying $\mathbb{E}_P m_1(W,\theta^*,s)<0$ for $\theta^*=\big(P(t=0),P(t=1)\big)$ and $\theta_1>P(t=1)$.

[11] Note that $s_0+s_1=1$ means that the reference test is independent of the underlying health status,

define $\Theta_P(s) = \emptyset$.[12] 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:

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

Note that, if Assumptions 2 and 3 hold, then $s \in \mathcal{D}_P$ if and only if the index test is a dilation by means of Theorem 1.[13] 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.

Taking the data in Table 9 as the true data-generating process, Figure 8 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, the machine learning algorithm of Abakarim et al. (2018) for loan applications would be a dilation. For instance, $s = (0.7, 0.8)$ would result in a dilation, whereas $s = (0.9, 0.9)$ would not. Notably, the latter performance measure was considered in Subsection 4.2. In that section, assuming this level of reference test performance and accounting for sampling variation, we also concluded that the algorithm is (most likely) not a dilation.

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 below. This allows us to apply techniques from moment inequality models also in this extension. For example, one could use the approach from Romano et al. (2014) to construct a confidence set for the dilator set that is uniformly valid in size across all distributions considered in Subsection 3.1.

Proposition 3.

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

and therefore it does not provide any information about the prevalence. Thus, any $P_s(y = 1) \in [0, 1]$ is possible. If $P_s(y = 1) \in \{0, 1\}$, then either sensitivity or specificity of the index test is not well-defined. For example, if $P_s(y = 1) = 0$, then θ_0 in Equation 2 is not properly defined. The natural more general definition would say that θ_0 needs to satisfy $\theta_0 P(y = 0) = P(t = 0, y = 0)$ and then any $\theta_0 \in [0, 1]$ would be consistent with this more general definition. In this case, however, θ_1 in Equation 1 remains well-defined. Thus, we could also set $\Theta_P(s)$ to be a proper subset of $[0, 1]^2$ with the first dimension being $[0, 1]$. This would not affect any of our discussion or results.

[12] If the reference test's performance measure s results in $P_s(y = 1) > 1$ or $P_s(y = 1) < 0$, the assumption about the reference test is refuted by the data, and thus the identified set is empty. See Manski (2007).

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

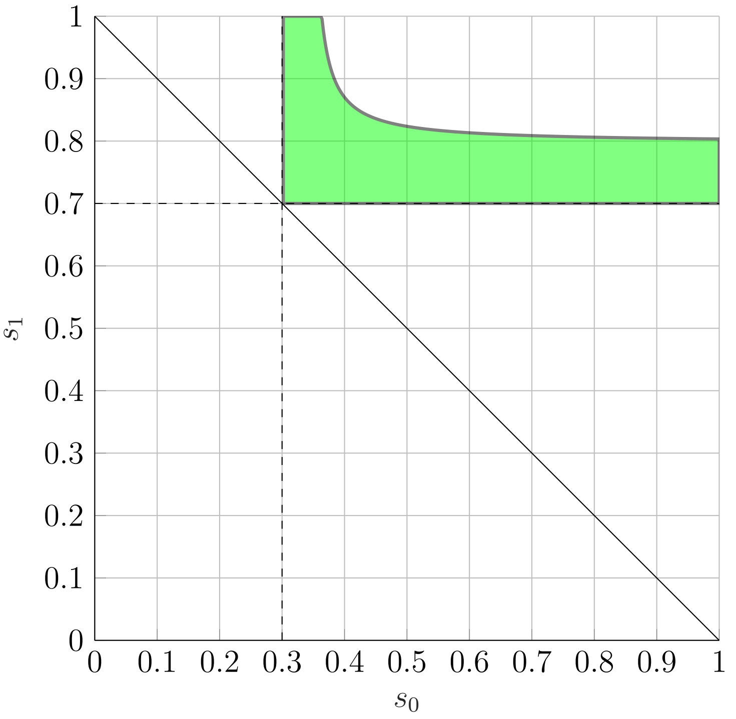


Figure 8: Dilator set $\mathcal{D}_P$ for the machine learning algorithm classifying loan applications taking Table 9 as the true data-generating process. The dashed lines correspond to $s_0 = P(r = 0)$ and $s_1 = P(r = 1)$.

where

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

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

Recall that we extended $\Theta_P(s) = [0,1]^2$ for the case where $s_0 + s_1 = 1$. Alternatively, we could have avoided this extension and simply taken the closure of the resulting set. In that case, the previous proposition would yield only $\mathcal{D}_P^{\llcorner}$ as the fully equivalent set. However, we opted for the current version because we believe it is reasonable to declare a dilation when $s_0 + s_1 = 1$. In either case, 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

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

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

5.3 Policy implications

We conclude this paper with a potential direct policy implication for regulatory agencies, illustrated by a simple example in the context of diagnostic testing. When approving a new diagnostic test—the index test—there are often minimum requirements for specificity and sensitivity. For example, typical thresholds for approval are 97% specificity and 80% sensitivity.[14] However, it is commonly assumed, either explicitly or implicitly, that the reference test is perfect, i.e., $s = (1, 1)$, even when the reference test is actually imperfect. This can lead to significant discrepancies in the evaluation of the index test's true performance as illustrated by Obradović (2024).

The hypothetical data in Table 10 would (exactly) meet these minimum requirements, which might lead a regulator to consider approving the index test. However, if the reference test is imperfect and these assumptions are not accounted for, the performance of the index test might be overestimated. Even worse, the index test could be entirely uninformative—in the sense of being a dilation—despite appearing to perform relatively well and meeting the minimum requirements. Therefore, in such cases, it is worthwhile to supplement the minimum requirements with an explicit test to determine whether the index test is a dilation, using the procedure proposed in Section 3.

Here's the revised version incorporating suggestions 2 and 3:

Another way to view this issue is through the dilator set, as introduced in Subsection 5.2. Taking the data as the actual data-generating process, Figure 9 shows the

[14] Examples include rapid antigen tests for COVID-19 (ECDC, 2021) or influenza (Green and StGeorge, 2018).

Table 10: Potentially worrisome example of a diagnostic test.

	$r=0$	$r=1$	
$t=0$	388	120	508
$t=1$	12	480	492
	400	600	$n=1000$

dilator set for this example. It reveals that even if the reference test has perfect sensitivity ($s_1 = 1$), there is a range of specificity, from 40% to just above 50%, that would still result in the index test being a dilation. On the other hand, even with perfect specificity ($s_0 = 1$), the index test could still be a dilation if the sensitivity is relatively low and around 62%.

Although these cases may seem extreme, as they require a significantly imperfect reference test, they could be relevant for specific applications. For instance, some forms of PCR tests for COVID-19 fall into the latter category.[15] If such a PCR test were used as the reference, the index test could be a dilation and therefore would be uninformative in an extreme sense, despite appearing relatively satisfactory and meeting the minimum thresholds mentioned above. This example therefore demonstrates that relying solely on the minimum requirements might be insufficient and could lead to the approval of an entirely uninformative test. The contribution of our paper provides a framework that can help avoid such mistakes by offering a statistical test with desirable properties.

A Details about the tests

A.1 Critical values for the proposed test

Here, we elaborate on how the minimum resampling critical value $\hat{c}_n^{1-\alpha}$ is calculated following Bugni et al. (2017, Section 2). $\hat{c}_n^{1-\alpha}$ is the $1-\alpha$ of the statistic $T_n^{MR} = \min\{T_n^{DR}, T_n^{PR}\}$, where T_n^{DR} and T_n^{PR} are given as follows.

First, for $j \in \{1, \ldots, 7\}$ define the following stochastic process for

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

[15] For example, based on point estimates, Alcoba-Florez et al. (2020) found that the lowest sensitivity for the tests they considered was only 60.2%. Specificity for these tests is typically close to 100% as mentioned above.

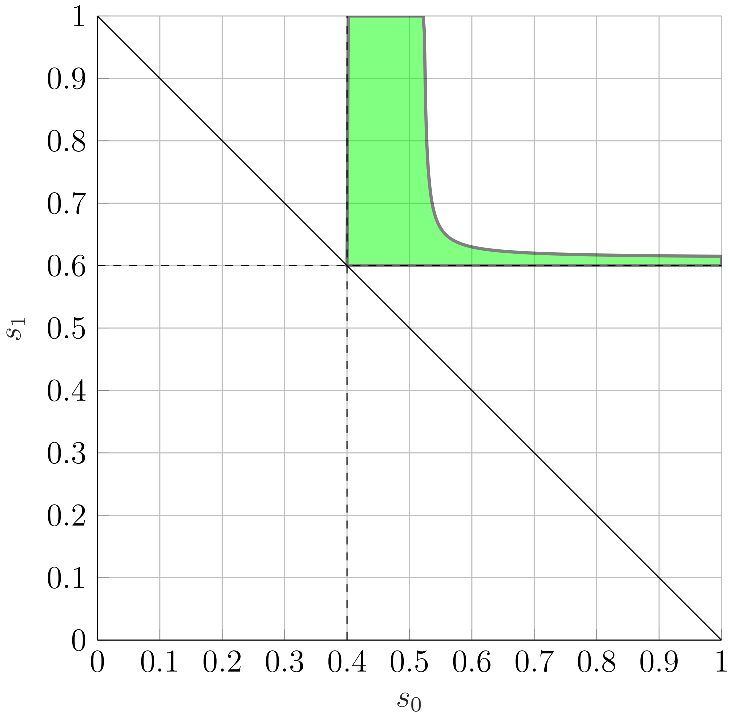


Figure 9: Dilator set $\mathcal{D}_P$ for data-generating process given by Table 10. The dashed lines correspond to $s_0 = P(r = 0)$ and $s_1 = P(r = 1)$.

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[16]

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

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 first statistics is given by

$$T_n^{DR} := \inf_{\theta \in \Theta_0 : Q_n(\theta; s) \leq T_n} \left\{ \sum_{j=1}^{6} \min\left\{0, \nu_{n,j}(\theta; s) + \varphi_j(\theta; s)\right\}^2 + \nu_{n,7}(\theta; s)^2 \right\}, \tag{10}$$

and the second is

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

In our actual implementation, we take the infimum over $\{\theta \in \Theta_0 \mid Q_n(\theta; s) \leq T_n + 10^{-4}\}$ in Equation 10 and also use $\hat{c}_n^{1-\alpha} + 10^{-6}$ as the actual critical value. The reason for intro-

[16] In our implementation, we use the tuning parameter $\kappa_n := \sqrt{\ln n}$ as suggested by Andrews and Soares (2010) and Bugni et al. (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.

duction of these constants are explained by Bugni et al. (2017, Remark 4.1) and Bugni et al. (2017, Remark B.2), respectively. The exact value for the latter follows the suggestion of Andrews and Shi (2013, p.625).

A.2 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 Subsection 3.3, are outlined here.

First, $P(t,r)$ 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_1,θ_0) arises only from the uncertainty in estimating $P(t=j,r=k)$.

Next, recall from Theorem 1, we test whether there exists a pair $(\theta_1,\theta_0)\in\Theta_P(s)$ such that $\theta_1+\theta_0=1$. To do this, we form bounds on the sum $\theta_1+\theta_0$ using the sharply partially identified set given by equations (3), (4), and (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 a dilation. 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 6 demonstrates that while the test provides adequate coverage, it exhibits significantly lower power for any given sample size n and design compared to our preferred approach based on subvector inference. This is because subvector inference more effectively exploits the structure of the problem. Our simulations also show that the test based on Goodman (1965) is substantially less computationally demanding than the inference procedure based on Bugni et al. (2017). We measure resource demand in terms of computation time, which might be relevant for large-sample application. Therefore, while the test based on Goodman (1965) achieves asymptotic coverage, our preferred inference procedure offers higher power and uniform asymptotic coverage despite being more resource-intensive.

B PROOFS FOR SECTION 2

Proposition 1. *Let Θ be a connected identified set for the performance measure of the index test. The index test is a dilation 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. We start with a preliminary observation: $\pi \leq v_1(\theta;\pi)$ and $\pi \geq v_0(\theta;\pi)$ if and only if $\theta_0 + \theta_1 \geq 1$ holds for any pre-test probability $\pi \in (0,1)$. 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). The conclusion follows now directly from the preliminary observation. □

Lemma 1. *Suppose $s \in [0,1]^2$ satisfies Assumption 1 and maintain Assumption 2. Then $\Theta_P(s)$—as defined in Equation 3—is non-empty. Furthermore, if Assumption 3 holds additionally, then $\Theta_P(s)$ is not a singleton set if and only if $s = (s_0, s_1) \neq (1, 1)$.*

Proof. Recall that under Assumptions 1 and 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:

$$
\begin{aligned}
P(t = 1, r = 0) - s_0 P_s(y = 0) &= (1 - s_1) P_s(y = 1) - P(t = 0, r = 0), \\
P(t = 1, r = 1) - (1 - s_0) P_s(y = 0) &= s_1 P_s(y = 1) - P(t = 0, r = 1), \\
P(t = 1, r = 1) - s_1 P_s(y = 1) &= (1 - s_0) P_s(y = 0) - P(t = 0, r = 1).
\end{aligned}
$$

Using these expressions together with definitions in Equation 4 and Equation 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 as argued in Remark 1, no matter whether Assumption 3 holds. Thus, it remains to show that if Assumption 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 the same as $|\Theta_P(s)| = 1$ 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 must be that $\min\{P(t = 1, r = 1) + s_0 P_s(y = 0), s_1 P_s(y$

0 which is not possible by Assumption 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 $-P(t=0,r=0)=\min\{P(t=1,r=1),s_1P_s(y=1))\}$, 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 $P(t=1,r=1)\leq s_1P_s(y=1)$. 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 2. Next, notice $\theta_1^L(s)=s_1P_s(y=1)-P(t=0,r=1)$ and let $P(t=1,r=1)>s_1P_s(y=1)$. 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 2, we have $s_1\geq P(r=1)$, and therefore $s_1\geq P(t=1,r=1)$. Hence, the case contradicts Assumption 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 $-P(t=0)=\min\{P(t=1,r=1)-s_1P_s(y=1),0\}=\min\{(1-s_0)P_s(y=0)-P(t=0,r=1),0\}$. 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.

□

Theorem 1. *Maintain Assumptions 1, 2 and 3, and let $\Theta_P(s)$ be the resulting identified set as in Equation 3. Then t is a dilation if and only if (1) $s\neq(1,1)$ and (2) there exists $\theta\in\Theta_P(s)$ such that $\theta_1+\theta_0=1$.*

Proof. By Proposition 1 and, in particular Corollary 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 $H(s)$ would be a singelton by Lemma 1, contradicting the existence of θ and θ', because they need to be different. 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 1, $\Theta_P(s)$ is a non-singelton, non-empty set. Furthermore, $\Theta_P(s)$ is a line segment as argued in Remark 1. It is then immediate, *cf.* Remark 4, that there exists $\theta'' \in \Theta : \theta''_1 + \theta''_0 = 1$.

For sufficiency, suppose that $s \neq (1,1)$ and there exists $\theta'' \in \Theta_P(s)$ with $\theta''_1 + \theta''_0 = 1$. By Lemma 1 there exists $\theta \in \Theta_P(s)$ such that $\theta \neq \theta''$. As discussed in Remark 1, $\Theta_P(s)$ 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. □

C Proofs for Section 3

Lemma 2. *If Assumption 1, Assumption 2' and Assumption 3' hold, then $\mathcal{P}$ is compact.*

Proof. 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' 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 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. □

C.1 Proof of Theorem 2

We will prove our Theorem 2, by means of an application of Bugni et al. (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 4 holds for all $P \in \mathcal{P}$, (ii) Assumption 1 holds, (iii) $\mathcal{P}$ satisfies Assumption 2' (and *a fortiori* Assumption 2), and (iii) $\mathcal{P}$ satisfies Assumption 3' (and *a fortiori* Assumption 3).

The following Lemmata verify Bugni et al. (2017, Definition 4.2).

Lemma A1. *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 A2. *There exists $\underline{\sigma}, \overline{\sigma} \in (0, \infty)$ 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 A1 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 2), there exists a uniform upper bound which is also finite. □

Lemma A3. *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 A2. □

Henceforth, define $\sigma_{P,j}(\theta; s) = \sqrt{\sigma^2_{P,j}(\theta; s)}$.

Lemma A4. *There exists a constant $a > 0$ such that for all $j = 1, \ldots, 7$, we have*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\sup_{\theta \in [0,1]^2} \left| \frac{m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)} \right|^{2+a} \right] < \infty \tag{11}$$

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 A1 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

$$\begin{aligned}\sup_{\theta \in [0,1]^2} \left| \frac{m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)} \right|^{2+a} &\leq \max_{(W',\theta,s) \in \{0,1\}^2 \times [0,1]^2} |M_j m_j(W', \theta; s)|^{2+a}, \\ &= \left|M_j m_j(W_j^*, \theta_j^*; s)\right|^{2+a} < \infty,\end{aligned}$$

where (W_j^*, θ_j^*) is an element of the arg max. Since the maximizer depends on j only, we conclude that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\sup_{s\in[0,1]^2} \left| \frac{m_j(W,\theta;s)}{\sigma_{P,j}(\theta;s)} \right|^{2+a} \right] \leq \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\left| M_j m_j(W_j^*, \theta_j^*; s) \right|^{2+a} \right]$$
$$= \left| M_j m_j(W_j^*, \theta_j^*; s) \right|^{2+a} < \infty.$$

□

For $i, j = 1, \ldots 7$, $P \in \mathcal{P}$, and $\theta, \theta' \in [0,1]^2$ define

$$\Omega_P(\theta,\theta')_{i,j} :=$$
$$\mathbb{E}_P \left[\left(\frac{m_i(W,\theta;s) - \mathbb{E}_P[m_i(W,\theta;s)]}{\sigma_{P,i}(\theta;s)} \right) \left(\frac{m_j(W,\theta';s) - \mathbb{E}_P[m_j(W,\theta';s)]}{\sigma_{P,j}(\theta';s)} \right) \right]$$

and let $\Omega_P(\theta,\theta')$ denote the 7×7 matrix with row $i = 1, \ldots, 7$ and column $j = 1, \ldots, 7$ given by $\Omega_P(\theta,\theta')_{i,j}$.

Lemma A5.

$$\lim_{\delta\downarrow 0} \sup_{\|(\theta,\theta')-(t,t')\|<\delta} \sup_{P\in\mathcal{P}} \|\Omega_P(\theta,\theta') - \Omega_P(t,t')\| = 0$$

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 A2 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 2; 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 $v_n(\theta) := \left(v_{n,j}\,(\theta)\right)_{j=1}^{7}$ 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 A6 (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 A1 and Assumption 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 A2 and Lemma A4. 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 Van der 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 Van der Vaart (2000, Theorem 19.5). □

Lemma A7 (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 A4.

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 A5 show the required continuity properties. This proves the lemma. □

Lemma A8. *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 A6 and Lemma A7 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 et al., 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 Subsection A.1 satisfies the needed properties as explained just after Equation (4.3) and Remark B.1 in Bugni et al. (2017).[17] Second, Assumption A.2 is not needed in our implementation because, as suggested by Bugni et al. (2017, Remark B.2), we adjusted the critical value by a small constant as mentioned in the last paragraph of Subsection A.1. Third, we verify their Assumption A.3, which requires the introduction of some further notation first.

$$\begin{aligned}
\mathcal{P}_0 &:= \{P \in \mathcal{P} : \Theta_0 \cap \Theta_P(s) \neq \emptyset\} \\
Q_P(\theta; s) &:= \sum_{j=1}^{6} \left[\min\left\{0, \frac{\mathbb{E}_P m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)}\right\}\right]^2 + \left[\frac{\mathbb{E}_P m_7(W, \theta; s)}{\sigma_{P,7}(\theta; s)}\right]^2 \\
g_{P,j}(\theta; s) &:= \frac{\mathbb{E}_P m_j(W, \theta; s)}{\sigma_{P,j}(\theta; s)} \\
\mathcal{P}_* &:= \{P \in \mathcal{P} : \Theta_P(s) \neq \emptyset\}
\end{aligned}$$

Assumption A.3. *The following conditions hold.*

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:*

$$\lim_{\delta \to 0} \sup_{P \in \mathcal{P}_*, (\theta, \theta') : ||\theta - \theta'|| \leq \delta} ||(\nabla g_{P,j})_{j=1}^7(\theta; s) - (\nabla g_{P,j})_{j=1}^7(\theta'; s)|| = 0.$$

[17] More formally, this follows from Bugni et al. (2015, Lemma D.9). See also Remark B.1 *ibidem.* Note that Bugni et al. (2017, Remark B.1) incorrectly refers to Lemma D.*8* of Bugni et al. (2015).

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 A9. *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 A2, 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 A10. *Assumption A.3(3) holds.*

Proof. By the same argument as in the proof of Lemma A6 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 2 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 et al. (2017, Theorem 4.1), which then proves our Theorem 2.

D Proofs for Section 5

Lemma A1. *The image of a path-connected space under a path-connected valued correspondence which admits a continuous selector is path-connected.*

Proof. Let $f : X \rightrightarrows Y$ be the correspondence with the properties stated.

If $f(X)$ is empty, the statement is vacuously true. Otherwise, take $y, y' \in f(X)$. By definition, there exist $x, x' \in X$ such that $y \in f(x)$ and $y' \in f(x')$. Since X is path-connected, there exists a path $p_X : [0,1] \to X$ with $p_X(0) = x$ and $p_X(1) = x'$. Furthermore, by assumption, there exists a continuous selection of f, which we denote by g. Then the composition $g \circ p_X : [0,1] \to Y$ is continuous. Additionally, since $f(x)$ and $f(x')$ are path-connected, we know there exists paths $p : [0,1] \to Y$ and $p' : [0,1] \to Y$ from y to $g(x)$ and y' to $g(x')$, respectively.

Now define $p^* : [0,1] \to Y$ as follows:

$$p^*(a) = \begin{cases} p(3a) & a \in [0, 1/3] \\ g(p_X(3a-1)) & a \in (1/3, 2/3) \\ p'(3-3a) & a \in [2/3, 1], \end{cases}$$

which is continous because

$$\lim_{a\searrow 1/3} p^*(a) = \lim_{a\searrow 1/3} g(p_X(3a-1)) = g(p_X(0)) = g(x) = p(1) = \lim_{a\nearrow 1/3} p^*(a)$$

and

$$\lim_{a\nearrow 2/3} p^*(a) = \lim_{a\nearrow 2/3} g(p_X(3a-1)) = g(p_X(1)) = g(x') = p'(1) = \lim_{a\searrow 2/3} p^*(a). \tag{12}$$

Furthermore, $p^*(0) = p(0) = y$ and $p^*(1) = p'(0) = y'$. Thus, p^* is a path from y to y'. □

Lemma 3. *If Assumption 1S holds, then $\Theta_P(\mathcal{S})$ is a path-connected set.*

Proof. Using the notation and the arguments just before the statement of Lemma 3 in the main text, $\Theta_P(\cdot)$ is a correspondence mapping a path-connected set (*cf.* Assumption 1S) into the unit square (with the usual topology). Furthermore, since $\Theta_P(s)$ is non-empty (*cf.* Lemma 1) and a line-segment for every $s \in \mathcal{S}$ (*cf.* Remark 1), the correspondence has path-connected values and the boundaries of these line segments ($\theta^L(\cdot)$ and $\theta^H(\cdot)$) are continuous selectors of $\Theta_P(\cdot)$. Finally, $\Theta_P(\mathcal{S})$ being path-connected follows from an application of Lemma A1. □

Lemma 4. *Suppose $\mathcal{S} \subset [0,1]^2$ satisfies Assumption 1S and maintain Assumption 2S. Then $\Theta_P(\mathcal{S})$—as defined above—is non-empty. Furthermore, if Assumption 3 holds additionally, then $\Theta_P(\mathcal{S})$ is not a singleton set if and only if $\mathcal{S} \neq \{(1,1)\}$.*

Proof. Fix $\mathcal{S} \subset [0,1]^2$ that satisfies Assumption 1S and note that for any $s \in \mathcal{S}$, Assumption 1 is applicable. Then, Lemma 1 gives that $\Theta_P(s) \neq \emptyset$. By definition $\Theta_P(s) \subseteq \Theta_P(\mathcal{S})$ and therefore non-emptiness carries over.

Now, assume that Assumption 3 holds too. If $\mathcal{S} = \{(1,1)\}$, then $\Theta_P(\mathcal{S})$ is a singleton by Lemma 1. If $\mathcal{S} \neq \{(1,1)\}$, then there exists $s \in \mathcal{S}$ such that $s \neq (1,1)$ and, again by Lemma 1, $\Theta_P(s)$ is not a singleton. Then, clearly $\Theta_P(\mathcal{S})$ is not a singelton either. □

Theorem 3. *Maintain Assumptions 1S, 2S and 3, and let $\Theta_P(\mathcal{S})$ be the resulting identified set. Then t is a dilation if and only if (1) $\mathcal{S} \neq \{(1,1)\}$ and (2) there exists $\theta \in \Theta_P(\mathcal{S})$ such that $\theta_1 + \theta_0 = 1$.*

Proof. Start with necessity. By Proposition 1 and, in particular Corollary 2, 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, $\Theta_P(\mathcal{S})$ is a path-connected set and therefore there is a path from θ to θ', which implies that there exists $\theta'' \in \Theta_P(\mathcal{S})$ such that $\theta''_1 + \theta''_0 = 1$ (*cf.* Remark 4). Furthermore, $\mathcal{S} \neq \{(1,1)\}$ holds, because if not[18] $\Theta_P(\mathcal{S})$ would be a singelton as argued in Remark 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 1 t is a dilation.

2. If $s = (1,1)$, then by hypothesis, there exists $s' \in \mathcal{S}$ with $s' \neq s$ and then Lemma 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 1.

□

Proposition 3.

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

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

[18] Note that $\mathcal{S}$ is non-empty by Assumption 1S.

Proof. In the following (1), (2), (3), and (4) refer to the inequalities indicated by the same numbers in the definition of $\mathcal{D}_P^{\llcorner}$.

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 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 2 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 2). (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

$$\begin{aligned} & s_1 \geq P(r=1) && \text{(by (3))} \\ \Longrightarrow\ & s_1 \geq P(t=0,r=1) \\ \Longleftrightarrow\ & P(t=0) \geq P(t=0,r=1) - s_1 + (1-\theta_1) && (\pm P(t=0)) \\ \Longleftrightarrow\ & P(t=0,r=0) \geq 1-\theta_1-s_1. \end{aligned}$$

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

$$\begin{aligned} & s_1 \geq P(r=1) && \text{(by (3))} \\ \Longrightarrow\ & s_1 \geq P(t=1, r=1) && \\ \Longleftrightarrow\ & P(t=1) \geq P(t=1, r=1) - s_1 + \theta_1 && (\pm P(t=1)) \\ \Longleftrightarrow\ & P(t=1, r=0) \geq \theta_1 - s_1. && \end{aligned}$$

□

References

Abakarim, Y., M. Lahby, and A. 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)*, IEEE, 306–313.

Ai, T., Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia (2020): "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology*, 296, E32–E40.

Alcoba-Florez, J., H. Gil-Campesino, D. G.-M. de Artola, R. González-Montelongo, A. Valenzuela-Fernández, L. Ciuffreda, and C. Flores (2020): "Sensitivity of different RT-qPCR solutions for SARS-CoV-2 detection," *International Journal of Infectious Diseases*, 99, 190–192.

Aliprantis, C. and K. Border (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer Berlin Heidelberg.

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

Andrews, D. W. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157.

Arevalo-Rodriguez, I., D. Buitrago-Garcia, D. Simancas-Racines, P. Zambrano-Achig, R. Del Campo, A. Ciapponi, O. Sued, L. Martinez-Garcia, A. W. Rutjes, N. Low, et al. (2020): "False-negative results of initial RT-PCR assays for COVID-19: a systematic review," *PloS one*, 15, e0242958.

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

Binney, N., C. Hyde, and P. M. Bossuyt (2021): "On the Origin of Sensitivity and Specificity," *Annals of Internal Medicine*, 174, 401–407.

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

Bradley, S. (2019): "Imprecise Probabilities," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, Metaphysics Research Lab, Stanford University, Spring 2019 ed.

Buck, A. A. and J. 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, 586–592.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185, 259–282.

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8, 1–38.

Canay, I. A., G. Illanes, and A. Velez (2023): "A User's guide for inference in models defined by moment inequalities," *Journal of Econometrics*, 105558.

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

Deneef, P. (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, 92–96.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2022): "Minimum distance belief updating with general information," Tech. rep., Working paper.

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

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

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

Gart, J. J. and A. 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, 593–602.

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

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

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

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

Groemping, U. (2019): “South german credit data: Correcting a widely used data set,” *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, 4, 2019.

Herron, T., T. Seidenfeld, and L. Wasserman (1997): “Divisive conditioning: further results on dilation,” *Philosophy of Science*, 64, 411–444.

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

Hui, S. L. and X. H. Zhou (1998): “Evaluation of diagnostic tests without gold standards,” *Statistical methods in medical research*, 7, 354–370.

Kanji, J. N., N. Zelyas, C. MacDonald, K. Pabbaraju, M. N. Khan, A. Prasad, J. Hu, M. Diggle, B. M. Berenger, and G. Tipples (2021): “False negative rate of COVID-19 PCR testing: a discordant testing analysis,” *Virology journal*, 18, 1–6.

Kellner, C., M. T. Le Quement, and G. Riener (2022): “Reacting to ambiguous messages: An experimental analysis,” *Games and Economic Behavior*, 136, 360–378.

Kops, C. and I. Pasichnichenko (2023): “Testing negative value of information and ambiguity aversion,” *Journal of Economic Theory*, 105730.

Liang, Y. (2024): “Learning from unknown information sources,” *Management Science*.

Lin, Y.-H. and F. Payró (2024): “Updating Under Imprecise Information,” Tech. rep., Working paper.

Manski, C. F. (2007): *Identification for Prediction and Decision*, Harvard University Press.

——— (2018): “Credible ecological inference for medical decisions with personalized risk assessment,” *Quantitative Economics*, 9, 541–569.

——— (2020): "Bounding the accuracy of diagnostic tests, with application to COVID-19 antibody tests," *Epidemiology*, 32, 162–167.

——— (2021): "Bounding the accuracy of diagnostic tests, with application to COVID-19 antibody tests," *Epidemiology*, 32, 162–167.

MANSKI, C. F. AND J. V. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

MEI, X., H.-C. LEE, K.-Y. DIAO, M. HUANG, B. LIN, C. LIU, Z. XIE, Y. MA, P. M. ROBSON, M. CHUNG, ET AL. (2020): "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nature medicine*, 26, 1224–1228.

MOLINARI, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144, 81–117.

MULHERIN, S. A. AND W. C. MILLER (2002): "Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation," *Annals of internal medicine*, 137, 598–602.

OBRADOVIĆ, F. (2024): "Measuring diagnostic test performance using imperfect reference tests: A partial identification approach," *Journal of Econometrics*, 244, 105842.

PACHECO PIRES, C. (2002): "A rule for updating ambiguous beliefs," *Theory and Decision*, 53, 137–152.

ROGAN, W. J. AND B. GLADEN (1978): "Estimating prevalence from the results of a screening test," *American journal of epidemiology*, 107, 71–76.

ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2014): "A practical two-step method for testing moment inequalities," *Econometrica*, 82, 1979–2002.

SACKS, D. W., N. MENACHEMI, P. EMBI, AND C. WING (2022): "What can we learn about SARS-CoV-2 prevalence from testing and hospital data?" *Review of Economics and Statistics*, 1–36.

SEIDENFELD, T. AND L. WASSERMAN (1993): "Dilation for sets of probabilities," *The Annals of Statistics*, 21, 1139–1154.

SHISHKIN, D. AND P. ORTOLEVA (2023): "Ambiguous information and dilation: An experiment," *Journal of Economic Theory*, 208, 105610.

STAQUET, M., M. ROZENCWEIG, Y. J. LEE, AND F. M. MUGGIA (1981): "Methodology for the assessment of new dichotomous diagnostic tests," *Journal of chronic diseases*, 34, 599–610.

STOYE, J. (2022): "Bounding infection prevalence by bounding selectivity and accuracy of tests: with application to early COVID-19," *The Econometrics Journal*, 25, 1–14.

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

THIBODEAU, L. (1981): "Evaluating diagnostic tests," *Biometrics*, 801–804.

TOULIS, P. (2021): "Estimation of COVID-19 prevalence from serology tests: A partial identification approach," *Journal of Econometrics*, 220, 193–213.

VACEK, P. M. (1985): "The effect of conditional dependence on the evaluation of diagnostic tests," *Biometrics*, 959–968.

VALENSTEIN, P. N. (1990): "Evaluating diagnostic tests with imperfect standards," *American Journal of Clinical Pathology*, 93, 252–258.

VAN DER VAART, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

VAN DER VAART, A. W. AND J. A. WELLNER (1997): *Weak convergence and empirical processes: with applications to statistics*, Springer New York.

WALLEY, P. (1991): *Statistical reasoning with imprecise probabilities*, vol. 42, Springer.

WATSON, J., P. F. WHITING, AND J. E. BRUSH (2020): "Interpreting a covid-19 test result," *Bmj*, 369.

WILLIS, B. H. (2008): "Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies," *Family Practice*, 25, 390–396.

Yerushalmy, J. (1947): "Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques," *Public Health Reports (1896-1970)*, 1432–1449.

Zhou, X.-H., D. K. McClish, and N. A. Obuchowski (2009): *Statistical methods in diagnostic medicine*, vol. 569, John Wiley & Sons.

Ziegler, G. (2021): "Binary Classification Tests, Imperfect Standards, and Ambiguous Information," *arXiv preprint arXiv:2012.11215.*