Robust Combination Testing: Methods and Application to COVID-19 Detection

Sanjay Jain
Department of Economics, University of Oxford
sanjay.jain@economics.ox.ac.uk

Jónas Oddur Jónasson
MIT Sloan School of Management
joj@mit.edu

Jean Pauphilet
Management Science and Operations, London Business School
jpauphilet@london.edu

Kamalini Ramdas
Management Science and Operations, London Business School
kramdas@london.edu

Simple and affordable testing tools are often not accurate enough to be operationally relevant. For COVID-19 detection, rapid point-of-care tests are cheap and provide results in minutes, but largely fail policymakers accuracy requirements. We propose an analytical methodology, based on robust optimization, that identifies optimal combinations of results from cheap tests for increased predictive accuracy. This methodological tool allows policymakers to credibly quantify the benefits from combination testing and thus break the trade-off between cost and accuracy. Our methodology is robust to noisy and partially missing input data and incorporates operational constraints—relevant considerations in practice. We apply our methodology to two datasets containing individual-level results of multiple COVID-19 rapid antibody and antigen tests, respectively, to generate Pareto-dominating receiver operating characteristic (ROC) curves. We find that combining only three rapid tests increases out-of-sample area under the curve (AUC) by 4% (6%) compared with the best performing individual test for antibody (antigen) detection. We also find that a policymaker who requires a specificity of at least 0.95 can improve sensitivity by 8% and 2% for antibody and antigen testing, respectively, relative to available combination testing heuristics. Our numerical analysis demonstrates that robust optimization is a powerful tool to avoid overfitting, accommodate missing data, and improve out-of-sample performance. Based on our analytical and empirical results, policymakers should consider approving and deploying a curated combination of cheap point-of-care tests in settings where 'gold standard' tests are too expensive or too slow.

Key words: Diagnostic operations; Combination testing; Knapsack; Robust optimization; Healthcare analytics.

History: January 23, 2023

1. Introduction

In diagnostic testing, there is a fundamental trade-off between diagnostic accuracy on the one hand, and testing cost and speed on the other hand. For COVID-19, for instance, the laboratory-based 'gold standard' Reverse Transcription Polymerase Chain Reaction (RT-PCR) test for detection of current infection requires expensive lab capacity and a supply chain connecting testing locations and labs, which increases both processing time and costs. In contrast, point-of-care (POC) lateral flow antigen tests are cheap and provide results in minutes, but largely fail policymakers accuracy criteria (Wise 2020). The same dilemma is faced in the detection of acquired immunity after infection, where lateral flow antibody tests offer lower accuracy than lab tests. This drawback is exacerbated in the early phases of spread of a novel pathogen, when the first tests available may be of lower quality. The inferior quality of POC tests has greatly limited their use for COVID-19 (Shuren and Stenzel 2021) and has resulted in millions wasted on purchasing tests subsequently deemed unfit for use (Armstrong 2020).

Despite these shortcomings, experts argue that frequent, fast, and large-scale antigen testing schemes, for which RT-PCR tests are clearly unsuitable, should be implemented to enable the safe re-opening of economic activity (Mina et al. 2020, Larremore et al. 2021), especially in use cases involving entry into public spaces such as airports, care homes, universities, and entertainment or sporting venues. Similarly, efficient and widespread antibody testing is useful to assess global immunity and prioritize vaccination delivery (Bubar et al. 2021). Beyond the specific case of COVID-19 diagnostics, the usefulness of combining multiple, cheap diagnostic options to improve overall accuracy extends to other healthcare contexts, as well as to applications such as maintenance or recruitment, where results of multiple diagnostic or classification methods are aggregated to make a single recommendation.

In this paper, we propose an optimization-based methodology to combine the results of different POC tests to achieve higher accuracy without compromising affordability or convenience. Intuitively, there are potentially both systematic and stochastic mechanisms that motivate the idea of combining results obtained from multiple tests. First, different tests can use different technologies or detection strategies—e.g., the target protein, chemical reagents used, or how the tests are administered may differ. In this case, each test provides a different perspective on a patient's underlying condition and combining them will lead to a more comprehensive and accurate picture. Second, test results are inherently random due to uncertainty in test administration, chemical processes, and unobserved patient characteristics. In this case, if the correlation between test outcomes is low, combining them can help reduce noise and improve accuracy. For both these reasons, we expect that combining tests can improve accuracy. For a stylized example of the systematic or stochastic mechanisms driving the benefits of combining tests, see Appendix A1.

Any methodology for combining multiple test outcomes must overcome three fundamental challenges in the design of combination testing policies. The first is how to interpret an *outcome sequence* of discordant yet potentially correlated test results—e.g., $(+,-,+)$ in the simple case of combining three tests—to classify a person as positive or negative. Heuristics, such as classifying an outcome sequence as positive only if all the test outcomes are positive, are simple but can have suboptimal performance.[1] The second challenge is to select which tests to combine, when multiple tests are available. A naïve approach of combining the tests with highest individual accuracy may result in suboptimal outcomes if their results are not complementary when combined. Test selection must also account for operational and budget constraints, affecting how many tests can feasibly be combined in a given setting. The third challenge is ensuring an empirical evidence base for the proposed combination testing policy—by using available data on the performance of individual tests—while avoiding overfitting, which would result in poor out-of-sample classification and therefore limit applicability.

Analytically, we formulate the policymaker's problem as one of selecting and combining a subset of available POC tests to maximize diagnostic accuracy, measured by sensitivity and specificity – i.e., the probability that a condition-positive or condition-negative individual is classified as positive or negative, respectively. First, for a given set of tests that are used in combination, we prove that a simple ranking procedure identifies all classification rules with Pareto-optimal accuracy (§3.1), by formulating this problem as a version of the well-studied knapsack problem. The resulting set of optimal classification rules yields a dominating receiver operating characteristic (ROC) curve – with the highest attainable area under the curve (AUC). Of potentially independent interest, our proof provides an optimization-based derivation of the Neyman-Pearson lemma (McIntosh and Pepe 2002). Note that our approach is non-parametric and does not require any assumptions about the correlation structure among tests results. Second, we extend our formulation to incorporate the selection of tests to be included in the combination, while accounting for the cost of adding each test and the policymaker's overall budget (§3.2). Third, we present a robust variant of the model to account for noise in the input parameter estimates, such as the sensitivity and specificity of individual POC tests (§3.3). Indeed, in practice, the data available for calibrating these parameters is often limited, leading to finite-sample estimation error. For emergency use authorization, for instance, manufacturers can use as few as 30 positive samples (Fitzpatrick et al. 2021), which can result in overoptimistic estimates of test performance (Armstrong 2020). Our robust model

[1] We note that standard Bayesian approaches might be used to interpret discordant outcomes in cases where priors are available and the testing accuracy (sensitivity and specificity) are known for each test. However, such approaches would not apply in settings where prevalence and accuracy must be estimated from noisy data, and would not allow for efficiently optimizing which tests to combine.

prevents such overfitting, and we provide theoretical guarantees on its out-of-sample diagnostic performance. Fourth, data gathered for manufacturer-independent head-to-head comparisons of tests often contains a significant number of missing entries, because all tests are not necessarily applied on all samples. Importantly, we extend our robust formulation so that it can be directly applied on data with missing entries (§3.4).

Empirically, we apply the four steps of our methodology to two datasets shared with us by medical research institutions in the U.K. and Germany, which record the performance of multiple commercially available point-of-care antibody (Adams et al. 2020) and antigen (Corman et al. 2021) tests for COVID-19 (§5.1–§5.4). Through this analysis we make several contributions. First, we demonstrate that optimally combining multiple tests using our robust optimization method improves diagnostic accuracy compared to using single tests or standard heuristics for combining tests. Notably, we find that combining three tests achieves a 4% (6%) relative improvement in out-of-sample AUC for antibody (antigen) testing, compared to using the best individual test. We also find that if a policymaker requires specificity of at least 0.95, our methodology achieves 8% and 2% relative improvements in sensitivity for antibody and antigen detection, respectively, compared with standard heuristics. Second, our analysis demonstrates that with limited available data the risk of overfitting is significant, an issue that is efficiently mitigated by our robust formulation. Similarly, regarding missing data, incorporating an imputation step into our robust combination algorithm substantially outperforms policies generated using maximum likelihood imputation in out-of-sample comparisons. Finally, we provide an illustrative example of how a decision maker who is responsible for designing testing protocols in a given community (e.g., for managing access to public spaces or outbreak detection) would apply our methodology to develop a testing policy, even in the realistic situation where data is limited.

Our work has direct and immediate policy implications. We introduce an actionable method for selecting and combining POC tests to achieve superior diagnostic accuracy. For detecting COVID-19 infection, a combination of two or three POC tests is likely to be cheaper and faster than RT-PCR, which requires both a lab infrastructure and a sample transportation supply chain (Böger et al. 2021, Rubin 2020). Our method offers policymakers a tool to credibly quantify the benefits from combination testing and quickly identify accurate combinations of tests if they exist. In addition, our combination testing approach delivers an entire ROC curve, which allows policymakers to emphasize either sensitivity or specificity using the same combination of tests, depending on their relative priorities for a particular population. Given our results, policymakers should consider approving and deploying combinations of tests in settings where RT-PCR testing is too costly or too slow, but individual POC tests do not provide enough diagnostic accuracy. Although directly implementable by performing multiple tests made by different manufacturers, combination testing

can also be implemented by creating a single diagnostic device that combines results from multiple tests 'under the hood'.

From a research standpoint, we believe this is the first paper to explore the value of optimally combining multiple diagnostic tests, through analytical modeling and empirical analysis. The medical literature has considered simple heuristics for combining a given set of tests (Marshall 1989, Sox et al. 2013), which can be sub-optimal when combining even as few as three tests. Our method not only identifies *how to* combine tests but also *which* subset of tests to combine, without conducting an exhaustive search. Furthermore, we provide much needed empirical evidence on the potential benefits of combination testing. While we illustrate our findings on data from COVID-19 testing, our model development is general and provides an application-agnostic methodology for combining the results of multiple tests. Therefore, our methodology can be applied in any setting where data on different classifiers is collected on the same subject, be it for HR recruitment purposes or machine maintenance. Our analysis demonstrates the value of structuring data collection such that it allows for both head-to-head comparison of different tests as well as the development of combination testing policies.

2. Literature Review

Our key contribution is to the ongoing debate on how to expand COVID-19 testing in a cost-effective way (§2.1). Our analysis also contributes to the operations management literature on diagnostics, clinical decision making, and operational innovations to battle COVID-19 (§2.2). More generally, our analysis relates to the machine learning literature on combining predictions for improved classification accuracy (§2.3) as well as prior work on the predictive power of using multiple tests in the context of test sequencing (§2.4).

2.1. COVID-19 Testing

The 'gold standard' test for COVID-19, RT-PCR (Böger et al. 2021), requires laboratory equipment, which increases the cost per test and induces sample-to-results delay of the order of 24-48 hours (Larremore et al. 2021). In addition, molecular laboratories in some countries have sometimes stopped accepting samples for RT-PCR tests during demand surges, due to manpower shortages (The Economic Times 2021).

At the same time, a plethora of POC tests have been developed for COVID-19 testing, both for antibody and antigen identification. In fact, the U.S. Food & Drug Administration (FDA) has approved over 200 different POC diagnostic tests for COVID-19 since the start of the pandemic (FDA 2020). The majority of POC tests for COVID-19 rely on lateral flow technology. Cost estimates for this type of POC tests (often around $6) vary across studies and range from one-third to one-tenth the cost of an RT-PCR test (Böger et al. 2021, Rubin 2020). In addition, some progress

has also been made in bringing the RT-PCR technology closer to the point of care. For example, portable devices that achieve high sensitivity have been deployed at hospitals, to alleviate pressure on laboratories (Gibani et al. 2020). However, these devices do not eliminate the need for expensive reagents and often need to be operated by professionals (Gupta et al. 2021). The methods presented in this paper would apply to any diagnostic technology.

Governments have committed significant funds to expand diagnostic testing during COVID-19, including POC testing. The U.S. has pledged $50 billion for this purpose (Politico 2021) and the U.K. has allocated 22 billion to its national testing program (National Audit Office 2020). However, the use of POC tests in the battle against COVID-19 is widely debated. Deeks et al. (2021) argue that POC testing can be useful, but only if highly accurate. They summarize trial results from the U.K. and conclude that the potential harm of false negatives outweighs the potential benefits. In contrast, Mina et al. (2020) argue in favor of POC test deployment, particularly for routine testing of asymptomatic individuals, to maximize the overall sensitivity of the testing regime, despite potentially low sensitivity for each individual.

Our methodological contribution and its empirical validation present a way to improve the accuracy of POC tests in a cost-efficient manner. Our data-driven approach reveals that the combination of even two POC tests can significantly improve diagnostic accuracy over that of an individual POC test, while remaining much faster and cheaper than RT-PCR testing.

It is worth noting that the overall cost of a test depends on a number of factors including volume ordered, public subsidies, method of administration, and market competition. For RT-PCR tests, for example, there is a wide variability in prices within and between countries (International Air Transport Association 2021). There is also a lack of reliable and objective data on test costs. In a comparison of test prices among 566 top U.S. hospitals, for example, Xiao and Rathi (2021) note that two thirds disclosed none.

2.2. Diagnostic Operations and Clinical Decision Making

The clinical decision making literature has explored the utility and cost-effectiveness of conducting additional tests (Weinstein and Fineberg 1980, Hershey et al. 1986, Sox et al. 2013), by considering simple rules of thumb, e.g., classifying an outcome as positive if any test is positive (the `any` rule), if all tests are positive (the `all` rule), or if the majority of tests is positive (the `majority` rule). This literature examines the net utility from choosing each of the available courses of action by taking as inputs a set of available tests; the diagnostic accuracy of each test; a set of alternative decisions (e.g., "conduct two tests in parallel and treat if both are positive"); and the utility associated with each course of action. However, no prior work provides a systematic methodology for classifying the outcome sequences of m tests for best achievable sensitivity and specificity. A key contribution

of our work is to develop such a method and demonstrate its effectiveness on real data. Using our method, we identify the set of classification rules that dominate all other rules in terms of sensitivity and specificity, for a given combination of tests. This set of best achievable classification rules for a given set of tests can be summarized in a dominating ROC curve.

In the operations management literature, the value of increasing testing accuracy is ambiguous. While performing additional tests can improve diagnostic accuracy, hence benefiting downstream decision making, it can also generate additional delays. Early modeling work in the context of maintenance combines maintenance interventions with inspections to account for stochastically failing equipment (McCall 1965, Wang 2002). Alizamir et al. (2013) analyze the trade-off between improving testing accuracy (e.g., by performing additional examinations) and treatment delays. They present an optimal policy which depends on prior beliefs about the patient type and the congestion level of the system. Kremer and de Véricourt (2022) conduct controlled laboratory experiments to assess decision makers' biases when balancing accuracy and congestion. Linking operational implementation and testing, Levi et al. (2019) study the scheduling trade-off between task diagnosis and task processing and Aprahamian et al. (2019) develop an optimal scheme for testing groups of patients using an imperfect single test, assuming subject-specific risk characteristics.

Recent papers have examined the operational implementation of testing to curb the spread of COVID-19. Several have examined allocation of limited testing capacity – e.g., to health centers in the Philippines (Buhat et al. 2021), patients with different probability of being infected (Calabrese and Demers 2022), or to balance the immediate benefit of diagnosing infected individuals with the benefit of learning population-level infection rates (Kasy and Teytelboym 2020). Bastani et al. (2021) propose and implement a bandit framework to specifically target the allocation of tests to incoming international travelers at the Greek border. Yang et al. (2022) develop a game-theoretic queueing model to identify improved scheduling and pricing policies for COVID-19 testing facilities. Pulia et al. (2020) propose a multi-tiered sequential testing policy in which patients are first triaged at the point of care using rapid antigen tests, before using RT-PCR technology for diagnostic confirmation. Drakopoulos and Randhawa (2021) consider tests as information products with limited availability and demonstrate that if targeted allocation of tests is impossible, moderately accurate tests are preferable in terms of social outcome. Dai and Singh (2021) explore the risk of overtesting in situations where laboratories can decide on the diagnostic criterion (which controls sensitivity) to apply for RT-PCR testing. These papers focus on the operational implementation of a given testing policy with a given diagnostic accuracy. In contrast, we propose a methodology for improving diagnostic accuracy.

In a conceptual article, Ramdas et al. (2020) suggest using multiple tests with independent outcomes to increase the accuracy of COVID-19 testing. In this paper, we develop the methodology

required to analyze data on the comparative performance of tests with correlated outcomes and identify powerful and cost-effective combination strategies using data from two medical institutions.

2.3. Combining Predictions for Improved Accuracy

Our analysis relates to aggregation methods in machine learning—particularly *stacking* and *bagging*. The idea of combining predictions from different models, originally proposed by Wolpert (1992) and referred to as *stacking* (see Džeroski and Ženko 2004, for a review), is widely used for binary classification. For instance, Barreno et al. (2007) leverage the Neyman-Pearson lemma (Neyman and Pearson 1933) to optimally combine binary predictions obtained from different classifiers, by ranking outcome sequences by their likelihood ratios. In medical research, and cancer screening in particular, the lemma has been applied to improve diagnostic accuracy by combining different biomarkers (McIntosh and Pepe 2002). Our method recovers the likelihood ratio ranking rule, using an optimization rather than statistical lens. In practice, however, each test is costly and combining all of them would be impractical. Our optimization perspective naturally accounts for such operational constraints and enables identification of the best subset of tests to be combined together, without exhaustive enumeration.

Breiman (1996) proposes bootstrap aggregation or *bagging*, in which weak classification algorithms (referred to as base learners) are trained on bootstrap samples of training data and then aggregated into a single prediction (usually via a heuristic decision rule such as classifying according to the majority of base learners or averaging across them). Our work shares the same aggregation flavor. However, unlike bagging, our base learners are tests developed by different manufacturers and not obtained on bootstrap samples. Hence, we learn the optimal aggregation policy directly from data on test performance, without any knowledge about the data generating process for each individual index test.

Finally, overfitting and out-of-sample performance are important practical issues in our setting, which the standard Neyman-Pearson lemma and the likelihood ratio ranking rule (McIntosh and Pepe 2002, Barreno et al. 2007) do not account for. In statistics, robust versions of the lemma have been developed to account for distribution misspecification (Huber 1965, Huber and Strassen 1973), but have not been used in the context of combining tests. We formulate a robust knapsack optimization problem to directly mitigate the impact of finite sample estimation error and missing data on combination testing, while accounting for operational constraints. Compared with previous work on robust knapsack problems with an uncertain objective (Monaci and Pferschy 2013, Cheng et al. 2014), uncertainty affects both the objective and the constraints of our problem.

2.4. Sequencing of tests

In medical decision making, there is a broad literature on contingent testing strategies. In this approach patients are screened with a first test. If the result is positive, then the final diagnosis is confirmed with a second test. These sequencing strategies have been studied, for instance, in the context of mammography examinations (Alagoz et al. 2013) and COVID-19 (Smith et al. 2021, Paltiel et al. 2020). For this to be useful in practice, the second (confirmatory) test should be more reliable than the first one, which usually means that it is also more expensive or invasive. In the context of COVID-19 detection, this would correspond to first performing a rapid antigen test, and confirming any positive result with RT-PCR, e.g., as piloted in the UK in Liverpool (Wise 2020).

One insight from this literature is that the first-step 'screening' test needs to be accurate enough, for these policies to be effective. A low false negative rate ensures that condition-positive individuals are detected. A low false positive rate ensures that the second (expensive) test is not wasted on many condition-negative individuals. However, with COVID-19 antigen tests, many have argued that they do not meet such standards (Wise 2020). With regard to this literature, our key contribution is to propose a solution to improve the accuracy of the first-step test, with little compromise on their price so that such contingent testing strategies can be effective in practice. We also believe that there are use cases, such as entry to public places, where one cannot afford a second-step confirmation test because of monetary or time considerations. Therefore, improving the accuracy of cheap and fast tests is of independent interest.

3. Model

In this section, we present a methodology for combining multiple tests for improved diagnostic accuracy. We first demonstrate that the problem of optimally classifying outcome sequences from multiple tests can be formulated as a knapsack problem (§3.1). Under this lens, we provide a new proof of the Neyman-Pearson lemma from first principles. We then extend the knapsack model for a given set of tests to the problem of selecting which tests to include in the combination, subject to a budget constraint (§3.2). Finally, we present a robust version of the model to account for in-sample estimation error (§3.3) and missing data (§3.4). Additionally, Appendix A3 provides a detailed optimization formulation for how to sequence tests, once the policy has been determined.

In the following, we use lowercase non-bold faced (x), lowercase bold faced ($\boldsymbol{x}$), uppercase bold faced ($\boldsymbol{X}$), and uppercase script faced ($\mathcal{X}$) characters to denote scalars, vectors, matrices, and sets respectively. We use hats to denote sample estimates of population parameters. For $\boldsymbol{x} \in \mathbb{R}^n$, its ith component is denoted x_i and $\boldsymbol{e}$ is the vector of all ones.

3.1. A Knapsack Perspective on Classifying Outcome Sequences

Let $\boldsymbol{s}$ denote an outcome sequence arising from performing m tests. For instance, when combining 2 tests, $\boldsymbol{s}$ can be one of $2^2 = 4$ possibilities: $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$. Let $\mathcal{S}_m$ denote the set of all possible outcome sequences given m tests and let $n = |\mathcal{S}_m| = 2^m$. We identify $\mathcal{S}_m$ as $\{1,\ldots,n\}$.

Let $\hat{a}_{\boldsymbol{s}}$ ($\hat{b}_{\boldsymbol{s}}$) denote the fraction of condition-positive (condition-negative) individuals in the data sample with outcome sequence $\boldsymbol{s}$. Observe that $\hat{\boldsymbol{a}}$ ($\hat{\boldsymbol{b}}$) has $n = 2^m$ coordinates and fully characterizes the joint distribution of the m test results on the condition-positive (condition-negative) individuals in a non-parametric manner. A classification rule is a vector $\boldsymbol{x} \in [0,1]^n$, where $x_{\boldsymbol{s}}$ indicates the fraction of samples with outcome sequence $\boldsymbol{s}$ classified as *positive*. If $x_{\boldsymbol{s}} = 1$ (resp. 0), the rule classifies outcome sequence $\boldsymbol{s}$ as *positive* (resp. *negative*). For instance, in the case of two tests, the `and` rule, in which only individuals with both tests positive are classified as positive, corresponds to

$$x_{(1,1)} = 1, \quad \text{and } x_{\boldsymbol{s}} = 0, \text{ for } \boldsymbol{s} \in \{(0,0),(1,0),(0,1)\}.$$

We use the term *classification* rule for $\boldsymbol{x}$ on purpose since the problem can be viewed as a binary classification problem of predicting the underlying condition given the outcome sequence $\boldsymbol{s}$. For a given rule $\boldsymbol{x}$, its sensitivity is given by $\sum_{\boldsymbol{s}} \hat{a}_{\boldsymbol{s}} x_{\boldsymbol{s}}$ and its specificity by $1 - \sum_{\boldsymbol{s}} \hat{b}_{\boldsymbol{s}} x_{\boldsymbol{s}}$. Given a target specificity level $t_0 \in [0,1]$, the rule $\boldsymbol{x}$ that maximizes sensitivity while achieving a specificity of at least t_0 is the solution of the knapsack problem:

$$\max_{\boldsymbol{x} \in [0,1]^n} \hat{\boldsymbol{a}}^\top \boldsymbol{x} \text{ s.t. } \hat{\boldsymbol{b}}^\top \boldsymbol{x} \leq (1 - t_0). \tag{1}$$

From knapsack theory (Dantzig 1957), we immediately derive the optimal solution of (1):

LEMMA 1. *The solution obtained by sorting outcome sequences by decreasing values of $\hat{a}_{\boldsymbol{s}}/\hat{b}_{\boldsymbol{s}}$ and greedily filling the knapsack is optimal for* (1).

In the context of disease screening, the Neyman-Pearson lemma implies that ordering outcome sequences according to their likelihood ratios is optimal (McIntosh and Pepe 2002). Indeed, consider a person whose outcome sequence is $\boldsymbol{s}$ and the null hypothesis that this person is condition-negative. In this case, setting $x_{\boldsymbol{s}} = 1$ is analogous to rejecting the null hypothesis. The Neyman-Pearson lemma states that the likelihood ratio rule, which rejects the null hypothesis whenever the likelihood ratio $\hat{a}_s/\hat{b}_s$ exceeds some desired threshold, is the uniformly most powerful test. In particular, this ordering and the resulting optimal classification rule do not depend on the prevalence of the disease. Lemma 1 recovers this simple result (see McIntosh and Pepe 2002, Section 3), yet by invoking tools from optimization rather than statistics. This new perspective lends itself to natural extensions

to account for operational and statistical estimation issues that we will present in the rest of this section.

In Problem (1), we did not constrain the components of $\boldsymbol{x}$ to be binary. If $x_{\boldsymbol{s}} \in (0,1)$, the rule is no longer deterministic and a person with outcome $\boldsymbol{s}$ will be classified as *positive* with probability $x_{\boldsymbol{s}}$. However, Lemma 1 shows that there always exists an optimal solution with at most one fractional entry. Also, for certain specific values of t_0, the optimal rule is integral. In the remainder of the paper, we focus our attention on binary vectors $\boldsymbol{x} \in \{0,1\}^n$, i.e., deterministic mappings from sequences $\boldsymbol{s} \in \mathcal{S}_m$ to a positive/negative classification. Indeed, probabilistic classification rules, while theoretically feasible, could prove harder to implement in practice.

By solving (1) for different values of t_0 (as described in Algorithm 1 in Appendix A5), we construct a piecewise linear ROC curve that represents the set of Pareto-dominating classification rules.

3.2. Accounting for Operational Constraints

We leverage the knapsack optimization perspective from §3.1 to extend the analysis to account for cost and operational constraints. To do so, we introduce a binary variable $y_k \in \{0,1\}$ indicating whether test $k = 1, \dots, m$ is used by the classification rule $\boldsymbol{x}$. Budget constraints can then be expressed linearly as $\sum_{k=1}^m c_k y_k \leq K$, where c_k is the cost of each individual test k and K is the per-person total allowable budget.

We capture the logical relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ via linear constraints. For every outcome sequence $\boldsymbol{s}$, denote by $\bar{\boldsymbol{s}}^k$ the outcome sequence obtained by switching the result of test k, i.e, $\bar{s}^k_j = s_j$, for $j \neq k$, and $\bar{s}^k_k = 1 - s_k$. For a given classification rule $\boldsymbol{x}$, if $|x_{\boldsymbol{s}} - x_{\bar{\boldsymbol{s}}^k}| = 0$, then, the results of the $m-1$ other tests being equal, the result of test k does not impact the classification. If this is the case for all sequences $\boldsymbol{s}$, then test k is not used, and we should have $y_k = 0$. In other words, we have the following constraints:

$$\frac{1}{n} \sum_{\boldsymbol{s} \in \mathcal{S}_m} |x_{\boldsymbol{s}} - x_{\bar{\boldsymbol{s}}^k}| \leq y_k \leq \sum_{\boldsymbol{s} \in \mathcal{S}_m} |x_{\boldsymbol{s}} - x_{\bar{\boldsymbol{s}}^k}|.$$

The first inequality enforces the constraint "$y_k = 0 \implies \forall \boldsymbol{s} \in \mathcal{S}_m, x_{\boldsymbol{s}} = x_{\bar{\boldsymbol{s}}^k}$", while the second enforces the reverse implication. As a result, we consider combination rules in the following set:

$$\mathcal{X} := \left\{ \boldsymbol{x} \in \{0,1\}^n \;\middle|\; \exists \boldsymbol{y} \in \{0,1\}^m, \text{ s.t. } \begin{array}{c} \boldsymbol{c}^\top \boldsymbol{y} \leq K, \\ \frac{1}{n} \sum_{\boldsymbol{s} \in \mathcal{S}_m} |x_{\boldsymbol{s}} - x_{\bar{\boldsymbol{s}}^k}| \leq y_k \leq \sum_{\boldsymbol{s} \in \mathcal{S}_m} |x_{\boldsymbol{s}} - x_{\bar{\boldsymbol{s}}^k}|, \forall k \end{array} \right\}.$$

Additional constraints can be added to the definition of $\mathcal{X}$ to account for multiple resources or enforce some structure on the classification rule $\boldsymbol{x}$. For instance, in Appendix A2, we propose adding linear constraints on $\boldsymbol{x}$ and $\boldsymbol{y}$ to force the aggregation rule $\boldsymbol{x}$ to correspond to one of the

available heuristic rules (`any`, `all`, or `majority` rule). Tests might also follow different protocols (e.g., saliva tests vs. nasal swab) and we can add constraints on $\boldsymbol{y}$ to capture patient preferences (e.g., a limit on the number of nasal swabs required).

The above modeling of testing costs allows for heterogeneous costs across tests to be included in the analysis. For simplicity and due to the absence of reliable data on test costs (Xiao and Rathi 2021), we will set $c_k = 1$ in the implementation of the approach for our empirical results. Setting $c_k = 1$ for all k restricts the number of tests being used to at most K. As a result, our empirical analysis focuses on simultaneously selecting K out of m tests to combine ($\boldsymbol{y}$) and the decision rule for that combination ($\boldsymbol{x}$). In terms of terminology, we refer to $\boldsymbol{x}$ as the "(classification) rule" and to $\boldsymbol{y}$ as the "test combination", while we use the term "policy" to refer to both the tests selected and how their results are to be interpreted.

3.3. Accounting for Noisy Estimation

Problem (1) uses in-sample estimates $\hat{a}_{\boldsymbol{s}}$ and $\hat{b}_{\boldsymbol{s}}$ and is not protected against estimation error. Our optimization perspective enables us to use robust optimization to address this concern.

First, we construct an *uncertainty set* around $\hat{\boldsymbol{a}}$, denoted as $\mathcal{U}_{\hat{\boldsymbol{a}}}$, which describes a set of plausible values for $\boldsymbol{a}$. Then, we replace the sensitivity term $\hat{\boldsymbol{a}}^\top \boldsymbol{x}$ in the objective of (1) by its the worst-case value over all $\boldsymbol{a} \in \mathcal{U}_{\hat{\boldsymbol{a}}}$, i.e., solve

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left[\min_{\boldsymbol{a}\in\mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x} \right] \text{ s.t. } \hat{\boldsymbol{b}}^\top \boldsymbol{x} \leq 1 - t_0. \tag{2}$$

For clarity of exposition, we detail our approach to sensitivity estimation (based on $\hat{\boldsymbol{a}}$) only, but also implement a conceptually similar approach for specificity (based on $\hat{\boldsymbol{b}}$).

We decompose the constraints defining $\mathcal{U}_{\hat{\boldsymbol{a}}}$ into three categories: (i) $\boldsymbol{a}$ should define valid probabilities on $\mathcal{S}_m$, (ii) the components of $\boldsymbol{a}$ should be 'close' to their in-sample estimates $\hat{\boldsymbol{a}}$, and (iii) the test sensitivities computed from the components of $\boldsymbol{a}$ should be consistent with confidence intervals on the performance of each test. We formalize these constraints below:

First, $\boldsymbol{a}$ is a vector of probabilities: $\boldsymbol{a} \geq 0$ and $\sum_{\boldsymbol{s}\in\mathcal{S}_m} a_{\boldsymbol{s}} = \boldsymbol{e}^\top \boldsymbol{a} = 1$.

Second, since $\hat{\boldsymbol{a}}$ are in-sample estimates of the underlying probabilities, any candidate probability vector $\boldsymbol{a}$ could be different, yet close to $\hat{\boldsymbol{a}}$. Accordingly, we bound the total absolute deviation $\sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}|$ with the constraint $\sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \leq \Gamma$, where Γ controls the level of robustness, i.e., protection against noise. Observe that total absolute deviation constraints are a special case of ϕ-divergence constraints (Ben-Tal et al. 2013)[2]. In our experiments, we treat the robustness parameter Γ as a hyper-parameter and calibrate its value using cross-validation (see §4.3).

[2] Among ϕ-divergences, Van Parys et al. (2021) prove that uncertainty sets based on the KullbackLeibler divergence are optimal in terms of out-of-sample disappointment. However, the resulting robust optimization problem here would be a mixed-integer conic optimization problem over the exponential cone. For tractability considerations, we have favored total absolute deviation, which leads to a simple mixed-integer linear program.

Third, for each test $k \in \{1,\dots,m\}$ and any $\boldsymbol{a}$, $\sum_{\boldsymbol{s}:s_k=1} a_{\boldsymbol{s}}$ is the associated sensitivity of test k. Assume that $[\hat{p}_k - \hat{\rho}_k, \hat{p}_k + \hat{\rho}_k]$ defines a $(1-\alpha)$ confidence interval on the estimated sensitivity of test k:

ASSUMPTION 1. *The true probability vector $\boldsymbol{a}_0$ satisfies each constraint $\left|\sum_{\boldsymbol{s}:s_k=1} a_{0,\boldsymbol{s}} - \hat{p}_k\right| \le \hat{\rho}_k$, for $k=1,\dots,m$, with probability at least $1-\alpha$.*

Accordingly, it is reasonable to expect $\boldsymbol{a}$ to satisfy similar constraints:

$$\left|\sum_{\boldsymbol{s}:s_k=1} a_{\boldsymbol{s}} - \hat{p}_k\right| \le \hat{\rho}_k.$$

In practice, the values of $\hat{p}_k$ and $\hat{\rho}_k$ could be provided by the manufacturer or estimated from data.

Observe that the second and third categories of constraints described above serve different purposes. The second group of constraints controls the deviation of $\boldsymbol{a}$ from $\hat{\boldsymbol{a}}$ in absolute terms, whereas in the third group of constraints, an under-estimate in one component of $\boldsymbol{a}$ could compensate for an over-estimate in another. Furthermore, the bound $\sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \le \Gamma$ weights all components of $\boldsymbol{a}$ equally, while the sensitivity constraints exploit a particular problem structure, i.e., the set $\{\boldsymbol{s} : s_k = 1\}$, which corresponds to test k being positive. By leveraging the fact that we can reconstruct the individual test sensitivities (via partial summations), we notably reduce the size of the uncertainty set (hence the conservatism of the approach).

Taken together, the uncertainty set $\mathcal{U}_{\hat{\boldsymbol{a}}}$ is defined as:

$$\mathcal{U}_{\hat{\boldsymbol{a}}} := \left\{ \boldsymbol{a} \ge 0 \,:\, \boldsymbol{e}^\top \boldsymbol{a} = 1;\ \sum_{\boldsymbol{s}\in\mathcal{S}_m} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \le \Gamma;\ \left|\sum_{\boldsymbol{s}:s_k=1} a_{\boldsymbol{s}} - \hat{p}_k\right| \le \hat{\rho}_k, \forall k \right\},$$

and Problem (2) returns the combination rule with the best worst-case sensitivity. Under Assumption 1, our pessimistic estimate $\min_{\boldsymbol{a}\in\mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x}$ provides a valid lower bound on the true performance of a combination rule, $\boldsymbol{a}_0^\top \boldsymbol{x}$:

PROPOSITION 1. *Assume the true probability vector $\boldsymbol{a}_0$ satisfies Assumption 1, and that $\Gamma = \gamma\sqrt{n/N^+}$ where N^+ denotes the number of condition-positive observations in the training data and $\gamma \ge 1$. Then, with probability at least $1 - m\alpha - \exp\left(-\frac{n}{2}(\gamma-1)^2\right)$ we have*

$$\boldsymbol{a}_0^\top \boldsymbol{x} \ge \min_{\boldsymbol{a}\in\mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x}.$$

A proof of Proposition 1 is included in Appendix A4. Put simply, this proposition states that the probability of out-of-sample disappointment, i.e., the probability that the actual sensitivity of rule $\boldsymbol{x}$, $\boldsymbol{a}_0^\top \boldsymbol{x}$, is less than the conservative estimate $\min_{\boldsymbol{a}\in\mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x}$ is bounded above by the probability ($\le m\alpha$) that the sensitivity of at least one test included in rule $\boldsymbol{x}$ falls outside of its $(1-\alpha)$ confidence interval, plus a term that exponentially decreases in $\Gamma^2 \propto \gamma^2$. For the latter term

to remain constant, we need γ to scale as $1/\sqrt{n}$, i.e., Γ to scale as $1/\sqrt{N^+}$. Such probabilistic guarantees are pervasive in robust optimization (see Bertsimas et al. 2021, and references therein). For $\Gamma = \gamma\sqrt{n/N^+}$, Bertsimas et al. (2021, Corollary 3) leads to an exponential term $\exp\left(-\frac{1}{2nN^+}\gamma^2\right)$, which is worse than the exponential term in Proposition 1, yet holds without any restriction on γ.

Finally, using strong duality, we can convert the inner minimization problem in Problem (2) into one of maximization, and reformulate (2) as a convex discrete optimization problem, solvable by branch-and-bound, as stated in Lemma 2 (proof included in Appendix A4).

LEMMA 2. *Problem* (2) *is equivalent to a mixed-integer linear optimization problem.*

In practice, we apply a similar approach to $\hat{\boldsymbol{b}}$. By considering confidence intervals on the specificity of each test[3], $[\hat{q}_k - \hat{\theta}_k, \hat{q}_k + \hat{\theta}_k]$, we construct an analogous uncertainty set

$$\mathcal{U}_{\hat{\boldsymbol{b}}} = \left\{ \boldsymbol{b} \geq 0 \,:\, \boldsymbol{e}^\top \boldsymbol{b} = 1;\ \sum_{\boldsymbol{s}\in\mathcal{S}_m} |b_s - \hat{b}_s| \leq \Gamma;\ \left| \sum_{\boldsymbol{s}:s_k=0} b_{\boldsymbol{s}} - \hat{q}_k \right| \leq \hat{\theta}_k, \forall k \right\},$$

and consider a robust knapsack problem with uncertainty on both $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left\{ \min_{\boldsymbol{a}\in\mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x} \right\} \text{ s.t. } \boldsymbol{b}^\top \boldsymbol{x} \leq 1 - t_0, \forall \boldsymbol{b} \in \mathcal{U}_{\hat{\boldsymbol{b}}}. \tag{3}$$

3.4. Accounting for partial observations

Next, we address another important source of uncertainty: missing data. Indeed, the ability to accommodate missing data is crucial for practical impact since real data routinely contains missing entries. In particular, in most studies that compare the performance of tests, all tests cannot be conducted on all individuals due to experimental failure or limited volume of each biological sample (see §4.1). Hence, a complete-case analysis would lead to different sample sizes for every subset of tests, making apples-to-apples comparisons of sensitivity/specificity values difficult.

In the absence of missing data, we know all the test results for all individuals, i.e., we can associate each individual with a single outcome sequence $\boldsymbol{s} \in \mathcal{S}_m$. We have access to a matrix $\boldsymbol{Z} \in \{0,1\}^{N^+ \times n}$, where $Z_{i,\boldsymbol{s}} = 1$ if individual i obtains sequence $\boldsymbol{s}$. From $\boldsymbol{Z}$, we can compute the in-sample sensitivity estimates $\hat{\boldsymbol{a}} = \frac{1}{N^+}\boldsymbol{Z}^\top \boldsymbol{e}$, or equivalently

$$\hat{a}_{\boldsymbol{s}} = \frac{1}{N^+} \sum_{i=1}^{N^+} Z_{i,\boldsymbol{s}}.$$

However, some individuals might have taken fewer than m tests. We cannot associate such individuals with a single outcome sequence $\boldsymbol{s}$. Instead, for each individual i, we consider the subset

[3] Typically, more data is available for estimating specificity because condition-negative pre-COVID-19 freeze-dried samples are available (whereas sensitivity estimation requires condition-positive samples), so the confidence intervals on specificity estimates are smaller.

$\mathcal{S}^{(i)}$ of all outcome sequences that are consistent with the observed test results. If individual i took all the tests, $\mathcal{S}^{(i)}$ is a singleton. Formally, let $\boldsymbol{s}^{(i)} \in \{0, 1, \texttt{NA}\}^m$ denote the observed sequence of results for individual i. $s_k^{(i)} = 1$ (resp. 0) indicates that test k is positive (resp. negative), while $s_k^{(i)} = \texttt{NA}$ indicates that the result of test k is unavailable. Accordingly, $\mathcal{S}^{(i)}$ is defined as $\mathcal{S}^{(i)} = \{\boldsymbol{s} \in \mathcal{S}_m : \forall k, s_k^{(i)} \neq \texttt{NA} \implies s_k^{(i)} = s_k\}$. Without imposing additional restrictions, the matrix $\boldsymbol{Z}$ belongs to the set

$$\mathcal{Z}^+ := \left\{ \boldsymbol{Z} \in [0,1]^{N^+ \times n} : \sum_{\boldsymbol{s} \in \mathcal{S}^{(i)}} Z_{i,\boldsymbol{s}} = 1, \forall i;\ Z_{i,\boldsymbol{s}} = 0, \forall \boldsymbol{s} \notin \mathcal{S}^{(i)} \right\}.$$

Without loss of generality, we relax the constraint $Z_{i,\boldsymbol{s}} \in \{0,1\}$ to $Z_{i,\boldsymbol{s}} \in [0,1]$ and interpret $Z_{i,\boldsymbol{s}}$ as the likelihood that individual i gets sequence $\boldsymbol{s}$.

Since $\boldsymbol{Z}$ is no longer unique, neither is $\hat{\boldsymbol{a}}$. A standard imputation technique would select one $\boldsymbol{Z} \in \mathcal{Z}^+$ (e.g., by likelihood maximization), compute $\hat{\boldsymbol{a}} = \frac{1}{N^+} \boldsymbol{Z}^\top \boldsymbol{e}$, and solve (2) for this value of $\hat{\boldsymbol{a}}$. Instead, we consider all the possible values for $\boldsymbol{Z}$ and integrate the uncertainty in $\boldsymbol{Z}$ into our robust formulation by replacing the uncertainty set in (2) by

$$\tilde{\mathcal{U}}_{\boldsymbol{a}} = \left\{ \boldsymbol{a} : \exists \boldsymbol{Z} \in \mathcal{Z}^+ \text{ s.t. } \boldsymbol{a} \in \mathcal{U}_{\hat{\boldsymbol{a}}} \text{ with } \hat{\boldsymbol{a}} = \tfrac{1}{N^+} \boldsymbol{Z}^\top \boldsymbol{e} \right\}.$$

The resulting problem can be formulated as a linear mixed-integer optimization problem similar to Problem (2) (see Lemma A1 in Appendix A4). Above, we focused our exposition on the condition-positive individuals and the estimation of $\boldsymbol{a_0}$, but the same logic applies for $\boldsymbol{b_0}$. We will demonstrate the benefit of this robust imputation technique over maximum-likelihood imputation via numerical experiments.

4. Data and Implementation

In this section, we introduce our data and describe the application of our methods from §3 in the context of antibody and antigen testing.

4.1. Data

We apply our methodology to two datasets, provided to us by researchers on the National COVID Scientific Advisory Panel (NCSAP) in the U.K. and the Charite Berlin hospital in Germany (which have informed the U.K. governments and the WHOs recommendations on COVID-19 testing, respectively). The datasets record the results of a head-to-head comparison of the diagnostic accuracy of POC tests for COVID-19 antibody and antigen detection, respectively. Note that these studies use stored samples of COVID-positive and -negative individuals (as most studies in the field do). As a result, the lack of accuracy of any given test can most likely be attributed to a limitation in the testing technology rather than the administration of the test.

Table 1 Individual performance of 9 POC LFIA devices for antibody detection. The 95% confidence intervals (CI) are obtained using the Wilson score method with continuity correction.

Test	(N_+, N_-)	Sensitivity (95% CI)	Specificity (95% CI)
1	(33,60)	0.55 (0.47, 0.61)	1.00 (0.93, 1.00)
2	(38,91)	0.61 (0.53, 0.66)	0.99 (0.94, 1.00)
3	(33,60)	0.64 (0.55, 0.70)	0.97 (0.89, 0.98)
4	(38,60)	0.66 (0.58, 0.71)	0.98 (0.91, 1.00)
5	(31,60)	0.61 (0.52, 0.68)	0.97 (0.89, 0.98)
6	(31,60)	0.65 (0.55, 0.71)	0.98 (0.91, 1.00)
7	(33,60)	0.70 (0.6, 0.75)	0.95 (0.88, 0.97)
8	(32,60)	0.56 (0.48, 0.63)	1.00 (0.93, 1.00)
9	(40,142)	0.55 (0.48, 0.61)	0.97 (0.94, 0.98)

Antibody tests: The Wellcome Trust funded a performance review of 9 POC lateral flow immunoassay (LFIA) devices by the NCSAP (Adams et al. 2020). In particular, their point estimates for the sensitivity (resp. specificity) of LFIA devices range from 0.55 to 0.70 (resp. $0.95-1.00$) versus the benchmark lab test, which led them to conclude that "currently available commercial LFIA devices do not perform sufficiently well for individual patient applications". We report the performance of the individual LFIAs in Table 1. The 95% confidence intervals (CI) are obtained using the Wilson score method with continuity correction (Newcombe 1998)[4]. Observe that the authors could not apply all tests to all samples due to the limited number of LFIA devices made available to them.

Antigen tests: Corman et al. (2021) compared the performance of 7 antigen tests. They all target the same viral protein, namely the nucleocapsid protein. Table 2 reports 95% confidence intervals for the specificity and sensitivity of each test, using the same method. As noted by the authors, their experimental setting deviates from manufacturers' instructions (they conduct a predilution step for condition-positive samples). So, the measured sensitivities might be lower than in practice. Given the poor sensitivity of Test II (0.09) on 45 positive samples with high viral load, Corman et al. (2021) decided not to study its sensitivity further and not to apply it on the remaining samples. To avoid any selection bias in our analysis, we will exclude test II from our study. The summary statistics on our data, reported in Table 3, motivate two useful observations. First, the size of both datasets is limited so the estimates for the diagnostic accuracy of the individual tests are subject to finite-sample estimation error. This motivates our robust formulation which accounts for such uncertainty (applied in §5.3). Second, both datasets suffer from missing data. Indeed, in both cases the authors were unable to apply all the tests to all samples due to limited testing kit availability, limited volume per sample, or experimental failure. A naïve complete case analysis

[4] The Wilson score method leads to CIs that are not centered around the empirical mean and uses a biased estimator instead, which has lower variance and overall lower mean squared error. This is a more appropriate method when the sample mean is close to the unit interval end points.

Table 2 Individual performance of 7 antigen tests. The 95% confidence intervals (CI) are obtained using the Wilson score method with continuity correction.

Test	(N^+, N^-)	Sensitivity (95% CI)	Specificity (95% CI)
I	(105,135)	0.71 (0.68, 0.73)	0.99 (0.96, 1.00)
II	(45,135)	0.09 (0.08, 0.16)	1.00 (0.97, 1.00)
III	(105,135)	0.85 (0.81, 0.86)	0.89 (0.86, 0.90)
IV	(105,135)	0.28 (0.26, 0.31)	1.00 (0.97, 1.00)
V	(105,135)	0.76 (0.73, 0.78)	0.95 (0.92, 0.96)
VI	(105,135)	0.34 (0.33, 0.37)	0.99 (0.96, 1.00)
VII	(115,135)	0.70 (0.67, 0.72)	0.99 (0.95, 0.99)

would reduce the already limited sample size (and increase the estimation error). This motivates our robust formulation which additionally accounts for missing data (evaluated in §5.4).

Table 3 Characteristics of the diagnostic testing datasets used in our numerical experiments.

Dataset	N^+	N^-	m	Complete cases	Missing entries
Antibody tests (Adams et al. 2020)	40	142	9	89	41%
Antigen tests (Corman et al. 2021)	138	135	6	207	11%

Table notes: N^+ (resp. N^-) denotes the number of condition-positive (resp. negative) samples and m is the number of tests.

Finally, we note that our data contains no precise information on the cost of each of these different tests. As mentioned earlier, data on test costs is scarce (Xiao and Rathi 2021) and costs depend on many factors and vary widely. Since each dataset compares tests that are similar to manufacture and administer (e.g., LFIA devices), in the absence of more precise information, we will work under the assumption that all tests have comparable (yet unknown) cost. Accordingly, we can set $c_k = 1$ in the budget constraint defining $\mathcal{X}$ and, in our analysis, identify the overall cost of a combination testing policy with the number of tests involved, K. Nonetheless, as previously mentioned, combining a few POC tests would remain cheaper than conducting a lab-based test.

4.2. Motivation for Combining COVID-19 Tests

Intuitively, combining tests can be beneficial for systematic or stochastic reasons. First, each test might use a different underlying technology, hence providing its unique view on the patient's condition. Second, a test result can be viewed as a stochastic signal and combination testing as a denoising procedure. We illustrate these two mechanims on two stylized examples in Appendix A1. In this section, we provide examples of different underlying technologies or sources of randomness in the case of COVID-19 detection.

First, consider the systematic mechanism. Technological differences across tests may allow one test to detect a target protein for a given patient, even if another test does not. Antigen tests detect

the presence of virus proteins. For COVID-19, tests usually target the nucleocapsid protein or the spike protein, by mixing the patient's sample with synthetic antibodies that have high affinity towards one sub-unit of the protein of interest. Because these proteins might be present in infected patients at different concentrations, for instance, and because the synthetic antibodies used might be constructed differently, the tests might give discordant results. For antibody tests, two types of antibodies are typically used (IgM and IgG) and manufacturers implement different detection strategies. Combining multiple tests can therefore result in improved accuracy by leveraging the different testing technologies. Similarly, Blacksell et al. (2011), Veyrenche et al. (2021) consider detecting current infections using one antigen and one IgM antibody test.

Second, consider the stochastic mechanism. Each test result can be viewed as the realization of a random variable and conducting multiple tests can reduce the noise associated with each application of a test. At the patient level, this stochastic view is supported by the presence of inherent noise stemming from how the test is administered and analyzed. For example, even applying multiple tests that rely on the same technology on the same individual might not consistently give the same result. Combining the results from multiple tests can have higher accuracy because it reduces noise, in the same way that diversification reduces the risk of a portfolio. In addition, some individual characteristics (e.g., viral load, time since infection, presence of symptoms) can impact test results and accuracy. At a patient level, they are constant, but they contribute to randomness (and correlation) in test results at the population level because they are unobserved and heterogeneous across patients.

Our methodology aims to leverage both of the mechanisms described above in a disciplined way to achieve higher classification accuracy. In an attempt to disentangle and quantify the extent to which systematic or stochastic mechanisms are prevalent, we compute and report (in Appendix A6.1) two metrics for each pair of tests, on condition-negative and condition-positive samples separately. To measure the extent of agreement among tests, we use the probability of pairwise agreement. To measure stochastic association between tests, we report their pairwise correlations. These descriptive statistics are included in Appendix A6.1—Tables A1 and A3 describe the antibody test data while Tables A2 and A4 describe the antigen test data.

We observe that pairwise correlations are close to 0 on condition-negative individuals, for both datasets. This suggests that discrepancy in test results is mostly due to independent noise. On the contrary, for condition-positive samples, agreement probabilities are low, demonstrating that tests give discordant results more often on condition-positive samples than on condition-negative samples. Furthermore, correlations are higher as well, suggesting that such discordance in these test results cannot be solely explained by independent random noise and that there might be underlying factors (e.g., time since infection for antibody tests or viral load for antigen tests) or systematic technological differences that correlate test results on condition-positive samples.

4.3. Application of Methods to COVID-19 Datasets

In our implementation of the robust model of §3.3, we will consider uncertainty sets for $\boldsymbol{a}$ of the form:

$$\mathcal{U}_{\hat{\boldsymbol{a}}}(\Gamma) = \left\{ \boldsymbol{a} \geq 0 \,:\, \boldsymbol{e}^\top \boldsymbol{a} = 1;\ \sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \leq \Gamma;\ \left| \sum_{\boldsymbol{s}: s_k = 1} a_{\boldsymbol{s}} - \hat{p}_k \right| \leq \hat{\rho}_k, \forall k \right\},$$

where $\hat{p}_k$ and $\hat{\rho}_k$ are set equal to the middle and half-width of the CIs for the sensitivity of test k (Tables 1 and 2) and $\Gamma \geq 0$ is a hyper-parameter which we calibrate using 5-fold cross-validation. A detailed description of the cross-validation procedure is provided in Appendix A5. We similarly construct uncertainty sets $\mathcal{U}_{\hat{\boldsymbol{b}}}(\Gamma)$ for $\boldsymbol{b}$.

We make two notes about the presentation of our results. First, diagnostic accuracy is generally reported in terms of sensitivity and specificity. Improvements in one of these metrics (e.g., sensitivity) only make sense if the other (e.g., specificity) is capped at a given threshold. For instance, we report achievable improvement in sensitivity given varying specificity targets in Section 5.5. However, setting a relevant value for this threshold will depend on the use case. Accordingly, we use the area under the ROC curve (AUC) as an accuracy metric in the cross-validation procedure (Algorithm 2) and present our main results in terms of AUC, for it captures in a single number the sensitivity/specificity trade-off in a threshold-free manner[5]. AUC corresponds to the probability of correctly ranking a condition-positive individual higher than a condition-negative one, which is especially relevant when screening provides access to a scarce and expensive resource. Still, we would like to emphasize that our method is agnostic to the metric being used. Other measures of accuracy can be used, both for cross-validation and reporting of the results, if deemed more relevant to the application at hand.

Second, when selecting K out of m tests to combine, we solve a knapsack problem –(1) or (3)– for varying specificity thresholds t_0, hence generating a Pareto-dominating ROC curve. Each point on that curve corresponds to a policy, namely K tests and a rule to combine their results. Accordingly, the AUC of the Pareto-dominating ROC curve does not correspond to the AUC of a unique K-test combination (the $\boldsymbol{y}$ variable) but rather captures the overall predictive power of the available tests when combined, representing the state of technology. For K fixed, among all the subsets of K tests that lie on this Pareto-dominating ROC curve, we conservatively focus our attention on the best (i.e., AUC maximizing) one. This allows a fairer comparison with individual tests and emphasizes an application of our method that is simpler to implement in practice. Our results for the complete Pareto-dominating ROC curve (for the optimal combinations of K tests at any specificity threshold) are discussed in Appendix A6.

[5] For a single test, the AUC is equal to the arithmetic mean of the sensitivity and specificity.

5. Results

We demonstrate the benefits of implementing our combination testing policies using data on COVID-19 antibody and antigen tests, described in §4.1. We first illustrate the dominating ROC curve for a given subset of tests, in §5.1. We then discuss the improvement in accuracy achievable by optimally selecting the tests to include in a combination policy in §5.2, and the out-of-sample performance of our robust formulations that address estimation error and missing data in §5.3 and §5.4, respectively. We then compare the performance of a combination of tests selected using our methodology with the performance of heuristics previously proposed in the literature, in §5.5. Finally, we provide an illustration of how a policymaker would apply our methodology to obtain a specific combination testing policy, in §5.6. In addition, Appendix A7 contains results on synthetic data, exploring the impact of the sample size and the proportion of missing entries on our results.

5.1. Combining a Given Set of Tests

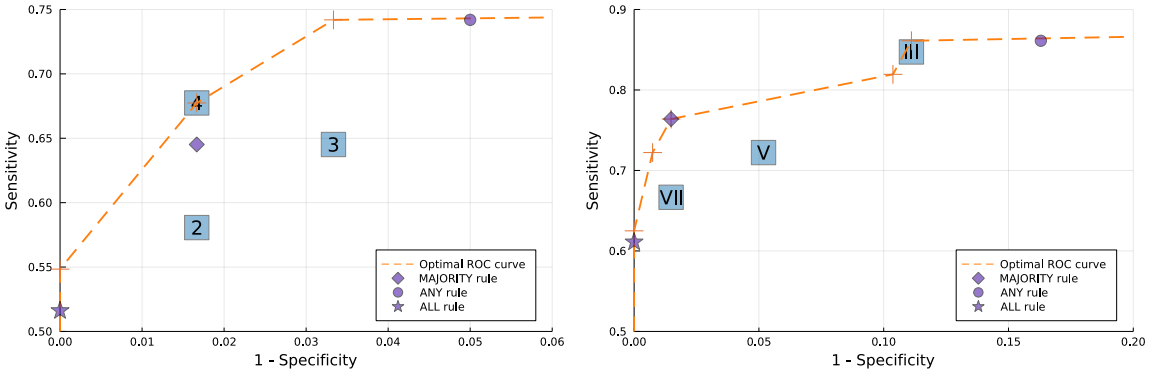
For illustration, we first restrict our attention to three arbitrarily chosen tests from each dataset. We obtain all Pareto-optimal combination rules for these tests, plotted as the break-points in the dashed (orange) ROC frontier in Figure 1, by solving the knapsack problem (1) for different values of t_0. As discussed in §3.1, the frontier can be computed by sorting the 8 possible outcome sequences by their $\hat{a}_{\boldsymbol{s}}/\hat{b}_{\boldsymbol{s}}$ values. For now, we consider only complete cases. We observe from Figure 1 that the dashed ROC frontier dominates, at least weakly, the performance of all the three individual tests, thus illustrating the benefit from combining tests. We also observe that the standard heuristics discussed in §2 are not guaranteed to lie on the frontier and that our methodology identifies points on the ROC frontier that the heuristics cannot (§5.5 and §A6.2 contain a more extensive comparison with these heuristics).

While Figure 1 demonstrates the performance of our methodology for a given set of tests, a key part of the practical challenge is to select which tests to combine, especially given the proliferation of rapid tests. For example, our antibody (resp. antigen) dataset comprises 9 (resp. 6) different tests, leading to 36 (resp. 15) possible 2-test combinations and 84 (resp. 20) 3-test combinations. In the following three sections, we efficiently search for the best combination of three antibody (or antigen) tests by solving our extended formulation with additional binary variables $\boldsymbol{y}$, and illustrate how robustness mitigates the issues of overfitting and missing data.

5.2. Selecting a Combination of Tests

We next illustrate the improvement in diagnostic accuracy from optimally selecting the tests to be included in the combination testing policy. We consider all available tests and solve (3) with $\Gamma = 0$ (i.e., obtaining only in-sample estimates of $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$) on all the complete cases and construct the ROC curve for the individual tests, as well as for the Pareto-optimal combinations of 2 and 3 tests.

Figure 1 **The trade-off between specificity and sensitivity (ROC curve) for 3 individual tests and all their potential combinations.**



(a) Antibody test data (b) Antigen test data

Figure 2 displays both the performance of the individual tests (labelled by their index number) as well as the ROC frontier of combinations of two (the orange dashed line with "×" breakpoints) and three (the green dashed line with "+" breakpoints) tests. We observe that the ROC frontiers for either combining two or three tests lies above the convex hull of the index tests.

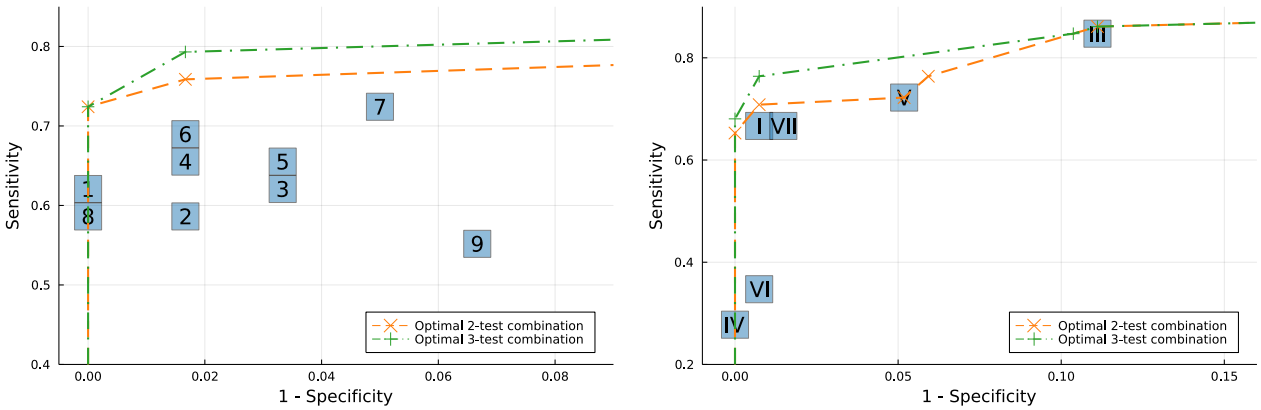
On the antibody test data, while the best performing test (test 7) has an in-sample AUC of 0.837, the set of Pareto-optimal 3-test combinations achieves 0.894 AUC (+7%). It also identifies two 3-test combinations – (6,7,8) and (1,6,8) – with AUCs of 0.862 and 0.888 respectively. Similarly, on the antigen test data, we observe a 5% increase in AUC between the best performing test (test III, AUC = 0.868) and the Pareto-optimal set of 3-test combinations (AUC = 0.921). Among the 3-test combinations on the Pareto-dominating ROC curve, (I,III,V) is the most accurate one (AUC = 0.878).

5.3. Robustness to Parameter Estimation

Due to finite-sample estimation error, our optimization model might find combination policies that improve accuracy on the training data but do not generalize to the entire population.

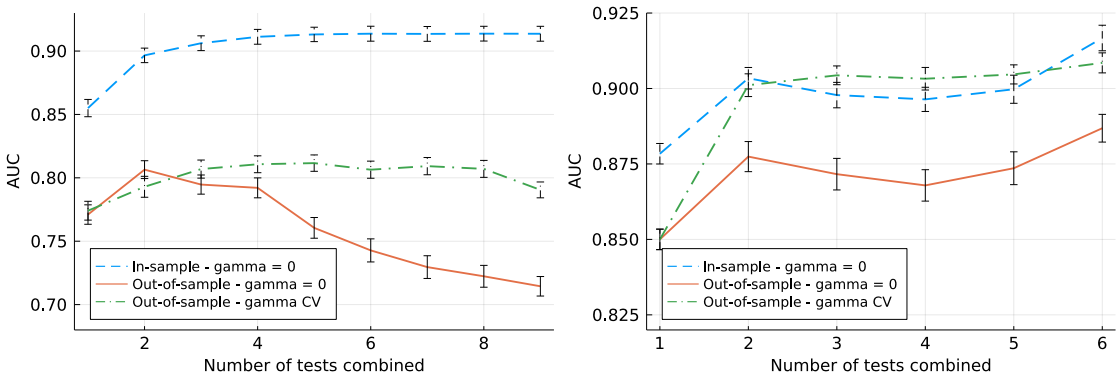
To illustrate the value of robust optimization in mitigating this concern, we repeatedly split the data into training and validation datasets using a 70%/30% split (robustness of the results with respect to the splitting proportion is presented in Appendix A6.2 (Figure A2), using stratified sampling to keep prevalence constant between the two datasets. For each training dataset, we apply the following methodology, for each K: we estimate the sensitivity and specificity of each test and their 95% CIs, construct uncertainty sets of the form $\tilde{\mathcal{U}}_{\hat{\boldsymbol{a}}}(\Gamma)$ (resp. $\tilde{\mathcal{U}}_{\hat{\boldsymbol{b}}}(\Gamma)$) for $\boldsymbol{a}$ (resp. $\boldsymbol{b}$), and compute the Pareto-dominating ROC curve. From this curve, we identify the most accurate K-test combination on the training data, and evaluate its performance on the validation data. We repeat this procedure over 50 random splits.

Figure 2 **The trade-off between specificity and sensitivity (ROC curve) for all the individual tests and the optimal combinations of 2 (orange line) and 3 tests (green line).**



(a) Antibody test data (b) Antigen test data

Figure 3 **Diagnostic accuracy (AUC) of the best K-test combination as a function of the number of tests combined (K).**



(a) Antibody test data (b) Antigen test data

Figure notes: Error bars (in all figures of the manuscript) correspond to ±1 standard error.

Figure 3 compares three performance curves. First, it reports in- and out-of-sample performance of the non-robust solution ($\Gamma = 0$) with missing data imputed using maximum likelihood (ML) imputation (Gelman and Hill 2006, Chapter 25). On both datasets, we observe that the out-of-sample AUC (solid orange line) is significantly lower than its in-sample value (dashed blue line). This discrepancy is starker on the antibody test data: while in-sample AUC unambiguously increases as more tests are used, out-of-sample performance deteriorates when using 3 tests or more. Second, the dash-dotted green line in Figure 3 displays the out-of-sample AUC of a robust policy where the value of Γ is calibrated using a 5-fold cross-validation procedure (see Appendix A5 for pseudo-code) on the training dataset and where missing data is imputed adversarially. The figure illustrates that adding robustness can improve out-of-sample performance. In both cases, the

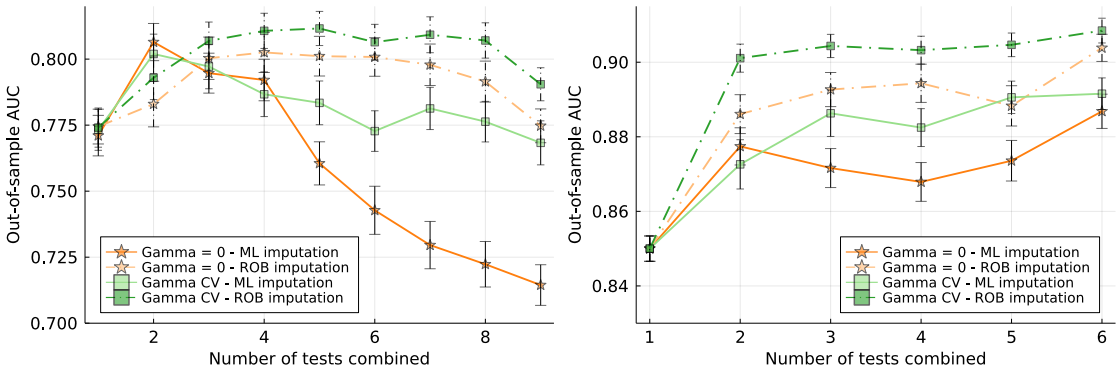
robust models exhibit comparable or stronger out-of-sample performance than the naïve approach. For instance, our method identifies 3-test combinations with an average AUC of 0.802 on the antibody test data (+4% compared with 0.774 for a single test) and of 0.902 on the antigen test data (+6%). For moderate values of K ($K \leq 4$), which are probably the most relevant values for policy implications, the benefit from robustness is mild on the antibody test data but stark on the antigen test data. In the appendix, we report similar yet smaller gains for the entire Pareto-dominating curve instead of the best K-test combination only, in Figure A3.

The behavior of the method displayed on Figure 3 is notably different on the antibody and antigen test data. On the antibody test data, the non-robust optimization formulation provides good solutions for up to 3 tests but its out-of-sample performance deteriorates for larger values of K. The robust formulation does not significantly improve on the non-robust formulation for small values of K but stabilizes the out-of-sample AUC for $K \geq 4$. On the contrary, on the antigen test data, the accuracy of the non-robust formulation generally increases with K, and the robust formulation provides substantial improvement for all $K \geq 2$. One explanation could be the difference in sample sizes. The total number of observations (and of condition-positive samples) in the antigen data is more than twice (resp. three times) as high as for the antibody data, and there is less missing data. An alternative explanation could be that among the tests in our data, the antibody tests might rely on more similar technologies than the antigen tests so combining tests helps mostly because it reduces noise, not because we gain new information. To support this intuition, we decompose the results for antigen tests on high- and low-viral load samples separately in Appendix A6.3. In summary, our (non)-robust combination policies display a similar behavior on the high-viral load samples to what we observe for the antibody test data. This observation corroborates the intuition that when various technologies are available and tests are systematically discordant (as in the low-viral load antigen tests data), our method is able to find particular classification rules that leverage these different signals, while when technologies are more mature and similar, combination of tests helps mainly as a denoising procedure.

5.4. Robustness to Missing Data

The robust optimization approach we consider in the previous section differs from the baseline approach in two ways. First, it imputes the missing values adversarially instead of using maximum likelihood. Second, it accounts for noisy estimates by considering uncertainty sets around the empirical values $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ (with Γ cross-validated instead of $\Gamma = 0$). Both ingredients provide some form of robustness and could explain the observed improvement. To isolate the relative benefit of each of them, Figure 4 compares the out-of-sample performance of two models non-robust to noisy estimates ($\Gamma = 0$): one where missing entries are imputed with their most likely values as in Figure 3 (orange solid line) and another where they are imputed adversarially (orange dash-dotted line).

Figure 4 Out-of-sample accuracy (AUC) as a function of the imputation method for missing data.



(a) Antibody test data (b) Antigen test data

Figure notes: The orange solid line and the dashed green line in this figure are the same as in Figure 3.

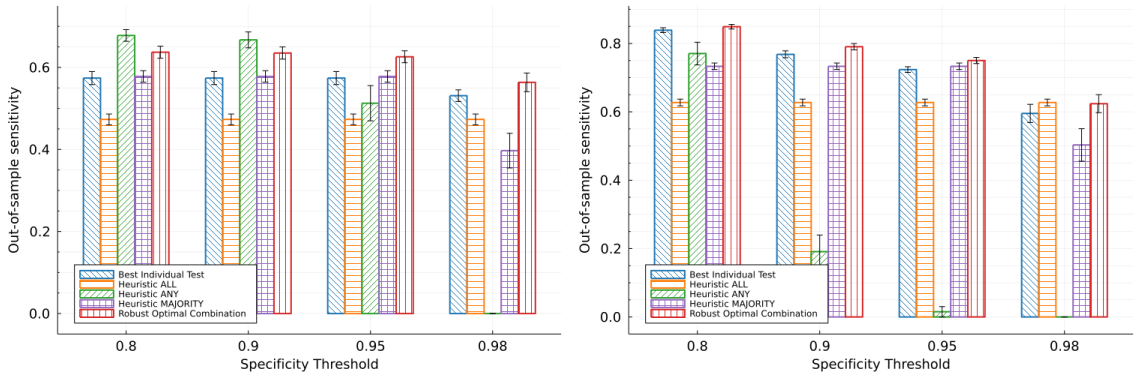
We observe that robust imputation provides a substantial improvement over ML imputation, especially as K increases. For comparison purposes, we also report the performance of both imputation methods with additional robustness to noisy estimates (Γ CV) in green. In summary, irrespective of the value of Γ (0 or CV) adversarial imputation improves out-of-sample predictive power over ML imputation (dash-dotted vs. solid). Irrespective of the imputation method, adding robustness (Γ CV) helps (green vs. orange). The robust formulation with adversarial imputation provides the best performance.

5.5. Comparison with Existing Heuristics

As discussed in §1 and §2, the prior literature has proposed various heuristics for interpreting the outcomes of multiple tests performed to detect the same condition. In particular, the `any` (if any test is positive), `all` (if all tests are positive), and the `majority` (if the majority of tests is positive) heuristics have been proposed.

We first compare the performance of the set of Pareto-optimal 3-test combinations with that of a single test and of the three aforementioned heuristics applied on the three individually best-performing tests for each dataset. On the training data, for each specificity target, we select the best test (or 3-test combination) in terms of sensitivity and compute its out-of-sample sensitivity on the validation set. We focus on 3-test combinations since Figure 3 demonstrates no additional benefits from adding more tests to a combination. For four specificity targets, Figure 5 reports the average out-of-sample sensitivity achievable by a single test, the heuristics, and our method. We observe that using our methodology to identify the best combination of tests and the appropriate classification rule to use at each threshold results in a significant improvement in sensitivity, for almost all specificity thresholds. In particular, if a policymaker requires specificity of at least 0.95, our methodology results in an 8% and 2% relative improvement in sensitivity for antibody and

Figure 5 Out-of-sample sensitivity achieved by combining 3 tests for varying specificity thresholds.



(a) Antibody test data - 3 test combinations (b) Antigen test data - 3 test combinations

Figure notes: The performance of the heuristics is obtained by identifying the 3 best performing individual tests for antibody (antigen) detection and applying the three heuristics.

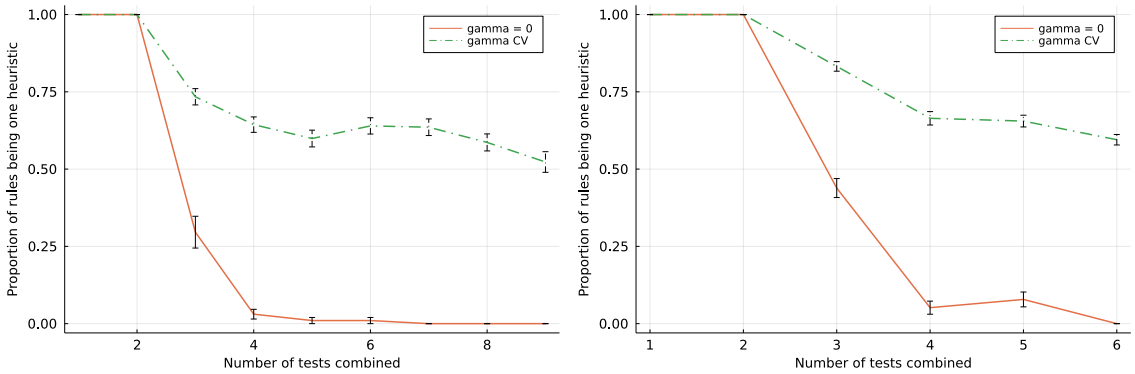
antigen detection, respectively, compared with the best performing heuristic (`majority` in these cases). Over the 8 cases (4 specificity targets $\times$ 2 datasets), our methodology significantly (p-value < 0.1) outperforms the available heuristics in 4 cases. Our method is only dominated (with a p-value < 0.1) by the `any` heuristic, on the antibody test data, and for a specificity threshold of 0.8. This behavior is most likely due to our cross-validation procedure for Γ (see pseudo-code in Appendix A5, Algorithm 2), which optimizes for the aggregate AUC measure. Optimizing for partial AUC (or sensitivity conditional on at least 0.80 specificity) would likely lead to improved performance for our approach. Similarly, our optimal 3-test combination systematically improves over using a single test, with the comparison being statistically significant in 4 out of the 8 cases.

As previously discussed, our optimization method achieves two concurrent objectives: to identify a subset of K tests (as encoded by the variable $\boldsymbol{y}$) and then to find the best classification rule to aggregate their results (as encoded by the variable $\boldsymbol{x}$). In contrast, heuristic rules solely address the issue of classification and not of test selection. We now try to disentangle the respective benefit of our method on the test selection and classification tasks separately.

In Figure 6, we consider the entire Pareto-optimal set of K-test combinations and count how many of them correspond to one of the aforementioned heuristics on the subset of tests they involve[6]. Obviously, for $K \leq 2$ tests, any classification rule can be seen as an `any` or `all` rule. As the number of tests involved, K, increases, the fraction of Pareto-optimal combinations that correspond to one of the heuristics decreases. However, we observe that our robust optimization formulation recovers heuristic rules substantially more frequently than the non-robust version does, suggesting that the heuristics are somewhat robust classification rules and that the optimal selection of tests is

[6] Figure A5 in appendix decomposes these numbers for each heuristic rule separately.

Figure 6 **Fraction of the Pareto-optimal K-test combinations corresponding to one of the heuristic rule (`all`/`any`/`majority`).**



(a) Antibody test data (b) Antigen test data

a central added value of our method. Still, for 3 tests, around 25% (resp. 20%) of the Pareto-optimal combinations are more complex than the heuristics on the antibody (resp. antigen) data.

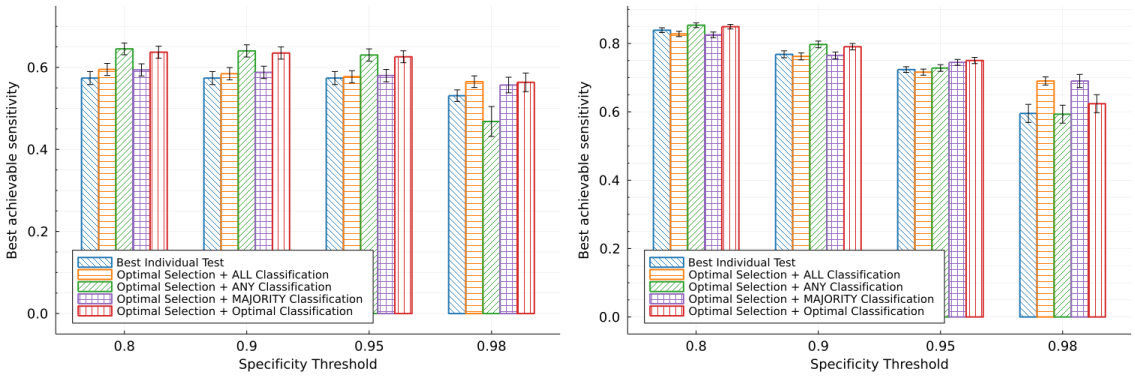
To further illustrate the benefit of our optimization formulation in optimally selecting which test to combine, we solve our robust optimization problem (3) with additional constraints that force the classification rule $\boldsymbol{x}$ to be one of the aforementioned heuristics (see detailed formulation in Appendix A2). Doing so provides the existing heuristics with a disciplined approach to identify the most powerful combination of tests while accounting for in-sample estimation error and missing data. Figure 7 replicates Figure 5 except that tests for the heuristic rules are selected optimally instead of heuristically. We observe that test selection has a first-order impact on the accuracy of a combination policy: optimally selecting the tests substantially increases the performance of all heuristic rules, up to the level of the optimal combination. For example, for a specificity of at least 0.98, the `majority` rule achieves a sensitivity of 0.56 (resp. 0.69) on the antibody (resp. antigen) data when tests are optimally selected, compared with 0.40 (resp. 0.50) when the best index tests are selected and compared with 0.56 (resp. 0.62) for the optimal combination strategy.

5.6. An illustration of a Specific Policy Recommendation

Up to this point, our analysis has been focused on evaluating our methodology for uncovering robust combination testing policies, hence the emphasis on cross-validation and out-of-sample accuracy. We now illustrate how a decision maker could apply our methodology in practice.

In the simplest use case, in which the decision maker has an abundance of data, they would apply our methodology on a training dataset and evaluate out-of-sample performance on a validation dataset. However, in realistic use cases of our methodology, the decision maker would have access to a similar dataset to ours, which compares the performance of various POC tests on a set of samples. Those datasets tend to be smaller in size, so omitting data for training purposes can significantly

Figure 7 **Out-of-sample sensitivity achieved by optimally combining 3 tests for varying specificity thresholds. Tests are selected using our robust knapsack formulation. Test results are classified using the `any`, `all`, `majority`, and optimal rules.**



(a) Antibody test data - 3 test combinations

(b) Antigen test data - 3 test combinations

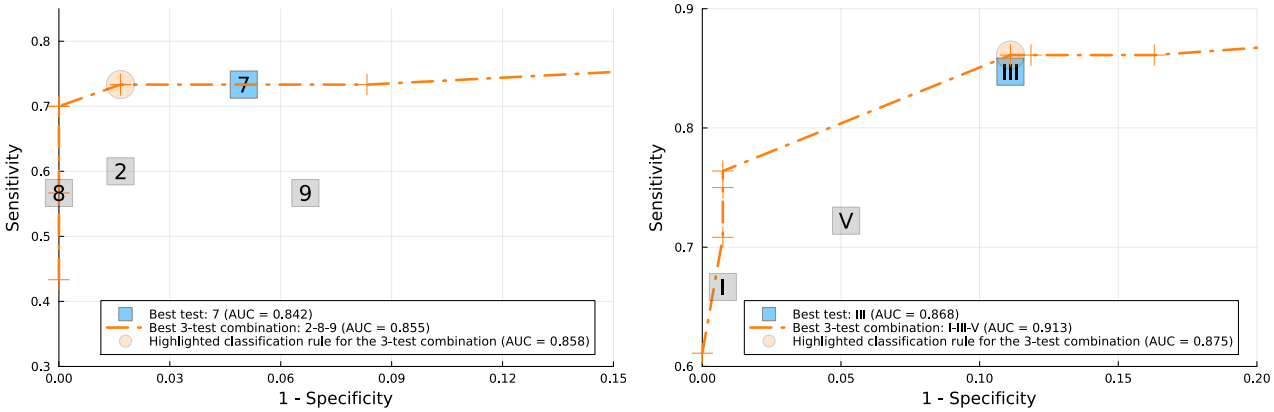
affect the solutions obtained and result in a too small validation dataset for a representative out-of-sample performance estimate. Therefore, for a realistic illustration of our methodology, we propose the following approach, based on the notion of *stability selection*, a method widely used for variable selection in statistics (Meinshausen and Bühlmann 2010).

As a first step, we leverage our dataset to identify the best single test and the best 3-test combination. We randomly split the data in to training (70%) and testing (30%) fifty times (the same approach as we use for cross validation). For each training dataset, we use our methodology to identify the best performing single test and the best performing three test combination—with an associated decision rule—as measured by their respective in-sample AUC. Using stability selection, we aggregate the results by counting how frequently each single test and each 3-test combination has the highest in-sample AUC across all the 50 training sets. As a second step, we evaluate the best single test and the best 3-test combination—identified using the stability selection approach described above—by calculating their respective AUC on the whole dataset.

Figure 8 shows the outcome of this process. We make four observations based on these results. First, for both the antibody and the antigen test datasets, a combination of three tests achieves a higher AUC than the best single test. For antibody testing (Figure 8(a)), a three test combination of Tests 2, 8, and 9 achieves an AUC of 0.855 while the best single test (Test 7) achieves an AUC of 0.842, a relative improvement of 1.5%. For antigen testing, (Figure 8(b)), a three test combination of Tests I, III, and V outperforms the best single test (Test III) by a relative 5% in terms of AUC.

Second, an important benefit of combination testing is that the same set of (e.g., three, for this illustration) tests gives the decision maker increased flexibility to emphasize either sensitivity or specificity, by choosing a specific point on the ROC curve to implement. For example, using

Figure 8 **ROC curve for best individual test and the best 3-test combinations (orange line), identified via stability selection.**



(a) Antibody test data

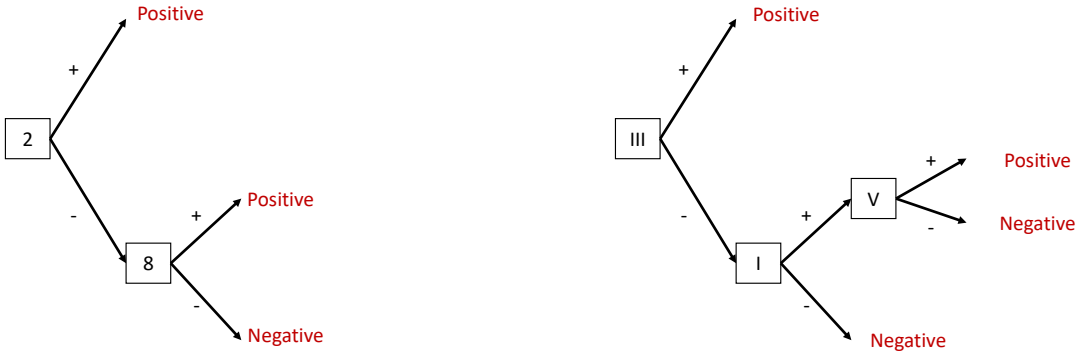
(b) Antigen test data

the same three antibody tests, the decision maker can achieve either a sensitivity of 0.73 (with a corresponding specificity of 0.92) or a sensitivity of 0.70 (with a corresponding specificity of 1.00). Similarly, for the selected antigen tests, the decision maker can choose between a sensitivity of 0.86 (with a corresponding specificity of 0.84) or a sensitivity of 0.61 (with a corresponding specificity of 1.00).

Third, this analysis provides an answer to the empirical question of *how much* sensitivity or specificity can be improved by combining tests in a specific application context. If the decision maker is satisfied with the sensitivity of the best performing antibody test (Test 7), specificity can be improved by 0.03 by combining tests (see the highlighted point in Figure 8(a)). On the other hand, if the decision maker is satisfied with the specificity of the best antigen test (Test III), sensitivity can be only modestly improved (by 0.01) by combining three tests (see the highlighted point in Figure 8(b)).

Fourth, each point on the ROC curve for a three test combination corresponds to a specific decision rule. If the decision maker wishes to improve antibody specificity by 0.03 (by choosing the highlighted point in Figure 8(a)) the associated decision rule is to apply the `any` rule using tests 2 and 8. In other words, even if we allow for up to three tests in the combination, this particular point on the ROC curve only requires the use of two tests. If the decision maker wishes to improve the sensitivity of antigen testing by 0.01 (by choosing the highlighted point in Figure 8(b)) they would classify a sample as positive if Test III returns a positive result or if both tests I and V return a positive result. The decision rules associated with these two points are illustrated in Figure 9. Note that such compact representations of a classification rule as a test sequence can be obtained systematically using optimization, as described in Appendix A3.

Figure 9 **Decision rules associated with the highlighted points in Figure 8.**



(a) Antibody test data

(b) Antigen test data

6. Conclusion and Implications

Despite advances in COVID-19 treatment and vaccination, the need for cheap, fast, and accurate testing remains. We contribute to the ongoing debate on the best use of POC devices for testing against COVID-19. We present a novel methodology, based on robust optimization, to best select and combine tests, hence providing a structured way for policymakers to quantify the benefit from combining multiple index tests and potentially break the trade-off between accuracy and cost in diagnostic testing. Our approach accounts for incomplete or noisy data, two significant practical considerations when working with limited data on emerging pathogens. We demonstrate the relevance and effectiveness of our algorithm on datasets, comprised of COVID-19 antibody and antigen test results, from two medical institutions.

Our results indicate that, by combining multiple POC tests, widespread COVID-19 testing can be simultaneously accurate, timely and cost-effective – and thus could help curb transmission while reducing unnecessary isolation. In a single week of June 2021, nearly 4% of British pupils missed school due to required isolation (U.K. Department of Education 2021). In the U.S., millions of passengers travel by air daily (Transportation Security Agency 2021). Initial guidelines imposed a negative RT-PCR result for all passengers at most 72 hours prior to departure, although they could be exposed and become infectious in the interim. Quick and accurate POC testing in these and other settings could reduce the costs of both false negatives and false positives.

With combination testing, a policymaker has access to a fine-grained ROC curve and can decide to emphasize either sensitivity (e.g., for vulnerable populations for which false negatives are costly) or specificity (e.g., in population-level seroprevalence studies) by selecting the appropriate classification rule. Designing combination rules that optimize for both downstream operational performance and predictive power is another interesting future direction. For example, our knapsack framework

can incorporate additional operational constraints such as limited availability of healthcare professionals to oversee or conduct testing. Overall, our method provides a practical and time-efficient method for policymakers to (re-)assess the benefit from combination testing, as new tests or new variants emerge.

Future research could apply our methodology to the repeated use of the same test as well. In that setting, instead of outcome sequences, one would describe test results by the number of positive or negative outcomes only. However, data on the repeated use of a single test is, to our knowledge, not commonly published.

In general, treatment policies and decisions based on statistics are ubiquitous. However, the underlying statistical tools rarely take operational constraints into account. By formulating a foundational result in statistics, the Neyman-Pearson lemma, as an optimization problem, we are able to account for important practical considerations, such as costs, and noisy or incomplete input data. Data is scarce in many healthcare applications, so conclusions based on simple statistical approaches might overfit the data. In our case, a naïve in-sample analysis would always conclude that adding more tests to a combination would improve diagnostic accuracy. Our robust formulation builds in safeguards against policy decisions based on overfitting to limited data. As such, our paper illustrates the power of optimization as a tool to integrate statistics and operations.

Beyond the COVID-19 pandemic, our method provides a theoretically-grounded and robust way to achieve higher diagnostic accuracy from combining cheap and fast predictors, and could be relevant in other areas of medicine, healthcare operations, manufacturing, or service systems.

Acknowledgments

We thank the Imperial team, etc.

References

Adams, Emily R, Mark Ainsworth, Rekha Anand, Monique I Andersson, Kathryn Auckland, J Kenneth Baillie, Eleanor Barnes, Sally Beer, John I Bell, Tamsin Berry. 2020. Antibody testing for COVID-19: a report from the national COVID scientific advisory panel. *Wellcome Open Research* **5**.

Alagoz, Oguzhan, Jagpreet Chhatwal, Elizabeth S Burnside. 2013. Optimal policies for reducing unnecessary follow-up mammography exams in breast cancer diagnosis. *Decision Analysis* **10**(3) 200–224.

Alizamir, Saed, Francis De Véricourt, Peng Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157–171.

Aprahamian, Hrayer, Douglas R Bish, Ebru K Bish. 2019. Optimal risk-based group testing. *Management Science* **65**(9) 4365–4384.

Armstrong, Stephen. 2020. Why COVID-19 antibody tests are not the game changer the UK government claims. *BMJ* **369**.

Barreno, Marco, Alvaro Cardenas, J Doug Tygar. 2007. Optimal ROC curve for a combination of classifiers. *Advances in Neural Information Processing Systems* **20** 57–64.

Bastani, Hamsa, Kimon Drakopoulos, Vishal Gupta, Ioannis Vlachogiannis, Christos Hadjicristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis, Sotirios Tsiodras. 2021. Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature* **599**(7883) 108–113.

Ben-Tal, Aharon, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59**(2) 341–357.

Ben-Tal, Aharon, Dick den Hertog, Jean-Philippe Vial. 2015. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming* **149**(1) 265–299.

Berend, Daniel, Aryeh Kontorovich. 2012. On the convergence of the empirical distribution. arXiv preprint arXiv:1205.6711.

Bertsimas, Dimitris, Dick Den Hertog, Jean Pauphilet. 2021. Probabilistic guarantees in robust optimization. *SIAM Journal on Optimization* **31**(4) 2893–2920.

Bertsimas, Dimitris, Jack Dunn. 2017. Optimal classification trees. *Machine Learning* **106**(7) 1039–1082.

Blacksell, Stuart D, Richard G Jarman, Mark S Bailey, Ampai Tanganuchitcharnchai, Kemajittra Jenjaroen, Robert V Gibbons, Daniel H Paris, Ranjan Premaratna, H Janaka de Silva, David G Lalloo. 2011. Evaluation of six commercial point-of-care tests for diagnosis of acute dengue infections: The need for combining NS1 antigen and IgM/IgG antibody detection to achieve acceptable levels of accuracy. *Clinical and Vaccine Immunology* **18**(12) 2095–2101.

Böger, Beatriz, Mariana M Fachi, Raquel O Vilhena, Alexandre de Fátima Cobre, Fernanda S Tonin, Roberto Pontarolo. 2021. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *American Journal of Infection Control* **49**(1) 21–29.

Breiman, Leo. 1996. Bagging predictors. *Machine Learning* **24**(2) 123–140.

Bubar, Kate M, Kyle Reinholt, Stephen M Kissler, Marc Lipsitch, Sarah Cobey, Yonatan H Grad, Daniel B Larremore. 2021. Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science* **371**(6532) 916–921.

Buhat, Christian Alvin H, Jessa Camille C Duero, Edd Francis O Felix, Jomar F Rabajante, Jonathan B Mamplata. 2021. Optimal allocation of COVID-19 test kits among accredited testing centers in the Philippines. *Journal of Healthcare Informatics Research* **5**(1) 54–69.

Calabrese, Justin M, Jeffery Demers. 2022. How optimal allocation of limited testing capacity changes epidemic dynamics. *Journal of Theoretical Biology* 111017.

Cario, Marne C, Barry L Nelson. 1997. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Tech. rep., Citeseer.

Cheng, Jianqiang, Erick Delage, Abdel Lisser. 2014. Distributionally robust stochastic knapsack problem. *SIAM Journal on Optimization* **24**(3) 1485–1506.

Corman, Victor M, Verena Claudia Haage, Tobias Bleicker, Marie Luisa Schmidt, Barbara Mühlemann, Marta Zuchowski, Wendy K Jo, Patricia Tscheak, Elisabeth Möncke-Buchner, Marcel A Müller. 2021. Comparison of seven commercial SARS-CoV-2 rapid point-of-care antigen tests: A single-centre laboratory evaluation study. *The Lancet Microbe* **2**(7) e311–e319.

Dai, Tinglong, Shubhranshu Singh. 2021. Overdiagnosis and undertesting for infectious diseases. Working paper. Available at `https://ssrn.com/abstract=3725057`.

Dantzig, George B. 1957. Discrete-variable extremum problems. *Operations Research* **5**(2) 266–288.

Deeks, Jon, Angela Raffle, Mike Gill. 2021. COVID-19: Government must urgently rethink lateral flow test roll out. *BMJ Opinion* **12**.

Drakopoulos, Kimon, Ramandeep S Randhawa. 2021. Why perfect tests may not be worth waiting for: Information as a commodity. *Management Science* **67**(11) 6678–6693.

Džeroski, Saso, Bernard Ženko. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine Learning* **54**(3) 255–273.

FDA. 2020. Coronavirus (COVID-19) Update: FDA Authorizes Antigen Test as First Over-The-Counter Fully At-Home Diagnostic Test for COVID-19. U.S. Food and Drug Administration News Release, December 15th.

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

Gelman, Andrew, Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gibani, Malick M, Christofer Toumazou, Mohammadreza Sohbati, Rashmita Sahoo, Maria Karvela, Tsz-Kin Hon, Sara De Mateo, Alison Burdett, KY Felice Leung, Jake Barnett, et al. 2020. Assessing a novel, lab-free, point-of-care test for SARS-CoV-2 (CovidNudge): a diagnostic accuracy study. *The Lancet Microbe* **1**(7) e300–e307.

Günlük, Oktay, Jayant Kalagnanam, Minhan Li, Matt Menickelly, Katya Scheinberg. 2021. Optimal decision trees for categorical data via integer programming. *Journal of Global Optimization* **81**(1) 233–260.

Gupta, Niharika, Shine Augustine, Tarun Narayan, Alan ORiordan, Asmita Das, D Kumar, John HT Luong, Bansi D Malhotra. 2021. Point-of-care PCR assays for COVID-19 detection. *Biosensors* **11**(5) 141.

Hershey, John C, Randall D Cebul, Sankey V Williams. 1986. Clinical guidelines for using two dichotomous tests. *Medical Decision Making* **6**(2) 68–78.

Huber, Peter J. 1965. A robust version of the probability ratio test. *The Annals of Mathematical Statistics* 1753–1758.

Huber, Peter J, Volker Strassen. 1973. Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics* 251–263.

International Air Transport Association. 2021. High testing costs could stall the recovery URL `https://www.iata.org/en/pressroom/pr/2021-05-04-03/`. Accessed on 2021-05-04.

Kasy, Maximilian, Alexander Teytelboym. 2020. Adaptive targeted infectious disease testing. *Oxford Review of Economic Policy* **36**(Supplement_1) S77–S93.

Kremer, Mirko, Francis de Véricourt. 2022. Mismanaging diagnostic accuracy under congestion. *Operations Research* doi:10.1287/opre.2022.2292.

Larremore, Daniel B, Bryan Wilder, Evan Lester, Soraya Shehata, James M Burke, James A Hay, Milind Tambe, Michael J Mina, Roy Parker. 2021. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science Advances* **7**(1) eabd5393.

Levi, Retsef, Thomas Magnanti, Yaron Shaposhnik. 2019. Scheduling with testing. *Management Science* **65**(2) 776–793.

Marshall, Roger J. 1989. The predictive value of simple rules for combining two diagnostic tests. *Biometrics* 1213–1222.

McCall, John J. 1965. Maintenance policies for stochastically failing equipment: A survey. *Management Science* **11**(5) 493–524.

McIntosh, Martin W, Margaret Sullivan Pepe. 2002. Combining several screening tests: optimality of the risk score. *Biometrics* **58**(3) 657–664.

Meinshausen, Nicolai, Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4) 417–473.

Mina, Michael J, Roy Parker, Daniel B Larremore. 2020. Rethinking COVID-19 test sensitivity—A strategy for containment. *New England Journal of Medicine* **383**(22) e120.

Monaci, Michele, Ulrich Pferschy. 2013. On the robust knapsack problem. *SIAM Journal on Optimization* **23**(4) 1956–1982.

National Audit Office. 2020. The government's approach to test and trace in england interim report URL `https://www.nao.org.uk/wp-content/uploads/2020/12/The-governments-approach-to-test-and-trace-in-England-interim-report.pdf`. Accessed on 2020-12-11.

Newcombe, Robert G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**(8) 857–872.

Neyman, Jerzy, Egon Sharpe Pearson. 1933. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**(694-706) 289–337.

Paltiel, A David, Amy Zheng, Rochelle P Walensky. 2020. Assessment of SARS-CoV-2 screening strategies to permit the safe reopening of college campuses in the united states. *JAMA Network Open* **3**(7) e2016818–e2016818.

Petherick, Anna. 2020. Developing antibody tests for SARS-CoV-2. *The Lancet* **395**(10230) 1101–1102.

Politico. 2021. Vaccines' success could undercut Biden's multibillion-dollar school testing plans URL `https://www.politico.com/news/2021/05/23/covid-vaccine-testing-schools-490213`. Accessed on 2021-05-03.

Pulia, Michael S, Terrence P OBrien, Peter C Hou, Andrew Schuman, Robert Sambursky. 2020. Multi-tiered screening and diagnosis strategy for COVID-19: A model for sustainable testing capacity in response to pandemic. *Annals of Medicine* **52**(5) 207–214.

Ramdas, Kamalini, Ara Darzi, Sanjay Jain. 2020. Test, re-test, re-test: Using inaccurate tests to greatly increase the accuracy of COVID-19 testing. *Nature Medicine* **26**(6) 810–811.

Rubin, Rita. 2020. The challenges of expanding rapid tests to curb COVID-19. *JAMA* **324**(18) 1813–1815.

Shuren, Jeffrey, Timothy Stenzel. 2021. The FDAs experience with COVID-19 antibody tests. *New England Journal of Medicine* **384**(7) 592–594.

Smith, Rebecca L, Laura L Gibson, Pamela P Martinez, Ruian Ke, Agha Mirza, Madison Conte, Nicholas Gallagher, Abigail Conte, Leyi Wang, Richard Fredrickson, Darci C Edmonson, Melinda E Baughman, Karen K Chiu, Hannah Choi, Tor W Jensen, Kevin R Scardina, Shannon Bradley, Stacy L Gloss, Crystal Reinhart, Jagadeesh Yedetore, Alyssa N Owens, John Broach, Bruce Barton, Peter Lazar, Darcy Henness, Todd Young, Alastair Dunnett, Matthew L Robinson, Heba H Mostafa, Andrew Pekosz, Yukari C Manabe, William J Heetderks, David D McManus, Christopher B Brooke. 2021. Longitudinal Assessment of Diagnostic Test Performance Over the Course of Acute SARS-CoV-2 Infection. *The Journal of Infectious Diseases* **224**(6) 976–982. doi:10.1093/infdis/jiab337. URL `https://doi.org/10.1093/infdis/jiab337`.

Sox, Harold C, Michael C Higgins, Douglas K Owenst. 2013. *Medical Decision Making*. 2nd ed. Wiley.

The Economic Times. 2021. Labs in Dehli halt services over delay in giving out tests results amid COVID-19 spike URL `https://economictimes.indiatimes.com/news/india/labs-in-delhi-halt-services-over-delay-in-giving-out-test-results-amid-covid-19-spike/articleshow/82173036.cms`. Accessed on 2021-04-21.

Transportation Security Agency. 2021. Tsa checkpoint travel numbers. URL `https://www.tsa.gov/coronavirus/passenger-throughput`. Accessed on 2021-06-28.

U.K. Department of Education. 2021. Attendance in education and early years settings during the coronavirus (COVID-19) outbreak URL `https://explore-education-statistics.service.gov.uk/find-statistics/attendance-in-education-and-early-years-settings-during-the-coronavirus-covid-19-outbreak/2021-week-26`. Accessed on 2021-06-28.

Van Parys, Bart PG, Peyman Mohajerin Esfahani, Daniel Kuhn. 2021. From data to decisions: Distributionally robust optimization is optimal. *Management Science* **67**(6) 3387–3402.

Veyrenche, Nicolas, Karine Bolloré, Amandine Pisoni, Anne-Sophie Bedin, Anne-Marie Mondain, Jacques Ducos, Michel Segondy, Brigitte Montes, Patrick Pastor, David Morquin. 2021. Diagnosis value of SARS-CoV-2 antigen/antibody combined testing using rapid diagnostic tests at hospital admission. *Journal of Medical Virology* **93**(5) 3069–3076.

Wang, Hongzhou. 2002. A survey of maintenance policies of deteriorating systems. *European Journal of Operational Research* **139**(3) 469–489.

Weinstein, Milton C, Harvey V Fineberg. 1980. *Clinical Decision Analysis*. W.B. Saunders Co.

Wise, Jacqui. 2020. COVID-19: Safety of lateral flow tests questioned after they are found to miss half of cases. *British Medical Journal (Online)* **371**.

Wolpert, David H. 1992. Stacked generalization. *Neural Networks* **5**(2) 241–259.

Xiao, Roy, Vinay K Rathi. 2021. Price transparency for COVID-19 testing among top us hospitals. *Journal of General Internal Medicine* **36**(1) 245–247.

Yang, Luyi, Shiliang Cui, Zhongbin Wang. 2022. Design of covid-19 testing queues. *Production and Operations Management* **31**(5) 2204–2221. doi:https://doi.org/10.1111/poms.13673.

Appendix for "Robust Combination Testing: Methods and Application to COVID-19 Detection"

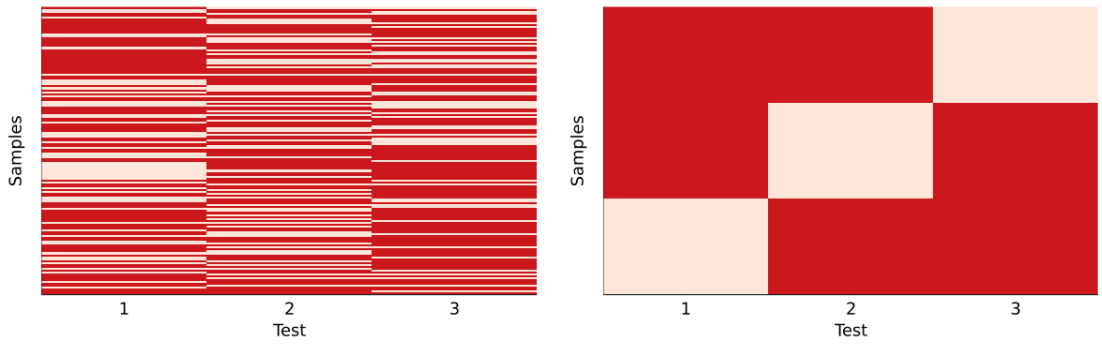
A1. Why combining tests helps: Two stylized examples for intuition

Intuitively, index tests may disagree for distinct reasons. On the one hand, they may not agree because of the different underlying technology they use (e.g., the target protein or the reagents applied) and because they have different underlying capability to identify the same target protein.[7] In this case test outcomes would have (a potentially negative) non-zero correlation. On the other hand, they may disagree due to stochastic noise associated with any testing process, in which case their results might disagree but have zero correlation.

Consequently, our data-driven approach for combining tests can improve diagnostic accuracy through two mechanisms, corresponding to the two types of potential reasons for disagreement. First, a systematic mechanism, since our approach leverages systematic differences in the technology of tests. Second, a stochastic mechanism, since combining tests can reduce the inherent noise associated with testing. To illustrate these two mechanisms, we consider two extreme situations, depicted in Figure A1. For illustration, we focus our attention on condition-positive samples. In both situations, the three index tests (depicted on the X-axis) each have a sensitivity of $2/3$. Each horizontal band in the figure (corresponding to a specific Y value) represents the results on a particular sample, from the 3 tests. Red indicates a positive and pink a negative test result.

- First consider the case in which there are no systematic or technological differences across tests, just stochastic noise in the test outcomes. We model the three tests as independent Bernoulli random variables, each with a sensitivity of $2/3$ (panel (a) in Figure A1). Hence, two particular tests (e.g., tests 1 and 3) provide the same result with probability $(1/3)^2 + (2/3)^2 = 5/9$ $(> 1/2)$, i.e., they agree on more than half the cases, but have zero correlation, by construction.
- Second, consider a case in which there are clear systematic differences across tests. Each test is completely ineffective on a (different) third of the population and perfectly accurate on the

[7] As an example, the specificity of an antibody test depends on its ability to (i) detect the biomarker of interest (i.e., an antibody that is developed in the human body in reaction to Sars-COV-2) and (ii) rule out other 'cross-reactive' biomarkers (i.e., antibodies developed in the body in reaction to other viruses). In the case of Sars-COV-2 antibody tests, prior coronavirus infections are the source of some of the potential cross-reactive biomarkers. Antibody tests operate by using a piece of the virus to detect antibodies in a blood sample. Many tests focus on detecting antibodies to the spike protein of Sars-COV-2. A recent Lancet article (Petherick 2020) notes that "There is a lot hanging on the uniqueness of the spike protein. In terms of the specificity of serological tests in which it is used, the more unique it is, the lower the odds of cross-reactivity with other coronavirusesfalse positives resulting from immunity to other coronaviruses. [...] the potential for cross-reactivity really depends on whether the new tests select sections of the spike protein that are particularly distinct across coronaviruses."



(a) Uncorrelated tests (b) Systematically discordant tests

Figure A1 Test results for condition-positive patients (red is positive) in two stylized situations.

remaining two thirds (panel (b) in Figure A1).[8] In this case, two particular tests (e.g., tests 1 and 3) provide the same results with probability $1/3$ ($< 1/2$) but the results are correlated (pairwise correlation between any two tests is $-1/2$).

A combination testing policy requires a) choosing the tests to combine and b) deciding how to interpret the outcome sequence of the selected tests. For illustration, we assume that we wish to combine all three tests and that the decision rule we apply is the `majority` rule.

- For the first case (uncorrelated but noisy test outcomes) applying the `majority` rule yields a sensitivity of $20/27 > 2/3$. In this situation, each test corresponds to a noisy observation of the quantity of interest and combining (in the case of the `majority` rule, 'averaging') test results achieves higher predictive power due to denoising.
- For the second case (technologically different tests) applying the `majority` rule yields a sensitivity of 1. In this situation, there is no stochasticity in the testing procedure per se but each test sees only a portion of the full picture. Combining the results helps because it aggregates different perspective on the same underlying phenomenon.

In summary, in both situations, combining 3 tests with a majority rule strictly improves diagnosic accuracy, yet through different mechanisms. Although useful for building intuition, the distinction between these two sources of disagreement is hypothetical. Reality is likely to be a mixture of the two extreme situations and it might be hard (if not impossible) to identify which factors are systematic vs. stochastic. Fortunately, doing this is not needed to apply our approach, which takes the correlation structure observed in the data as its primitive. Our two examples are also simplistic

[8] Referring back to the example provided in the prior footnote, consider a population of individuals who each have cross-reactive antibodies to three prior coronaviruses. A pattern similar to that shown in Figure A1(b) can emerge on condition negative samples if the Sars-COV-2 protein segment targeted by each of the three tests overlaps with that in (a different) one of these three prior coronaviruses. For a similar motivation related to antigen testing, please see Corman et al. [5], Table 3, where the tests examined clearly differ in their ability to rule out cross-reactive coronaviruses.

in the sense that, in practice, accuracy is a trade-off between sensitivity and specificity, i.e., performance on both infected and uninfected populations. Furthermore, a systematic methodology is required to decide which tests to combine and how to interpret their results. The tools developed in this paper specifically achieve these objectives.

A2. Optimal test selection with heuristic classification rules

As previously discussed, all existing heuristics (the `any`, `all`, and `majority` rules) are classification rules, i.e., they indicate how to combine results from different tests but not how to select which tests to combine. On the contrary, our knapsack optimization formulation achieves both objectives. In particular, the binary variables $\boldsymbol{y} \in \{0,1\}^m$ indicate which tests are involved used in the combination while the binary variables $\boldsymbol{x} \in \{0,1\}^n$ define the classification rule. Our methodology simultaneously optimize over $\boldsymbol{y}$ and $\boldsymbol{x}$. Accordingly, we can adapt our formulation to optimally select the tests while forcing the aggregation rule $\boldsymbol{x}$ to be one of the existing heuristics. In this section, we derive additional (linear) constraints that should be added to the definition of the feasible set $\mathcal{X}$ to do so.

Consider one outcome sequence $\boldsymbol{s} \in \{0,1\}^m$, the number of tests used in the combination $\boldsymbol{y}$ that return a positive result in this sequence is given by $\sum_{k\in[m]:s_k=1} y_k$.

Based on this observation, we can prove that the `any` rule corresponds to vectors $(\boldsymbol{x}, \boldsymbol{y})$ such that, for any sequence $\boldsymbol{s} \in \{0,1\}^m$,

$$\frac{1}{m} \sum_{k\in[m]:s_k=1} y_k \leq x_{\boldsymbol{s}} \leq \sum_{k\in[m]:s_k=1} y_k.$$

Indeed, if one of the selected test is positive, $\sum_{k\in[m]:s_k=1} y_k > 0$ and the left-hand side inequality forces $x_{\boldsymbol{s}} > 0$ (so $x_{\boldsymbol{s}} = 1$). Otherwise, the right-hand side imposes $x_{\boldsymbol{s}} = 0$. Note that, on the left-hand side inequality, any valid upper-bound on $\sum_{k\in[m]:s_k=1} y_k$ could be used instead of m, e.g., K in presence of a cardinality constraint $\sum_{k\in[m]} y_k \leq K$.

Classifying a patient as positive if all tests are positive is equivalent to classifying a patient as negative is any test is negative. Hence, we can restrict $\boldsymbol{x}$ to correspond to the `all` rule by imposing that, for any sequence $\boldsymbol{s} \in \{0,1\}^m$,

$$\frac{1}{m} \sum_{k\in[m]:s_k=0} y_k \leq 1 - x_{\boldsymbol{s}} \leq \sum_{k\in[m]:s_k=0} y_k.$$

Finally, we verify that the `majority` rule corresponds to the following set of constraints: For any sequence $\boldsymbol{s} \in \{0,1\}^m$,

$$\frac{1}{m}\left(\sum_{k\in[m]:s_k=1} y_k - \tfrac{1}{2}\sum_{k\in[m]} y_k\right) \leq x_{\boldsymbol{s}} \leq 1 + \frac{1}{m}\left(\sum_{k\in[m]:s_k=1} y_k - \tfrac{1}{2}\sum_{k\in[m]} y_k - \tfrac{1}{2}\right).$$

If $\sum_{k\in[m]:s_k=1} y_k > \frac{1}{2}\sum_{k\in[m]} y_k$, the left-hand side inequality forces $x_{\boldsymbol{s}} > 0$ (so $x_{\boldsymbol{s}} = 1$). In this case, observe that the left-hand side is always lower than 1 so $x_{\boldsymbol{s}} = 1$ is feasible. Moreover, $\sum_{k\in[m]:s_k=1} y_k \geq \frac{1}{2}[\sum_{k\in[m]} y_k + 1]$ so the right-hand side inequality is trivially satisfied. On the other hand, if $\sum_{k\in[m]:s_k=1} y_k \leq \frac{1}{2}\sum_{k\in[m]} y_k$, the right-hand side inequality forces $x_{\boldsymbol{s}} = 0$.

A3. Converting a classification rule into a sequencing strategy

In this section, we provide an optimization formulation that allows us to convert any classification rule $\boldsymbol{x}$ into a sequence of tests to apply. We represent a sequencing strategy as a tree: Each internal node corresponds to applying a test. If the result is positive (resp. negative), we continue to its left (resp. right) child. A terminal node (or leaf) corresponds to a 0/1 label (classify as positive/negative). At a high level, $\boldsymbol{x}$ maps every sequence of $\boldsymbol{s} \in \{0,1\}^m$ to a binary label $x_{\boldsymbol{s}} \in \{0,1\}$ and we represent this mapping as a tree. Hence, our formulation is analogous to mixed-integer formulations for binary classification trees in supervised learning (see, e.g., Günlük et al. 2021, Bertsimas and Dunn 2017).

For a given depth d, we consider the maximal tree of this depth, which has $T = 2(d+1) - 1$ nodes. We index each node by by an integer $t = 1, \ldots, T$, so that internal nodes are indexed by $t \in \mathcal{I} := \{1, \ldots, \lfloor T/2 \rfloor\}$ and leaf nodes by $t \in \mathcal{L} := \{\lfloor T/2 \rfloor + 1, \ldots, T\}$. In our case, d is bounded by the total number of tests involved in the combination $K (\leq m)$ so $T \leq 2K + 1$.

For each internal node $t \in \mathcal{I}$ and each test k we introduce a binary variable $z_{t,k}$ that indicates whether test k is used to split node t. Since we fixed the topology of the tree, we want to allow the option of not splitting at an internal node. We use an indicator variables d_t to indicate whether a split is applied at node t. If not, we use the convention that all the sequences are propagated to the left child-node. With this convention, we should impose the constraints:

$$\sum_{k=1}^{m} z_{t,k} = d_t, \quad \forall t \in \mathcal{I}, \tag{A1}$$

and a sequence $\boldsymbol{s} \in \{0,1\}^m$ is assigned to the left child-node if and only if $\sum_{k=1}^{m} z_{t,k} s_k \leq d_t/2$.

For each leaf node $t \in \mathcal{L}$ and each sequence $\boldsymbol{s}$, we identify the leaf node to which $\boldsymbol{s}$ is assigned to via the binary variable $\ell_{\boldsymbol{s},t}$. For a leaf node $t \in \mathcal{L}$, define $\mathcal{A}_{left}(t)$ (resp. $\mathcal{A}_{right}(t)$) as the set of ancestors of t whose left (resp. right) branch has been followed on the path from the root node to t. Then, we have the following constraints:

$$\begin{aligned}
&\sum_{k=1}^{m} z_{u,k} s_k \leq d_u/2 + (1 - \ell_{\boldsymbol{s},t}), \quad \forall \boldsymbol{s},\ \forall t \in \mathcal{L},\ \forall u \in \mathcal{A}_{left}(t), \\
&\sum_{k=1}^{m} z_{u,k} s_k \geq d_u/2 - (1 - \ell_{\boldsymbol{s},t}), \quad \forall \boldsymbol{s},\ \forall t \in \mathcal{L},\ \forall u \in \mathcal{A}_{right}(t).
\end{aligned} \tag{A2}$$

Finally, we want the tree to perfectly describe the rule $\boldsymbol{x}$ so we impose the constraint that each leaf is composed of sequences classified in the same way. Let us introduce a binary variable $x_t^{tree} \in \{0,1\}$ to encode for the classification of leaf t. We impose

$$\begin{aligned} x_t^{tree} &\leq x_{\boldsymbol{s}} \ell_{\boldsymbol{s},t} + (1 - \ell_{\boldsymbol{s},t}), \quad \forall \boldsymbol{s},\ \forall t \in \mathcal{L}, \\ x_t^{tree} &\geq x_{\boldsymbol{s}} \ell_{\boldsymbol{s},t} - (1 - \ell_{\boldsymbol{s},t}), \quad \forall \boldsymbol{s},\ \forall t \in \mathcal{L}. \end{aligned} \tag{A3}$$

All in all, we obtain the minimal tree representation of the classification rule via solving the following binary linear optimization problem

$$\min_{d_t, z_{t,k}, \ell_{\boldsymbol{s},t}, x_t^{tree}} \quad \sum_{t \in \mathcal{I}} d_t \text{ s.t. (A1)} - \text{(A2)} - \text{(A3)}$$

with $2K + K^2 + K2^K$ binary variables and $K + K^2 2^K + K2^K$ constraints.

A4. Proofs and Omitted Results

In this section we include proofs and additional results that were omitted in the main body of the manuscript. We first provide a proof for Proposition 1, then provide a proof for Lemma 2, before stating and proving Lemma A1, which is referred to at the end of §3 in the main body of the paper.

Proof of Proposition 1 Decompose $\mathcal{U}_{\hat{\boldsymbol{a}}}$ into $\mathcal{U}_{\hat{\boldsymbol{a}}} = \mathcal{U}_0 \cap \mathcal{U}_\Gamma$ with $\mathcal{U}_0 = \{\boldsymbol{a} : \left|\sum_{\boldsymbol{s}: s_k = 1} a_{\boldsymbol{s}} - \hat{p}_k\right| \leq \hat{\rho}_k, \forall k\}$ and $\mathcal{U}_\Gamma = \{\boldsymbol{a} \geq 0 : \boldsymbol{e}^\top \boldsymbol{a} = 1,\ \sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \leq \Gamma\}$. Then,

$$\mathbb{P}\left(\boldsymbol{a}_0^\top \boldsymbol{x} \geq \min_{\boldsymbol{a} \in \mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x}\right) \geq \mathbb{P}\left(\boldsymbol{a}_0 \in \mathcal{U}_0 \cap \mathcal{U}_\Gamma\right) \geq 1 - \mathbb{P}\left(\boldsymbol{a}_0 \notin \mathcal{U}_0\right) - \mathbb{P}\left(\boldsymbol{a}_0 \notin \mathcal{U}_\Gamma\right),$$

from Boole's inequality. Assumption 1 yields $\mathbb{P}\left(\boldsymbol{a}_0 \notin \mathcal{U}_0\right) \leq m\alpha$. Regarding $\mathbb{P}\left(\boldsymbol{a}_0 \notin \mathcal{U}_\Gamma\right)$, Berend and Kontorovich (2012, Theorem 2) states that, for any $r \geq \sqrt{\frac{n}{N^+}}$,

$$\mathbb{P}\left(\|\hat{\boldsymbol{a}} - \boldsymbol{a}_0\|_1 > r\right) \leq \exp\left(-\frac{N^+}{2}\left[r - \sqrt{\frac{n}{N^+}}\right]^2\right).$$

Setting $r = \Gamma$ concludes the proof. □

Proof of Lemma 2 As in the proof of Proposition 1, we decompose $\mathcal{U}_{\hat{\boldsymbol{a}}}$ into $\mathcal{U}_0 \cap \mathcal{U}_\Gamma$. From Ben-Tal et al. (2015, Lemma 6.4),

$$\min_{\boldsymbol{a} \in \mathcal{U}_0 \cap \mathcal{U}_\Gamma} \boldsymbol{a}^\top \boldsymbol{x} = \max_{\boldsymbol{x}^0, \boldsymbol{x}^\Gamma : \boldsymbol{x}^0 + \boldsymbol{x}^\Gamma = \boldsymbol{x}} \left[\min_{\boldsymbol{a} \in \mathcal{U}_0} \boldsymbol{a}^\top \boldsymbol{x}^0\right] + \left[\min_{\boldsymbol{a} \in \mathcal{U}_\Gamma} \boldsymbol{a}^\top \boldsymbol{x}^\Gamma\right],$$

and we can treat the two uncertainty sets separately. For $\mathcal{U}_0$, stong linear duality yields

$$\begin{aligned} \min_{\boldsymbol{a} \in \mathcal{U}_0} \boldsymbol{a}^\top \boldsymbol{x}^0 &= \min_{\boldsymbol{a}} \boldsymbol{a}^\top \boldsymbol{x}^0 \text{ s.t. } \hat{p}_k - \hat{\rho}_k \leq \sum_{\boldsymbol{s}: s_k = 1} a_{\boldsymbol{s}} \leq \hat{\rho}_k + \hat{p}_k, \forall k = 1, \ldots, m, \\ &= \max_{\boldsymbol{q}^1, \boldsymbol{q}^2 \geq 0} -(\hat{\boldsymbol{\rho}} - \hat{\boldsymbol{p}})^\top \boldsymbol{q}^1 - (\hat{\boldsymbol{\rho}} + \hat{\boldsymbol{p}})^\top \boldsymbol{q}^2 \text{ s.t. } \sum_{k: \boldsymbol{s}_k = 1} q_k^1 - q_k^2 = x_{\boldsymbol{s}}^0, \forall \boldsymbol{s}, \end{aligned}$$

where $\boldsymbol{q}^1$ and $\boldsymbol{q}^2$ are the dual variables associated with the lower and upper bound constraints, respectively.

For $\mathcal{U}_\Gamma$, Corollary 1 of (Ben-Tal et al. 2013) applied to total absolute deviation states that

$$\begin{aligned}\min_{\boldsymbol{a}\in\mathcal{U}_\Gamma} \boldsymbol{a}^\top \boldsymbol{x}^\Gamma &= \min_{\boldsymbol{a}\geq 0, \boldsymbol{e}^\top \boldsymbol{a}=1} \boldsymbol{a}^\top \boldsymbol{x}^\Gamma \text{ s.t. } \sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \leq \Gamma, \\ &= \max_{\lambda\geq 0, r, \boldsymbol{v}} -r - \Gamma\lambda + \hat{\boldsymbol{a}}^\top \boldsymbol{v} \text{ s.t. } \begin{cases} \boldsymbol{v} \leq \lambda, \\ \boldsymbol{v} \leq \boldsymbol{x}^\Gamma + r, \\ -\lambda \leq \boldsymbol{x}^\Gamma + r. \end{cases}\end{aligned}$$

Combining the two yields the following formulation:

$$\begin{aligned}\max_{\substack{\boldsymbol{x}\in\mathcal{X},\ \boldsymbol{x}^0, \boldsymbol{x}^\Gamma \\ \lambda, r, \boldsymbol{v}, \boldsymbol{q}^1, \boldsymbol{q}^2}} \ & \hat{\boldsymbol{a}}^\top \boldsymbol{v} - r - \Gamma\lambda - (\hat{\boldsymbol{\rho}} - \hat{\boldsymbol{p}})^\top \boldsymbol{q}^1 - (\hat{\boldsymbol{\rho}} + \hat{\boldsymbol{p}})^\top \boldsymbol{q}^2 \text{ s.t. } \boldsymbol{x} = \boldsymbol{x}^0 + \boldsymbol{x}^\Gamma, \\ & \boldsymbol{v} \leq \min(\lambda, \boldsymbol{x}^\Gamma + r), \\ & -\lambda \leq \min(\boldsymbol{x}^\Gamma + r, 0), \\ & x^0_{\boldsymbol{s}} = \sum_{k:\boldsymbol{s}_k=1} q^1_k - q^2_k, \ \forall \boldsymbol{s}, \\ & \boldsymbol{b}^\top \boldsymbol{x} \leq 1 - t_0, \\ & \lambda \geq 0, \boldsymbol{q}^1, \boldsymbol{q}^2 \geq \boldsymbol{0}.\end{aligned}$$

The above optimization problem is linear, with mixed-integer variables whenever $\boldsymbol{x}$ is binary. □

Lemma A1. *The robust problem*

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left[\min_{\boldsymbol{a}\in\tilde{\mathcal{U}}_{\boldsymbol{a}}} \boldsymbol{a}^\top \boldsymbol{x}\right] \ \textit{s.t. } \boldsymbol{b}^\top \boldsymbol{x} \leq 1 - t_0,$$

is equivalent to the following linear optimization problem

$$\begin{aligned}\max_{\substack{\boldsymbol{x}\in\mathcal{X},\ \boldsymbol{x}^0, \boldsymbol{x}^\Gamma \\ \lambda, r, \boldsymbol{v}, \boldsymbol{q}^1, \boldsymbol{q}^2, \boldsymbol{z}}} \ & \frac{1}{N^+}\boldsymbol{e}^\top \boldsymbol{z} - r - \Gamma\lambda - (\hat{\boldsymbol{\rho}} - \hat{\boldsymbol{p}})^\top \boldsymbol{q}^1 - (\hat{\boldsymbol{\rho}} + \hat{\boldsymbol{p}})^\top \boldsymbol{q}^2 \ \textit{s.t. } \boldsymbol{x} = \boldsymbol{x}^0 + \boldsymbol{x}^\Gamma, \\ & \boldsymbol{v} \leq \min(\lambda, \boldsymbol{x}^\Gamma + r), \\ & -\lambda \leq \min(\boldsymbol{x}^\Gamma + r, 0), \\ & x^0_{\boldsymbol{s}} = \sum_{k:\boldsymbol{s}_k=1} q^1_k - q^2_k, \ \forall \boldsymbol{s}, \\ & z_i \leq v_{\boldsymbol{s}}, \forall \boldsymbol{s} \in \mathcal{S}_i \\ & \boldsymbol{b}^\top \boldsymbol{x} \leq 1 - t_0, \\ & \lambda \geq 0, \boldsymbol{q}^1, \boldsymbol{q}^2 \geq \boldsymbol{0}.\end{aligned}$$

Proof of Lemma A1 Observe that

$$\min_{\boldsymbol{a}\in\tilde{\mathcal{U}}_{\boldsymbol{a}}} \boldsymbol{a}^\top \boldsymbol{x} = \min_{\hat{\boldsymbol{a}} = \frac{\boldsymbol{Z}^\top \boldsymbol{e}}{N^+}, \boldsymbol{Z}\in\mathcal{Z}} \ \min_{\boldsymbol{a}\in\mathcal{U}_{\hat{\boldsymbol{a}}}} \boldsymbol{a}^\top \boldsymbol{x}.$$

From Lemma 2, we reformulate the inner-minimization problem as a maximization problem in $\lambda, r, \boldsymbol{v}, \boldsymbol{q}^1, \boldsymbol{q}^2$. Then, strong duality applies to the minimization with respect to $\hat{\boldsymbol{a}}$:

$$\min_{\hat{\boldsymbol{a}}=\frac{1}{N^+}\boldsymbol{Z}^\top \boldsymbol{e}, \boldsymbol{Z}\in\mathcal{Z}} \hat{\boldsymbol{a}}^\top \boldsymbol{v} = \min_{\boldsymbol{Z}\in\mathcal{Z}} \tfrac{1}{N^+}\boldsymbol{e}^\top \boldsymbol{Z}^\top \boldsymbol{v} = \max_{\boldsymbol{z}} \tfrac{1}{N^+}\boldsymbol{e}^\top \boldsymbol{z} \text{ s.t. } z_i \le v_{\boldsymbol{s}}, \forall \boldsymbol{s}\in\mathcal{S}_i,$$

and the result follows. □

A5. Pseudo-code and implementation details

The knapsack formulation (1) (or its robust variant) returns the feasible combination of tests achieving the best sensitivity subject to having a specificity of at least t_0. In other words, such a classification rule $\boldsymbol{x}^\star(t_0)$ is Pareto-dominating, namely no other classification rule can achieve higher sensitivity without compromising specificity. By solving the same problem for different values of t_0, as described in Algorithm 1, we obtain the set of all Pareto-dominating classification rules.

Algorithm 1: Computation of the set of Pareto-dominating combination policies

Input: Uncertainty sets $\tilde{\mathcal{U}}_{\hat{\boldsymbol{a}}}$ and $\tilde{\mathcal{U}}_{\hat{\boldsymbol{b}}}$

Parameters: Feasible set $\mathcal{X}$ (e.g., maximum number of tests to combine K), grid of values for the minimum required specificity t_0

ParetoPolicies $\leftarrow$ [ ];

for *each value of* t_0 **do**
  Compute $\boldsymbol{x}^\star(t_0)$ the solution of (3) (or (1) in the non-robust case);
  Add $\boldsymbol{x}^\star(t_0)$ to *ParetoPolicies*;
end

return *ParetoPolicies*

In practice, a robust formulation is needed to mitigate the risk of overfitting and account for missing data. In particular, in Section 4.3 and in our numerical experiments, we parametrize the size of the uncertainty sets, $\mathcal{U}_{\hat{\boldsymbol{a}}}(\Gamma)$ and $\mathcal{U}_{\hat{\boldsymbol{b}}}(\Gamma)$, by an hyper-parameter Γ that controls how much we allow the true sensitivity and specificity to deviate from their in-sample estimates. The "right" value of Γ is a priori unknown so we calibrate it using an F-fold cross-validation procedure (we take $F=5$ in our experiments) described in Algorithm 2. Observe that our procedure involves an accuracy metric `accuracy` that can be adapted to the specific application at hand. To keep our exposition as generic as possible, we consider the AUC as our metric of interest but it is not a requirement imposed by our method. In addition, we need to specify a grid of values for Γ. In our implementation, we try uniformly spaced values of Γ of the form $\Gamma = \lambda\Gamma_0$, where $\lambda \in \{0,\dots,10\}$ and Γ_0 is a scaling factor. In particular, we decided to set Γ_0 equal to the smallest (positive) value

for which the set $\{\boldsymbol{a} \geq 0 : \boldsymbol{e}^\top \boldsymbol{a} = 1;\ \sum_{\boldsymbol{s}} |a_{\boldsymbol{s}} - \hat{a}_{\boldsymbol{s}}| \leq \Gamma_0;\ \left|\sum_{\boldsymbol{s}:s_k=1} a_{\boldsymbol{s}} - \hat{p}_k\right| = 0, \forall k\}$ is nonempty. Note that, since we are using Wilson CIs, $\sum_{\boldsymbol{s}:s_k=1} \hat{a}_{\boldsymbol{s}} \neq \hat{p}_k$ and this set is empty for $\Gamma_0 = 0$. In other words, our value of Γ_0 scales like $1/N^+$ and captures the distance between the empirical data, $\hat{\boldsymbol{a}}$, and a vector $\boldsymbol{a}$ that recovers exactly the center of the Wilson CIs.

Algorithm 2: Cross-validation loop for the size of the uncertainty sets

Data: Patient-level results on m tests, $\boldsymbol{S}^{\pm} \in \{0, 1, \texttt{NA}\}^{N^{\pm} \times m}$.

Parameters: Size of the combination K, number of folds F, accuracy metric `accuracy`

Input: Partition of the data into F folds, grid of values for Γ

for *each value of* Γ **do**
  for *each fold* f **do**
    Use the data of the $F-1$ remaining folds to construct the sets $\tilde{\mathcal{U}}_{\hat{\boldsymbol{a}}}(\Gamma)$ and $\tilde{\mathcal{U}}_{\hat{\boldsymbol{b}}}(\Gamma)$;
    Compute $ParetoPolicies(\Gamma, f)$, the set of Pareto-dominating combination rules (Algorithm 1);
    Compute $\texttt{accuracy}(\Gamma, f)$, the accuracy of $ParetoPolicies(\Gamma, f)$ on the f-th fold;
  end
  Compute $\texttt{accuracy}(\Gamma) = \frac{1}{F}\sum_{f \in F} \texttt{accuracy}(\Gamma, f)$;
end
Pick $\Gamma^\star \in \arg\max_\Gamma \texttt{accuracy}(\Gamma)$;
Construct the sets $\tilde{\mathcal{U}}_{\hat{\boldsymbol{a}}}(\Gamma^\star)$ and $\tilde{\mathcal{U}}_{\hat{\boldsymbol{b}}}(\Gamma^\star)$ on the entire data;
Compute $ParetoPolicies(\Gamma^\star)$ for the entire data (Algorithm 1);
return $ParetoPolicies(\Gamma^\star)$

A6. Supporting evidence on COVID-19 testing data

This section provides additional numerical evidence from the COVID-19 antigen and antibody detection datasets.

A6.1. Additional summary statistics

Our optimization-based approach takes advantage of potential systematic disagreement of tests on the same sample and/or lack of correlations between tests, i.e., dependency between the sources of randomness in the results at a patient and population level.

To measure agreement, we compute, for each pair of tests, the probability that they return the same result, which we refer to as the probability of pairwise agreement or simply agreement probabilities. Tables A1 and A2 report the probability of pairwise agreement between tests for the antibody and antigen test data, respectively. On both data, we observe agreement probabilities

close to 1 on condition-negative individuals, suggesting that there is more agreement among tests on samples that are indeed negative. On condition-positive samples, however, agreement probabilities between tests are lower than on condition-negative samples, demonstrating that there is less agreement on how to classify these patients.

Table A1 Probability of pairwise agreement of the 9 antibody tests

Condition-positive individuals

	1	2	3	4	5	6	7	8	9
1	1.0	0.87	0.79	0.76	0.9	0.81	0.85	0.75	0.85
2		1.0	0.81	0.81	0.86	0.83	0.84	0.7	0.97
3			1.0	0.97	0.94	0.84	0.88	0.88	0.82
4				1.0	0.9	0.81	0.85	0.84	0.82
5					1.0	0.9	0.94	0.84	0.9
6						1.0	0.9	0.74	0.87
7							1.0	0.75	0.82
8								1.0	0.72
9									1.0

Condition-negative individuals

	1	2	3	4	5	6	7	8	9
1	1.0	0.98	0.97	0.98	0.97	0.98	0.95	1.0	0.93
2		1.0	0.95	0.97	0.95	0.97	0.93	0.98	0.95
3			1.0	0.98	0.93	0.95	0.92	0.97	0.9
4				1.0	0.95	0.97	0.93	0.98	0.92
5					1.0	0.95	0.98	0.97	0.93
6						1.0	0.93	0.98	0.95
7							1.0	0.95	0.92
8								1.0	0.93
9									1.0

Table A2 Probability of pairwise agreement of the 6 antigen tests

Condition-positive individuals

	I	III	IV	V	VI	VII
I	1.0	0.79	0.61	0.88	0.68	0.96
III		1.0	0.41	0.85	0.48	0.77
IV			1.0	0.56	0.91	0.59
V				1.0	0.62	0.9
VI					1.0	0.67
VII						1.0

Condition-negative individuals

	I	III	IV	V	VI	VII
I	1.0	0.88	0.99	0.94	0.99	0.98
III		1.0	0.89	0.85	0.88	0.89
IV			1.0	0.95	0.99	0.99
V				1.0	0.94	0.93
VI					1.0	0.98
VII						1.0

Regarding correlations, Tables A3 and A4 report the pairwise correlation matrices for the antibody and antigen test data respectively, on condition-positive and condition-negative individuals. For both test types, we observe that the pairwise correlation in test outcomes depends on whether the sample in question is indeed positive or negative. In particular, on condition-negative individuals, pairwise correlations are mostly very close to 0 on both the antigen and antibody test data. Combined with the fact that individual test specificities are high on this sub-population, this observation suggests that, on condition-negative individual samples, all tests are effective and that variability in their individual performance is mostly due to random measurement error and noise, which are independent from one test to another. In contrast, on condition-positive samples, we observe higher pairwise correlation, especially for the antibody test data.

As we mention at the beginning of this section, our methodology is designed to uncover complementary sets of tests that provide better diagnostic accuracy, either by removing noise by

Table A3 Pairwise correlation matrices for the 9 antibody tests

Condition-positive individuals

	1	2	3	4	5	6	7	8	9
1	1.0	0.73	0.58	0.51	0.79	0.59	0.72	0.49	0.7
2		1.0	0.6	0.6	0.7	0.63	0.66	0.39	0.94
3			1.0	0.93	0.88	0.67	0.74	0.75	0.66
4				1.0	0.8	0.58	0.65	0.69	0.65
5					1.0	0.79	0.88	0.66	0.83
6						1.0	0.8	0.46	0.77
7							1.0	0.5	0.68
8								1.0	0.44
9									1.0

Condition-negative individuals

	1	2	3	4	5	6	7	8	9
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2		1.0	-0.03	-0.02	-0.03	-0.02	-0.04	0.0	-0.03
3			1.0	0.7	-0.03	-0.02	-0.04	0.0	-0.08
4				1.0	-0.02	-0.02	-0.03	0.0	-0.05
5					1.0	-0.02	0.81	0.0	0.49
6						1.0	-0.03	0.0	0.74
7							1.0	0.0	0.37
8								1.0	0.0
9									1.0

Table A4 Pairwise correlation matrices for the 6 antigen tests

Condition-positive individuals

	I	III	IV	V	VI	VII
I	1.0	0.54	0.46	0.69	0.54	0.9
III		1.0	0.2	0.63	0.25	0.46
IV			1.0	0.41	0.81	0.42
V				1.0	0.48	0.73
VI					1.0	0.52
VII						1.0

Condition-negative individuals

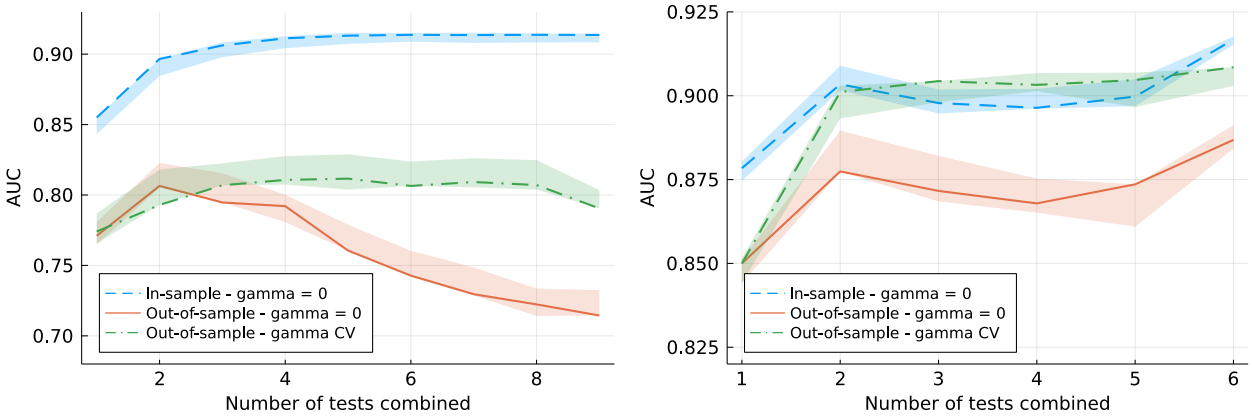
	I	III	IV	V	VI	VII
I	1.0	-0.03	0.0	-0.02	-0.01	-0.01
III		1.0	0.0	0.02	-0.03	0.15
IV			1.0	0.0	0.0	0.0
V				1.0	-0.02	-0.03
VI					1.0	-0.01
VII						1.0

aggregating multiple test outcomes or by exploiting the patterns by which individual tests disagree. For illustration, one would expect the noise-reduction mechanism to be the main driver of improved accuracy in settings where individual tests are independent (i.e., correlation is low) but agreement is high. In the other extreme, one would expect the ability to exploit patterns in test disagreement to be the main driver of improved accuracy in settings with high correlation but low agreement (which would suggest that systematic differences across tests might be driving disagreement). The summary statistics presented in this section illustrate that for both test types (antibody and antigen) both mechanisms are likely to be at play (since the level of agreement and correlation depends on the true condition of the sample, which is not observable in practice). As a result, it is not obvious which mechanism drives the results we observe, just by looking at summary statistics. Rather, for each dataset, one must apply our methodology to learn to what extent combining different tests will improve diagnostic accuracy.

A6.2. Additional results on COVID-19 testing data

First, we assess the robustness of the behavior depicted in Figure 3 to the splitting proportion. We replicate the experiments from Section 5.3, splitting the data into training and validation, in proportions $p_0/1-p_0$ for $p_0 \in \{55\%, 60\%, \ldots, 95\%\}$. For a fixed proportion p_0, one can represent AUC as a function of the number of tests combined, as in Figure 3. Figure A2 displays the area covered by the average AUC when varying p_0, for increasing values of K. As in Section 5.3,

Figure A2 **Robustness of the AUC of the best combination of K tests to the training/validation proportion, as the number of tests K increases, for the non-robust combination policy (in-sample in dash blue, out-of-sample in solid orange) and the robust policies with cross-validated γ (dash-dot green).**



(a) Antibody test data

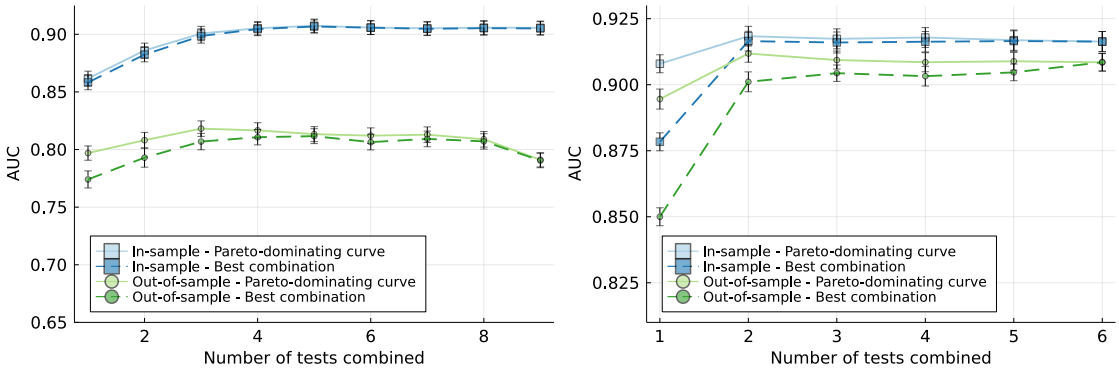
(b) Antigen test data

we observe that the non-robust combinations of tests suffer from overfitting, and that adding robustness improves out-of-sample AUC significantly.

Second, while Figure 3 displays the AUC of the best K-test combination, it only represents a fraction of the Pareto-dominating ROC curve. As discussed in Section 4.3, the AUC of the entire Pareto-dominating curve captures the current state of the technology, i.e., the combined predictive power of the available tests, and could be of independent interest. Figure A3 compares the AUC (both in- and out-of-sample) of the entire Pareto-dominating curve with the one for the best K-test combination only, for the robust formulations. We observe that most of the area covered by the Pareto-dominating curve is recovered by the best combination, and that the AUC of the entire curve follows a similar trend to that of the best combination. For instance, combining 3 tests increase total AUC from 0.797 to 0.818 (+3%) on the antibody test data, and yields a +2% improvement on the antigen test data (0.895 vs. 0.909)

In Section 5.2 and 5.5, we compared our knapsack formulation to individual test performance and heuristic rules. Tables A5-A6 report the average sensitivity values (and their standard errors) displayed on Figure 5. Heuristic rules are heuristic in the way they select which tests to combine and how the selected tests are combined (`majority`, `any` or `all` rule). For test selection, we decided to combine the 3 tests that individually achieve the highest individual AUC. In Figure A4, we replicate the in-sample complete-case analysis of Figure 2 and compare this heuristic test selection rule (in orange) with a situation where we choose to combine tests that individually lie on the frontier (tests 1, 6, and 7 for the antibody test data; and I, III, and IV for the antigen test data) in yellow, and a situtation where we choose the 3-test combination leading the the highest AUC in green. In

Figure A3 **Comparison of the diagnosis accuracy (AUC) of the best combination of K tests and the entire Pareto-dominating curve as a function of the number of tests K combined.**



(a) Antibody test data (b) Antigen test data

Figure notes: Each line corresponds to the estimated AUC as the number of tests K increases. The blue (green) curves are the in-sample (out-of-sample) estimate of the AUC. Dash lines correspond to the performance of the best K-test combination while solid lines correspond to the Pareto-dominating ROC curve. Error bars correspond to $\pm$ 1 standard error.

Table A5 **Out-of-sample sensitivity (and standard errors) achieved by the best single test, heuristic 3-test combinations, and optimal 3-test combinations on the antibody test data. For the single tests and the heuristics, we test whether the value is significantly different from the optimal 3-test combination and report p-values of a paired t-test with unequal variance.**

	Specificity threshold			
	0.8	0.9	0.95	0.98
Robust Optimal Combination	63.7% (1.5%)	63.5% (1.5%)	**62.6% (1.5%)**	**56.4% (2.3%)**
Best Individual Test	57.4% (1.6%) $< 4.4 \cdot 10^{-3}$	57.4% (1.6%) $< 6.0 \cdot 10^{-3}$	57.4% (1.6%) $< 1.8 \cdot 10^{-2}$	53.1% (1.4%) $< 2.3 \cdot 10^{-1}$
Heuristic `all`	47.3% (1.3%) $< 5.9 \cdot 10^{-13}$	47.3% (1.3%) $< 1.3 \cdot 10^{-12}$	47.3% (1.3%) $< 7.3 \cdot 10^{-12}$	47.3% (1.3%) $< 9.3 \cdot 10^{-4}$
Heuristic `any`	**67.8% (1.4%)** $< 5.0 \cdot 10^{-2}$	**66.7% (1.9%)** $< 2.0 \cdot 10^{-1}$	51.3% (4.3%) $< 1.6 \cdot 10^{-2}$	0.0% (0.0%) $< 0.1 \cdot 10^{-29}$
Heuristic `majority`	57.8% (1.4%) $< 4.5 \cdot 10^{-3}$	57.8% (1.4%) $< 6.3 \cdot 10^{-3}$	57.8% (1.4%) $< 2.0 \cdot 10^{-2}$	39.7% (4.2%) $< 8.9 \cdot 10^{-4}$

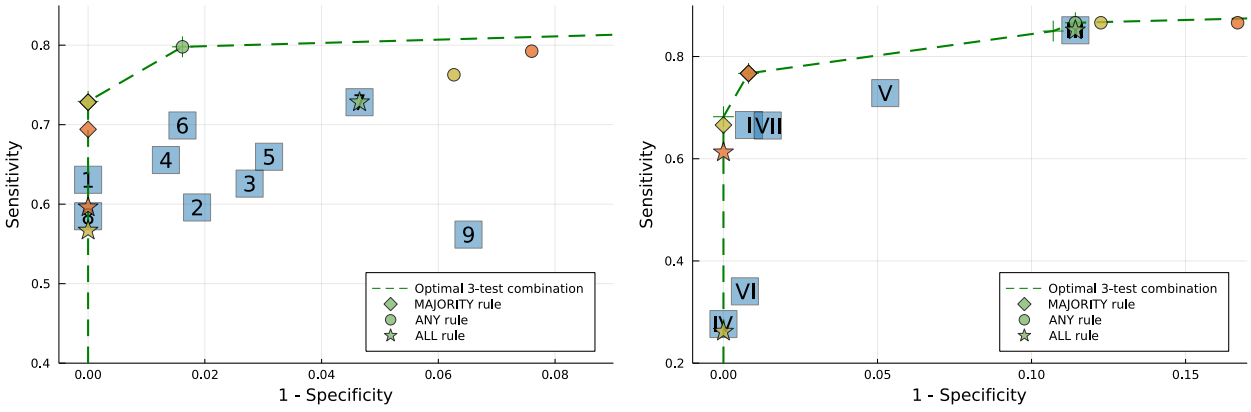
all cases, we observe that these heuristic rules do not systematically lie on the optimal ROC curve and are (at least weakly) dominated by our optimization-based approach. In addition, although the optimal selection of 3 tests can be easily computed in this case via exhaustive enumeration, this brute force approach is sensitive to noise and finite-sample estimation error and fails to account for missing data, unlike our robust knapsack formulation.

In Section 5.5, we tried to untwine the benefit of our approach coming from the test subset selection part and the result aggregation part. In Figure A5, we consider the entire robust Pareto-

Table A6 **Out-of-sample sensitivity (and standard errors) achieved by the best single test, heuristic 3-test combinations, and optimal 3-test combinations on the antigen test data. For the single tests and the heuristics, we test whether the value is significantly different from the optimal 3-test combination and report p-values of a paired t-test with unequal variance.**

	Specificity threshold			
	0.8	0.9	0.95	0.98
Robust Optimal Combination	**84.9% (0.7%)**	**79.0% (0.9%)**	**75.0% (0.9%)**	62.4% (2.6%)
Best Individual Test	83.9% (0.7%) $< 3.0 \cdot 10^{-1}$	76.8% (1.0%) $< 1.1 \cdot 10^{-1}$	72.4% (0.8%) $< 3.3 \cdot 10^{-2}$	59.5% (2.6%) $< 4.5 \cdot 10^{-1}$
Heuristic `all`	62.7% (1.0%) $< -0.0 \cdot 10^{-32}$	62.7% (1.0%) $< 0.1 \cdot 10^{-21}$	62.7% (1.0%) $< 6.7 \cdot 10^{-15}$	**62.7% (1.0%)** $< 9.1 \cdot 10^{-1}$
Heuristic `any`	77.0% (3.3%) $< 2.5 \cdot 10^{-2}$	19.1% (4.9%) $< 8.2 \cdot 10^{-17}$	1.5% (1.5%) $< 0.1 \cdot 10^{-56}$	0.0% (0.0%) $< 0.1 \cdot 10^{-28}$
Heuristic `majority`	73.3% (0.9%) $< 3.2 \cdot 10^{-16}$	73.3% (0.9%) $< 3.7 \cdot 10^{-5}$	73.3% (0.9%) $< 2.0 \cdot 10^{-1}$	50.3% (4.8%) $< 3.0 \cdot 10^{-2}$

Figure A4 **Performance of the `majority` (diamond), `any` (circle) and `all` (star) rules on a subset of 3 tests selected that achieve highest individual AUC (orange), are cherry-picked to lie on the frontier (yellow), or achieve the highest AUC when combined (green).**



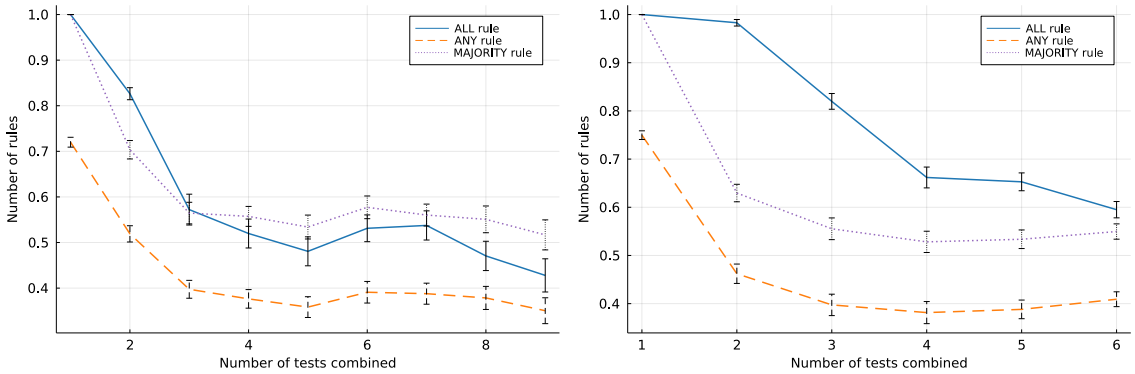
(a) Antibody test data

(b) Antigen test data

optimal set of K-test combinations, count how many of them correspond to one of the aforementioned heuristics on the subset of tests they involve, and report the frequency of occurrence of each heuristic rule separately.

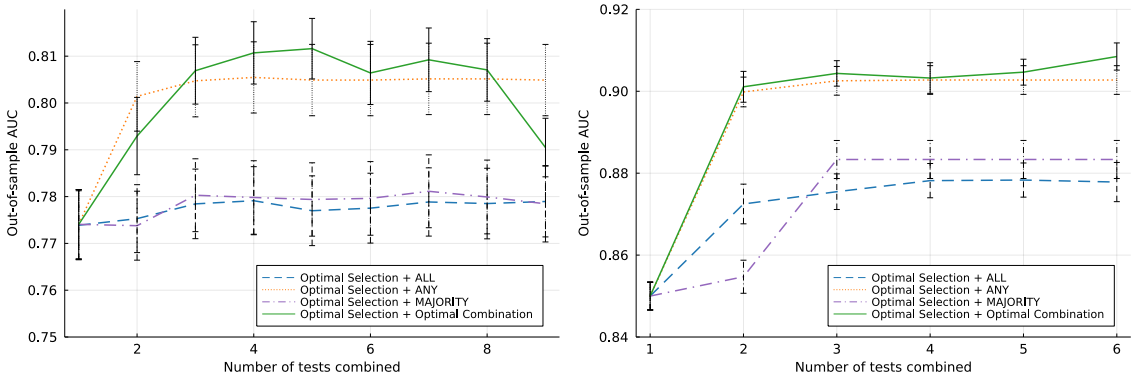
In a complementary direction, we solve the robust optimization problem (3) with the additional constrained described in Appendix A2 to constraint the classification rule $\boldsymbol{x}$ to be one of the aforementioned heuristics. Figure A6 compares the out-of-sample AUC of the three heuristics compared with the optimal rule, as the number of tests combined, K, increases. We observe that optimally combining tests substantially increases the performance of the heuristics, with the `any`

Figure A5 **Fraction of the robust Pareto-optimal K-test combinations that correspond to one of the `all`, `any`, and `majority` rule (`all`/`any`/`majority`).**



(a) Antibody test data (b) Antigen test data

Figure A6 **Out-of-sample AUC as the number of tests combined K increases, for three heuristic rules and the optimal classification rule. For each classification rule, tests are selected using our robust knapsack formulation.**



(a) Antibody test data - 3 test combinations (b) Antigen test data - 3 test combinations

rule achieving comparable AUC than the optimal rule. Note that these results correspond to the robust optimization formulation (Γ cross-validated) with adversarial imputation of missing data.

A6.3. Antigen test data: Stratified results by viral load

The data presented in Figure 1 of Corman et al. (2021) shows that, on condition-positive samples, the test result highly depends on the underlying viral load. Here, we stratify our condition-positive population into two groups corresponding to high and low viral load patients respectively (taking the median viral load as a threshold) and report the accuracy of our combination testing policy on each sub-group separately. By restricting our attention to high-viral load patients, we are partially controlling for viral load. Hence, the remaining sources of randomness in test results are more likely to be inherent to the chemical process involved in protein detection. We thus expect the results

Table A7 Sensitivity of the 6 antigens tests computed on the high and low viral load patients

High-viral load individuals			Low-viral load individuals		
Test	N_+	Sensitivity (95% CI)	Test	N_+	Sensitivity (95% CI)
I	54	93% (85%, 95%)	I	52	50% (45%, 55%)
III	48	98% (89%, 100%)	III	57	74% (68%, 77%)
IV	48	60% (54%, 65%)	IV	57	0% (0%, 6%)
V	54	94% (87%, 96%)	V	52	58% (52%, 62%)
VI	48	75% (68%, 79%)	VI	57	0% (0%, 6%)
VII	60	87% (80%, 89%)	VII	56	54% (49%, 58%)

Table A8 Probability of pairwise agreement of the antigen test results on the high- and low-viral load patients.

High-viral load individuals							Low-viral load individuals						
	I	III	IV	V	VI	VII		I	III	IV	V	VI	VII
I	1.0	0.94	0.64	0.94	0.79	0.96	I	1.0	0.67	0.59	0.81	0.59	0.96
III		1.0	0.58	0.97	0.73	0.87	III		1.0	0.26	0.74	0.26	0.67
IV			1.0	0.61	0.81	0.62	IV			1.0	0.51	1.0	0.56
V				1.0	0.76	0.94	V				1.0	0.51	0.85
VI					1.0	0.79	VI					1.0	0.56
VII						1.0	VII						1.0

Table A9 Pairwise correlation of the antigen test results on the high- and low-viral load patients.

High-viral load individuals							Low-viral load individuals						
	I	III	IV	V	VI	VII		I	III	IV	V	VI	VII
I	1.0	-0.04	0.15	0.55	0.21	0.74	I	1.0	0.48	0.0	0.62	0.0	0.93
III		1.0	-0.12	-0.02	-0.08	-0.06	III		1.0	0.0	0.58	0.0	0.46
IV			1.0	0.0	0.61	0.3	IV			1.0	0.0	0.0	0.0
V				1.0	0.0	0.63	V				1.0	0.0	0.69
VI					1.0	0.47	VI					1.0	0.0
VII						1.0	VII						1.0

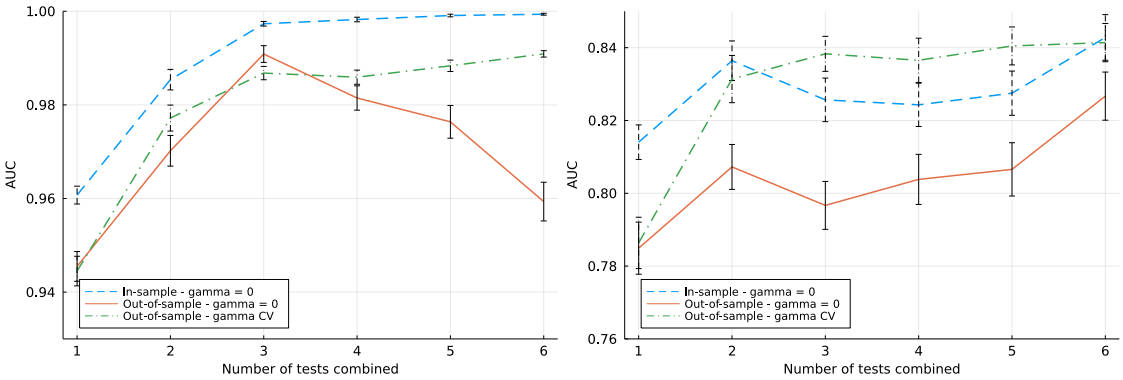
of different tests on the same individual and of the same test on different individuals to be less correlated.

Table A7 reports the sensitivity for each individual test on each stratum. As expected, high-viral load patients are easier to detect and sensitivities are noticeably higher on this sub-population. On below-median viral loads, two tests (namely IV and VI) even have a sensitivity of 0%.

In parallel, Tables A8 and A9 reports pairwise agreement probabilities and correlations of the tests on each sub-population. As mentioned above, by stratifying, we create more homogeneous sub-groups and effectively control for latent patient-level characteristics that might simultaneously impact the results of the tests. Accordingly, we expect and generally observe lower pairwise correlations than for the general population. Overall, the pairwise correlations on the high-viral load patients only (left panel of Table A9) seem lower than those on the low-viral load patients only (right panel of Table A9) – excluding tests IV and VI.

By inspecting the individual test performance and pairwise correlation of tests on these two subgroups of condition positive individuals, determining the impact of combining test results is not

Figure A7 **Diagnostic accuracy (AUC) of the best K-test combination of antigen test, as K increases, for high- and low- viral load patients respectively.**



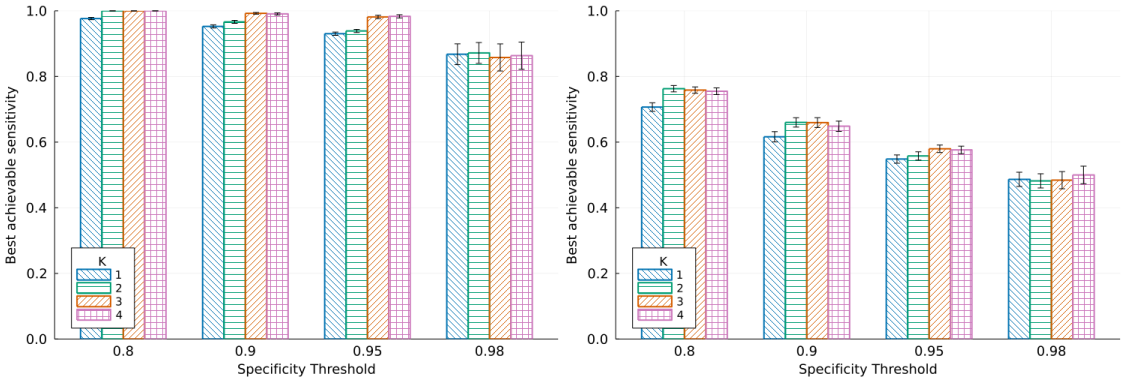
(a) High viral load (b) Low viral load

straightforward. On one hand, high-viral load patients might benefit more because the tests appear less correlated. On the other hand, sensitivity for these patients is already high and there is little room for improvement. Figure A7 replicates Figure 3, while reporting the AUC on the high (left panel) and low (right panel) viral load patients respectively, the condition-negative samples being the same in both cases. Quite surprisingly, we observe an improvement for both sub-populations and of comparable magnitude (+0.05 in AUC by combining 3 tests). Our intuition is that, on the high-viral load patients, tests are very concordant and combining helps primarly as a denoising procedure (i.e., the improvements are due to the stochastic mechanism discussed in §4.2). On the low-viral load samples, a smart combination policy can gain information by aggregating the results from different technologies (i.e., the improvements are due to the systematic mechanism, also described in §4.2).

Here, comparing AUC might be confusing because overall AUC also depends on the performance on the condition-negative population, which is the same for both panels. Accordingly, we also report in Figure A8 the improvement in sensitivity on each stratum of the condition-positive population, for different target specificity thresholds.

However, we should emphasize again that, in practice, viral load is not observable and such a stratification is impossible a priori. So, the actual performance of a test or combinations of test would be the performance we previously computed on the entire population, provided our study population is representative of the general population.

Figure A8 **Best achievable sensitivity for a K-test combination of antigen tests, for different target specificity levels, for high- and low- viral load patients respectively.**



(a) High viral load (b) Low viral load

A7. Additional numerical validation on synthetic data

In order to illustrate the behavior of our optimal combination policy as the number of observations $N = N^+ + N^-$ and the proportion of missing entries increase, we depart from our COVID-19 datasets and generate synthetic data.

A7.1. Synthetic data generation methodology

We describe here the data generation methodology we used for generating synthetic data of m correlated tests. We acknowledge the fact that our data generating process is simple and that more sophisiticated approaches could be implemented (e.g., Cario and Nelson 1997). However, we rely on this model for data generation only. Our robust combination algorithm, on the other hand, is non-parametric and can be applied to data with arbitrary correlation structure.

Given a target sample size N and prevalence level $\bar{p}$, we fix N^+ and N^-, the total number of condition-positive and -negative observations respectively, so as to match the target prevalence level $\bar{p} = N^+/(N^+ + N^-)$.

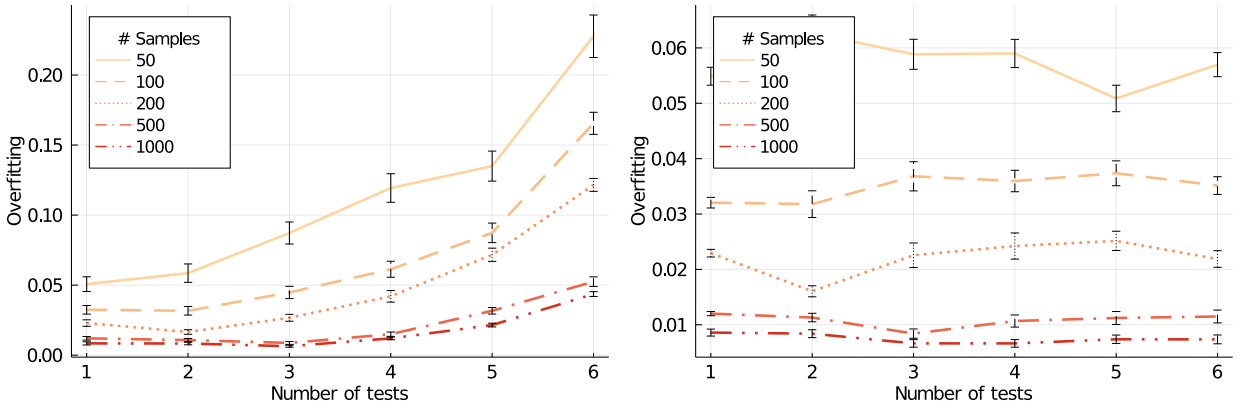
For condition-positive individuals, we consider an m-dimensional normal random variable $\boldsymbol{x} \in \mathbb{R}^m$ with mean $\boldsymbol{\mu}^+$ and covariance matrix $\boldsymbol{\Sigma}^+$. For each condition-positive observation i, we sample $\boldsymbol{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+)$ and set the result of test k equal to $\mathbb{I}(x_k^{(i)} > 0)$. Note that since the result of each test is a discrete version of $\boldsymbol{X}$, the resulting correlation matrix between test results is not exactly $\boldsymbol{\Sigma}^+$. In our numerical experiments, we consider $m = 9$ test as in the antigen test data from Corman et al. (2021). We calibrate the components of $\boldsymbol{\mu}^+$ so that $\mathbb{P}(x_j \geq 0) = p_k$, where p_k is the sensitivity for test k in Corman et al. (2021), and set $\boldsymbol{\Sigma}^+$ equal to the empirical correlation matrix between the 9 tests on the condition-positive observations in the data.

We proceed analogously for condition-negative observations.

A7.2. Impact of the sample size N

First, we evaluate the ability of our algorithm to detect relevant combinations as more data is available. To do so, we consider training data sets of increasing size, $N = N^+ + N^-$, and evaluate the resulting combination policies on the same test set, 1,000 held-out observations. Figure A9 represents overfitting, defined as the difference between in- and out-of-sample AUC, as K increases, for varying sample sizes. Due to the combinatorial nature of the problem – with K tests, there are 2^K potential outcome sequences – overfitting increases exponentially with K. On this matter, adding robustness (right vs. left panel) appears as an effective way to reduce overfitting, both in terms of dependency on K and magnitude (scale of the y-axis). Robustness is especially valuable in low-data regimes.

Figure A9 **Overfitting (relative difference between in-sample and out-of-sample AUC) as the number of tests K increases, for the non-robust (left) and robust (right) policy, on synthetic data without any missing entries, for varying sample size N. Results are averaged over 50 random training sets.**



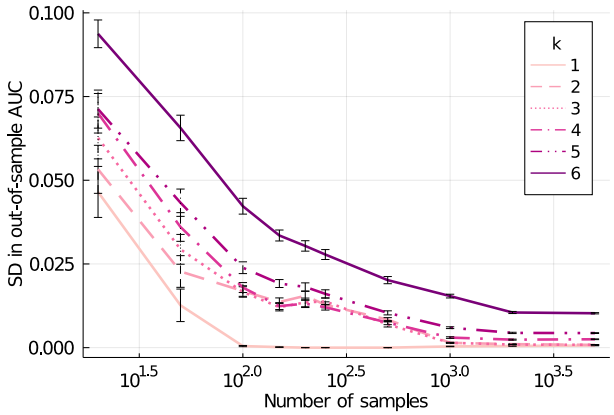
(a) $\Gamma = 0$ (b) Γ CV

Intuitively, different values of Γ lead to different aggregation rules, with varying levels of accuracy. Figure A10 displays the standard deviation of out-of-sample AUC between models with different values of Γ along the cross-validation grid, as sample size increases. A low standard deviation means that Γ has little impact and that the different values of Γ lead to very similar models as far as predictive accuracy is concerned. As depicted, as the number of samples grows, the impact of Γ on the resulting model and its performance shrinks.

A7.3. Impact of the proportion of missing entries

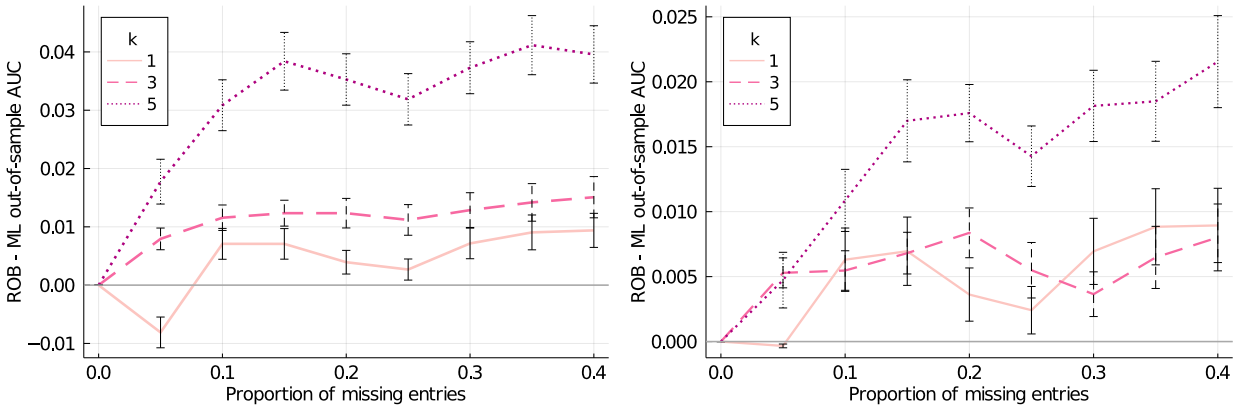
Second, we compare the relative performance of maximum likelihood vs. robust imputation, when the proportion of missing entries varies. To do so, we consider a training data set of 200 observations and randomly create missing entries, varying the proportion of missing entries between 0% and

Figure A10 **Standard deviation (SD) in out-of-sample AUC between models with $\Gamma \in \{0, \ldots, 10\}$ as the number of samples increases, for different values of K. Results are averaged over 50 random training sets.**



40%. We then compare the two imputation methods by computing the difference between the out-of-sample AUC with the robust and maximum likelihood imputation method as reported in Figure A11. A positive value indicates that the robust method improves over maximum likelihood. Figure A11 confirms the findings from Section 5.4: Robust imputation provides a significant gain in out-of-sample AUC, which increases as both the number of tests involved K and the proportion of missing entries increase. The edge of robust imputation persists when considering robust combinations of tests, although the magnitude of the improvement is smaller.

Figure A11 **Absolute difference in out-of-sample AUC between robust (ROB) and maximum likelihood (ML) imputation as the proportion of missing entries increases, for the non-robust (left) and robust (right) policy, on synthetic data with $N^+ + N^- = 200$. Results are averaged over 50 random training sets.**



(a) $\Gamma = 0$ (b) Γ CV