Drawing the line for risk stratifications:
Designing return-to-work policies that consider diagnostic error, costs, benefits and COVID-19

Roger M. Stein*

First Draft: April 4, 2020
This Draft:† April 29, 2020

*Stern School of Business, NYU. rstein@stern.nyu.edu (email).

†I am grateful to Daniel Arbess, Dina Aronson, Vasant Dhar, Ophir Falk, David Gershon, Thomas Hughes, David Katz, Alexander Lipton, Debbie Lucas, Foster Provost, Ariel Stein, Nicole Walden, and Tim Walsh for helpful comments on earlier drafts of this article. All errors are my own. Comments are welcome.

Reopening Communities Affected by COVID-19 is a Risk Management Issue

Governor Andrew Cuomo, April 15, 2020 NY (D)

"It's over when people know I'm 100% safe and I don't have to worry about this. When does that happen? When we have a vaccine...Until you have a vaccine, until you have the medical treatment, what do you do? How are you building the bridge? Well, it's going to be a phased reopening."

Governor Charlie Baker, April 22, 2020 MA (R)

"If we move too quickly, we risk moving the progress that we've made so far... We'll get through this and we will come out stronger on the other side but everyone needs to do their part and understand that we need the facts on the ground to drive our decision making."

Governor Gavin Newsom, April 15, 2020 CA (D)

"Let's not make the mistake of pulling the plug too early, as much as we all want to...I don't want to make a political decision that puts peoples' lives at risk and puts the economy at even more risk by extending the period of time before we can ultimately transition and get people moving again."

Governor Kim Reynolds, March 25, 2020 IA (R)

"While we'll look at it from a regional perspective and we'll talk about collectively the metrics we're using, each individual governor is going to look at their own state's metrics."

Governor Larry Hogan, March 25, 2020 MD (R)

"You can't put a timeframe on saving people's lives. We're going to make decisions based on the scientists and the facts."

Executive Summary

What this Paper Provides

This paper outlines an easy to implement toolbox that clinicians can use to help develop local risk policies to minimize the total harm from COVID-19.

How to Use this Paper

Use this paper to help develop and evaluate risk stratification policy options:

- **input information about a test, model or risk stratification policy rule, and your assumptions about the community needs and priorities**
- **receive back a customized risk stratification policy for that community.**

Key Points

1. Any policy to risk stratify individual involves implicit or explicit trade-offs, footnotee.g., harm to health from infection with COVID-19 vs. harm to health from economic losses regardless of how risk stratification is done.[a]

2. Policies can be designed to achieve maximum value for the trade-offs by making simple calculations, and this analysis provides a coherent, concise communication tool for and explaining policies to constituencies and the media.

3. Policies can be designed to take advantage of whatever information is available at a given time, and can incorporate different testing protocols within a single framework (e.g., virology tests where they are available; age-based criteria when they are not).

[a] e.g.,using diagnostic test, predictive model, expert judgement, or some other way.

Importantly, clinicians do not need to be mathematicians to use these techniques. The calculations may be done in a few lines of a spreadsheet or even on a hand calculator.[1] Most of the main content is delivered through real-world examples.

Templates provide guidance on how to make better decisions about reopening economies, screening patients, and other decisions and policy questions.

A **web-based tool** for exploring this analysis is publicly available at no charge at:
http://www.rogermstein.com/covid-19-resources/. (*see the Quick-reference Guide.*)

[1] Though not required, mathematical detail and technical references are provided in the Appendices.

Where to Find Techniques for Some Common Policy Questions

- How can policies be designed to minimize total harm, while also maximizing benefits, given economic and related health concerns?[a]
- What are the potential dangers of using risk stratification methods to stage the end of self-isolation?[b]
- How do these dangers change if tests and data are less or more reliable?[c]
- If different regions have different access to different tests, can these be combined within a single policy?[d]
- How can risk stratification using a single measure be made more flexible to allow for prioritization of specific community and economic objectives?[e]
- Is it worth it to spend extra money on better data or testing, and if it is, how much?[f]

---

[a] By doing a few calculations. See Example 5.
[b] It varies. See Example 3.
[c] It can make a BIG difference. See Example 4.
[d] Often, yes. See Example 8.
[e] Yes, by adjusting a few numbers for each objective. See Example 7.
[f] Sometimes. The decision can be explicitly structured however. See Example 6.

Contents

Abstract

The COVID-19 pandemic of 2019 and 2020 has brought to a standstill normal life and commerce in many parts of the world. This has left clinicians challenged to determine when and how to return workers to the workforce while also balancing critical public health issues and minimizing additional harm. A number of recent proposals have suggested using risk stratification strategies to classify members of the public into "higher" and "lower" risk tiers. This article describes practical methods for creating risk stratifications using currently available information which can be updated over time. In this context, we discuss three key observations: (1) Any policy to risk stratify individuals involves implicit or explicit trade-offs (e.g.,harm from morbidity/mortality vs. harm from economic impact on health due to self-isolation), regardless of whether risk is measured using diagnostic tests, predictive models, or expert judgement. (2) Policies can be designed to minimize the harm caused by such trade-offs and failing to do so will directly result in unnecessary harm for the population. (3) Policies can be designed to take advantage of whatever information is available at a given time, and can incorporate different testing protocols within a single framework (e.g., virology tests where they are available, and age-based criteria when they are not). We demonstrate each of these points.

We provide a set of simple *calculation templates* that clinicians can use to discuss risk stratification policies with the objective of developing *Total Harm Minimization* policies. We have also provided a *web-based* tool-set that implements these methods, and which is available publicly.

Using these approaches, for example, a policymaker can input assumptions about risk stratification needs and community priorities, and get as output a risk stratification policy that minimized health and economic harm, given the community priorities. This allows clinicians to assess the advisability of a range of important options for bringing communities back on-line, as well as for making decisions, such as whether to invest in more accurate tests or gather more detailed data.

Keywords *COVID-19; health policy; health economics; cost-benefit analysis; total harm minimization; base rate; Bayesian probability; ROC analysis*

1 Introduction

At the time of this writing, the vast majority of the U.S. population has been quarantined or is under a self-isolation or shelter-in-place order. This is the result of a set of strategies, implemented at the state level to reduce the immediate transmission of the novel novel coronavirus, and to slow its progression. In addition to public heath concerns, a key objective of these strategies is to reduce the risk of overwhelming the healthcare infrastructure of individual states and cities.

These drastic measures, while showing promise in reducing the spread of COVID-19, have also had the effect of severely depressed both local and global economies. Furthermore, a number of health advocates have observed that a prolonged period of attenuated economic and social activity has, itself, the potential to increase substantially negative health outcomes.

Recent quantitative models have suggested the merits of have advocated performing "risk stratification" in order to to classify members of the workforce and broader public into "higher" and "lower" risk tiers, to potentially allow some to return to work in advance of the introduction of an effective novel novel coronavirus vaccine. Recent recommendations (see, e.g., Katz et al., 2020; Stein et al., 2020) have also advocated this strategy.

The basic premise of such proposals is that, at the very least, many "low risk" individuals would not themselves be at severe health risk, even were they to be exposed to and contract COVID-19.[2]

These proposals assert that these "low risk" individuals should return to the workforce to re-energize the economy and reduce unnecessary harm from the negative health impacts of self-isolation. Under these proposals, those individuals who are at "high risk" for severe health outcomes if exposed to the novel coronavirus, in contrast, would remain in isolation pending the development of widespread screening programs and/or vaccines that reduce health risks to the levels of more common viral infections such as seasonal flu.

NY Governor Andrew Cuomo summed-up this concept up succinctly: "*Until you have a vaccine, until you have the medical treatment, what do you do? How are you building the bridge? Well, it's going to be a phased reopening.*"

It is inevitable that some people who are infected will end up being allowed to "cross the bridge" too early while some other healthy people will just as inevitably end up being asked to "stay back" by mistake. In the setting of a pandemic and the resulting lock-down, there are few easy decisions and **there are no attractive options** for managing the risks of infection, on the one hand, and the health and economic conse-

[2] There is often discussion as well about the risk to high risk individuals of low-risk individuals transmitting the disease while themselves showing no or mild symptoms.

quences of self-isolation and quarantine, on the other.

> ***All options currently before clinicians for managing the COVID-19 lockdown involve risks that may harm segments of the population. The charge of clinicians is to minimize this potential harm to the highest degree possible, given the risks that must be taken. Policy approaches that do this are called***
>
> ***Total Harm Minimization (THM) strategies.***

Most risk stratification approaches to policy rely implicitly on a policymaker's (PM) ability to segment individuals into different risk groups using some form of a "test," which may be a diagnostic test, a statistical analysis, a heuristic, etc. (see Section 1.1.1 for a discussion of the many forms that such tests may take). For example, a policy might consider only two risk groups, "high" and "low." The coarse test might be related age and medical status: all individuals above the age of 65 with certain comorbidities would be stratified into the "high risk" group, while those those below 65 or with no comorbidities would be stratified into the "low risk" group.

Risk-stratification recommendations, by construction, assume the widespread availability of some form of diagnostic instrument or predictive model that can be used as a test to grade the risk of individuals to classify them into risk groups (tiers). As a matter of practice, any current test criteria that clinicians must use to defining such stratifications are not perfect, in that the test may erroneously classify some COVID-19 infected individuals as "low risk," while also erroneously classifying some non-infected individuals as "high risk." [3]

1.1 Risk stratification

The techniques described in this paper are meant to be applied in cases in which a policymaker wishes to implement a **policy**: a set of criteria and a plan for allowing subsets of individuals, or specific segments of the population, to enter some venue or to return to some specific activity before an effective COVID-19 vaccination has been introduced.

Some examples are given below:

[3] For example, anecdotal evidence suggests that at the time of this writing, the COVID-19 viral tests that are used to determine whether an individual is infected or not have a false negative rate of about 15% meaning that in about 1 out of 7 cases in which an individual is infected, the test will give a negative result indicating that the patient is healthy.

Activity/Venue		*Policy Maker (PM)*
◦ End self isolation	→	National/local officials
◦ Reopen elementary schools	→	School admins; National/local officials
◦ Enter press conference	→	Political administration
◦ Enter factory	→	HR executives; National/local officials
etc.		etc.

The goal of a policy is to balance the health **risk** to the individual and public at large of contracting and spreading COVID-19, against the health and other risks to the individual and public at large of reduced economic capacity and other non-COVID-19-related morbidities that result from keeping individuals isolated unnecessarily.

In forming such policies, clinicians will consider two types of information about how widespread the COVID-19 is. For questions of capacity, epidemiologists typically consider **prevalence** which is the current number of active cases or the number of active cases as of a specific point in time. For considering contagion, the **incidence rate** describes the number of new cases that develop during a specific time.

For example, the prevalence of COVID-19 in New York on January 1, 2020 represents the total proportion of New Yorkers who were sick with COVID-19 in on January, 1, while the incidence rate of COVID-19 in New York in the month of January describes the rate at which healthy New Yorkers were contracted COVID-19 during that month. Prevalence is often useful for considering issues of capacity and health burden, while the incidence rate describes how quickly COVID-19 is spreading.

1.1.1 Tests and other forms of screening

The techniques we present are designed for cases in which a policy is anticipated to use some form of **test** to determine which individuals will be allowed to enter (do) the target venue (activity) and when. We use the term **screening** to describe the process of administering the test and then using the result as a decision criterion.

Tests may take a number of forms. Some examples include:

- A virology diagnostic
- A serology diagnostic
- Heuristics based on Age, travel, comorbidities, etc.
- Predictive analytics based on epidemiological data
- etc.

Such tests are generally **administered at the individual level** and the result of a test applies to an individual.[4]

Some tests produce **binary** outcomes while others produce results in a **continuous range** of possible outcomes. For example, the answer to a "yes/no" question such as, "Do you live alone?" is (generally) binary as are certain serology tests, while certain RNA-based virology test may return the number of amplicons or the viral load, with higher levels indicating more virus present. Another example of a test with continuous range outcomes would be the use of *Age* as a crude "test," with a different expected health outcomes associated with ages.

A **test result** may be delivered in a number of forms. Examples include:[5]

- RNA concentration level in a specimen
- The presence of IgG antibodies to SARS-CoV-2 in the blood
- Travel to a COVID-19 "hot-spot" within the past two weeks
- Absence of a fever above 98.7° F (37.0° C)
- An estimate of the probability of current infection
- etc.

In some cases, the preferred test may not be available to all individuals given limited availability or differences in test protocols at different facilities. This makes it necessary for a policymaker to integrate different types of tests into a single policy (see, Section 3.4).

There are many forms of **risk** related to COVID-19 for which a policymaker may wish to screen in order to implement some form of harm mitigation. Among the **risks** that may concern a policymaker may consider are:

- individuals who are *currently infected* with COVID-19 entering the general population;
- individuals who are *more likely to become infected* with COVID-19 becoming exposed;
- individuals who might experience *more severe prognoses were they to be infected* becoming infected with the coronavirus (e.g., older individuals or those with comorbidities);
- individuals who are *more likely to transmit* COVID-19 if infected, even if they themselves do not experience severe symptoms circulating in the general population;

[4] Note that although the results are given at the individual level, it is possible, in some cases, to administer the test itself to a group by, e.g., pooling individual specimens (see, for example, (Bilder and Tebbs, 2012)).

[5] These, and all examples in the paper are provide purely for illustrative purposes.

- individuals who do not meet the above criteria, being unnecessarily kept from working and conducting other daily activities;
- etc.

For purposes of our discussions, we generally use examples involving the first and last risks. For our examples, we primarily assume the goal of the policymaker is to isolate currently infected individuals and permit uninfected individuals to leave isolation).

However, we do this for exposition not because we suggest this to be or not to be main purpose of current policies. In fact, the techniques we discuss are quite general and can be applied to any of the risk stratification problems listed, as well as many others.[6]

In what follows, we use the term **risk** to mean the risk of that a policy decision will cause causing individuals who would otherwise have remained healthy to become infected with COVID-19. This may be refined, in some settings to contemplate cases in which the average prognosis for an exposed individual becomes worse as the result of a policy decision.

> An individual may be *at risk* for infection or at risk because if they were to become infected, their prognosis would likely be severe, perhaps resulting in significant hospitalization or death. Even if an individual were not personally likely to become severely ill if exposed to the novel coronavirus, that individual may still put others at risk if they are likely to spread COVID-19.

In what follows, we use the term **harm** to mean the consequences of policy decisions which result in reductions of health and economic security for an individual.

> An individual may be *harmed* if they suffer a reduction in the quality or duration of their health, as the result of either direct morbidity from COVID-19 or other conditions resulting from prolonged isolation, the effects of food insecurity, lack of access to medical services due to high utilization by COVID-19 patients, psychological trauma, domestic abuse, and loss of economic security.

1.1.2 Risk stratification systems, staging and test errors

For purposes of policies based on **risk stratification**, individuals may be grouped into **risk levels** (**risk tiers**) by assigned each individual or group a **risk score** based on the results of one or more tests.

[6] This is also not intended to imply that each of the challenges (concerns) we listed is equally easy to analyze. Assessing the impact of transmission by asymptomatic individuals is more complicated than screening individuals for the presence of the novel coronavirus RNA.

Each risk tier is defined by a set of **criteria** based on the results of one or more types of tests.

Importantly, it is not necessary for the inclusion criteria for a specific risk tier to be based on a single test, and not all tests must necessarily be required for inclusion in a specific tier. For example, a coarse "high risk" class might include individuals who are over 80 years old *or* have recently come into contact with an individual with a confirmed case of COVID-19 *or* who have tested positive for the novel coronavirus.

Risk tiers can be particularly useful for implementing **staging** strategies, designed to enable some segments of the population to "enter" earlier, while others do so at later points. For example, all else equal, individuals in a company that do similar jobs may be staged such that those who are deemed to be lower risk are the first to return to work, followed later by those who are deemed to be higher risk.

Risk stratification involves defining one or more **cutoffs**: threshold values for test results that determine into which of two adjacent risk tiers a result would place an individual. Returning to the earlier examples of various forms of test results, we can see that the criteria implicitly include cutoffs:

Test criterion	*"high" risk cutoff*
RNA concentration in a specimen	concentration $>x$
The presence of IgG antibodies	No antibodies found
Travel to a COVID-19 "hot-spot"	Two weeks & on "hot-spot" list
Fever	body temp $>98.8°$ F
A model estimate of a probability of infection	probability$>1\%$

Finally, for practical purposes, all tests have some level of **test error**. An error occurs when a test result disagrees with the truth. This most often occurs because most tests cannot be exactly tailored to each individual and may miss key information that determines an individual's risk, even if they do identify differences between individuals along more typical dimensions. In some cases, there may also be instances of testing error that go undetected.

By convention, we label those test results that erroneously classify a sick person as healthy, **false negatives** – high risk individuals who are erroneously stratified as "low risk" and reenter the general population. These are sometimes called **Type I errors**.

We label those results that erroneously classify a healthy person as sick **false positives** – uninfected or immune individuals who are erroneously classified high risk and thus prevented from reentering the general population and workforce. These are sometimes called **Type II errors**.

We call correct test results the **truth**, and we label a test result that correctly identifies a sick individual "positive," a **true positive**, Similarly, we label a result as a **true negative** if the test correctly identified an uninfected individual as "negative."

Collectively, we call the set of the true negative rates, true positive rates, false negatives rates, and false positive rates for a test, the test's **performance**.

It is worth noting that the terms "negative" and "positive" can sometimes cause confusion. In common usage, we think of negative outcomes as bad and positive outcomes as good. However, in our context, these terms refer to the outcome of a test for, e.g., the presence of COVID-19. In this case, if the test "comes back" negative, it means that no virus was found.

Continuing the "age test criterion" example, a policymaker may conclude that individuals above 40 years old are at higher risk than those at below 40. Indeed, if *"Age is lower than 40 years"* were used as a criterion to allow individuals to return to work, many infections might be avoided (true positive). However, in addition, many healthy individuals would be unnecessarily kept from returning to work (false positive). Furthermore, some individuals below age 40 will also contract COVID-19, despite passing through the screening (false negative), thought the many more healthy people would also correctly be able to return (true negative).

The *risk* is of infecting ones-self or others others, while the *errors* are any mis-stratification of high risk individuals into low risk tier, or the low risk individuals into the high risk tier. Figure 1 gives an example of how risk and errors related.

For clarity, we distinguish between the risk of COVID-19 infection and transmission, and the **costs** of errors that result from imperfect stratification. Costs relate to the **harm** caused by a policy with respect to test errors.

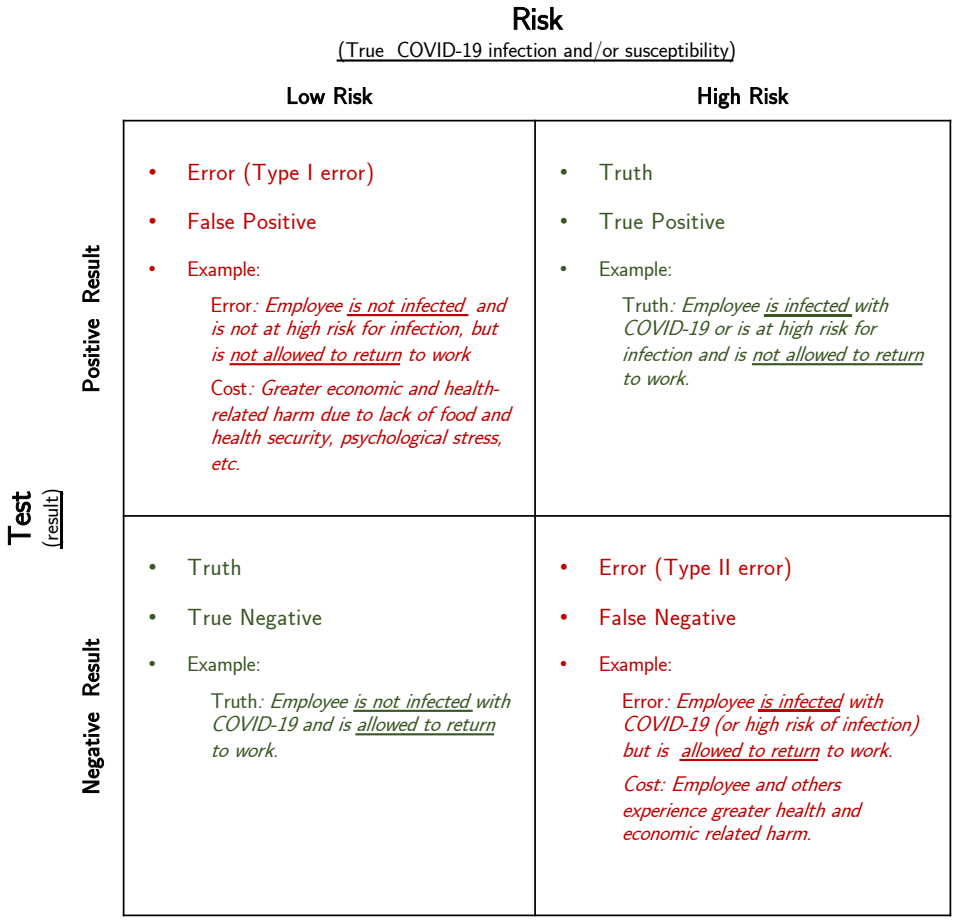


Figure 1: An example of the relationship between outcome risk (true outcome) and error (test result)

This table shows both the kinds of errors a diagnostic test may make in the context of COVID-19 screening, as well as correct (true) test results. Each cell notes whether the outcome is the truth or an error, the name of the type of error if applicable, and gives an example. An error occurs if when the test result (vertical) disagrees with the truth (horizontal). Benefits are not shown, though we discus these later on.

Returning again to the example of the "age test criterion,"

- The "errors" would be the mis-classifications (mis-stratifications) of high and low risk individuals (e.g., under-40-year-olds who actually are or become infected and gravely ill: false true negatives; and uninfected over-40-year-olds who are asked to continue self-isolating: false positives).
- The "costs" would be the harm that results from the mis-classifications (e.g., additional infections to others increasing morbidity and mortality from COVID-19; and additional health and financial problems due to extended self-isolation and economic loss, respectively).

The cells of Figure 1 give examples of the the costs of each type of error. (Also see Section 4.)

Note that later on, we will also discuss **benefits** that may be associated with correct classifications in the same way that costs are associated with errors.

1.2 Trade-offs in risk stratification: Painful questions and limited data

Implicit in any risk stratification proposal is an acceptance that there will be some number false negatives and that there will also be some number of false positives.

A common question that clinicians struggle with is therefore:

What level of health and economic of risk is acceptable?

Said differently, one may ask:

How should a policymaker determine what level of risk is "too high," given the needs of their community?

Determining such a cutoff (e.g., everyone under the age of 65 with no preexisting conditions may return to work) requires consideration of a number of factors relating to the overall risk and severity of infection:

- the costs to the healthcare system (in both financial and health terms) implied by severe infections;
- the benefits of workers returning to work;
- the costs (in both financial and health terms) of individuals remaining in isolation;
- the error rate of whatever test is used;
- the speed at which COVID-19 is spreading;
- etc.

These factors are often hard to assess, and the data available to clinicians is often limited, noisy, inconsistent, and anecdotal. However, regardless of whether a policy is explicit in how it trades off these factors or is silent, and regardless of whether the policymaker introduces them them deliberately or incidentally, these trade-offs are present and can often be inferred by observers simply by using public data.

Policy makers evaluate such trade-offs routinely.

Example 1. Routine trade-offs in policy

The U.S. FDA approves a drug's safety based on its analysts judgement that the adverse effects of the drug are outweighed by the benefits to the patients that respond well (as supported by data from clinical trials).

In general, this is the case. However, many patients are still harmed by these drugs. The The Agency For Healthcare Research and Quality, which is part of the US HHS, estimates that in 2017, there were over *720,000 adverse drug events while patients were under close medical supervision within a hospital* (AHQR, 2019). This does not include the adverse drug reactions that non-hospitalized individuals experienced. In 2017, adverse drug reactions accounted for about 3% of all visits to hospital emergency departments (Rui and Kang, 2018).

The experience of clinicians with drug approvals is not intended as a normative observation.[7] Rather, it is serves to highlight that much of policy making involves risk management, that risk management involves trade-offs, and thus trade-offs are inevitable in policy making.

Despite this, a policymaker's constituents may have little transparency into or appreciation of what these trade offs involve, or how they are determined.

[7] To the contrary: The Agency for Healthcare Research and Quality has instituted a program to aggressively reduce ADRs through a number of initiatives, including more transparent reporting and data collection. The incidence of ADRs declined by 25% between, 2010 and 2016.

The techniques that follow can also provide a language for communicating to constituents, rather than leaving the interpretation of policy priorities to community members,policymaker who may have less context than the policymaker.

1.3 An approach to making flexible, rational decisions with incomplete data and tools

clinicians are charged with responding quickly in crisis situations. Their charge requires them to use of whatever data and tests are available, regardless of the quality. In this context, it is logical that a policymaker should use these data and tests to minimize the total harm caused by imperfect information when determining how best to make required trade-offs.

A policymaker should seek to minimize the total harm the the highest degree possible for the incidental risk that accompanies the use of imperfect test, criteria and data.

Example 2. How not to choose an irrational policy

Imagine that a policymaker is asked to determine a policy for determining which individuals will be able to return immediately to their jobs and which individuals will remain in isolation. The definition of such a policy involves a trade-off between the benefits to workers of workers returning to the workforce, on the one hand, and the costs of the harm harm that will result from some of those workers becoming infected or infecting coworkers, on the other.

The policymaker is considering three policy proposals. Each policy will return some workers to the workforce, but will also l also likely create additional harm by causing an increase in COVID-19 cases in the community, because to some of the infected workers will not be screened out due to test error.

For simplicity at this point, we assume that the only information that the policymaker has to inform this decision relates to the number of cases of COVID-19 that each policy is likely to produce due to test error and the number of workers that will be permitted to return to work.

	Expected # employees back to work	Expected # additional cases of COVID-19
Plan A	15,000	500
Plan B	10,000	500
Plan C	15,000	400

The three proposals listed above are the policy choices plans available to the PM.

Under any state of the world, the PM should prefer Plan A to Plan B (assuming he wishes to maximize the number of employees returning to the workforce and minimize the number of additional COVID-19 infections).

He would deliver more benefit with Plan A, for the same cost of additional harm. (More benefit, same cost.)

By similar reasoning, the PM should prefer Plan C *even more* than Plan A: Plan C gives the same benefit as Plan A, but with lower cost. (Same benefit, less cost.)

In this case, these preferences will hold regardless of how he values the harm of additional COVID-19 cases relative to the benefit of additional employees returning to work.[8]

If these were the only factors in his decision, it would be irrational to choose any plan other than Plan C. ♢

In what follows, we discuss issues related to choosing policies when errors in risk stratification are to be expected due to imperfect information and imperfect tests.

We examine

- the impact of the error rate of a diagnostic instrument on the real-world population of COVID-19 susceptible individuals; and
- the potential impacts on policy evaluation that data and risk stratification errors may precipitate; and
- a set of methods for evaluating potential policies that are designed to minimize the total harm to the population, given the economic, social and health costs and benefits.

The specific costs of different types of errors (infected employees returning to work or healthy employees being asked not to), and benefits of true test results (economic security, access to services, social well being; mitigating the spread of COVID-19) will vary greatly across communities and domiciles. We are not qualified to opine on these quantities, and do not address their determination in this paper.

[8] This ignores in the improbable case that he views the harm as irrelevant.

However, the leadership of a community (country, state, city, firm) *is* well positioned to opine on these topics. Assuming that leaders are able to determine, even in broad terms, these factors and costs, the techniques we describe can provide them with a mechanism for choosing the "best" risk tiers and cutoffs in order to minimize total harm.

To be useful, these methods must be accessible. To make this paper accessible, most of the main body of each section is given over to examples, rather than detailed mathematical derivations. (However, for those interested in the technical details, more in-depth mathematical results are available in the Appendices.)

This paper is not intended be theoretical. It is meant to be used by non-mathematicians. To that end, most of the main results may be calculated on a hand calculator or in a few cells of a spreadsheet.

In many cases the results may also be calculated online at no charge at: http://www.rogermstein.com/covid-19-resources/

The remainder of this paper is organized as follows:

The Organization of the Paper

- **Section 2** (the next section) discusses the policy implications of model and test error and of sparse or unreliable data. Some of these implications are surprising to readers not familiar with probability theory, but are nonetheless directly relevant to the issues that can arise, even when a test has a high reported accuracy, as well as some potential solutions.
- **Section 3** describes how a policymaker can use even imperfect tests and data to determine risk stratification policies, given a set of priorities, costs, and benefits. This section is divided into a number of sub-sections, that demonstrate:
  - how to coherently incorporate different decisions and priorities into a policy;
  - how observers can infer a policy's trade-offs by examining the policy and doing a few calculations;
  - how clinicians can assess whether more expensive or extensive tests or data collection efforts are justified and, if so, what their impact could be;
  - how to create flexible risk stratification policies to adjust for different priorities and community objectives
  - how to tailor such policies to different groups by considering things like the criticality of an individual's job function or the risk of harm that their return to the workforce may pose to the community or to themselves.
  - how to incorporate tests given to different individuals, using different testing tools, can be coherently incorporated into a single consistent policy.
- **Section 4** discusses some implications of the results, and also discusses:
  - some of the challenges in developing local cost and benefit assessments;
  - how more advanced, flexible and efficient risk stratification policies may be constructed;
- For those who are interested, a number of technical Appendices provide mathematical details that are used in the main paper, including the main formulae.

2 Risk policy in an imperfect world: Bad options + Rational Analysis = Total Harm Minimization

Consider a test that produces a score: $\mathcal{HIGH}$ or $\mathcal{LOW}$, indicating that an individual is either in a "high" or "low" risk tier, respectively, for serious COVID-19 morbidity. This test could be a medical diagnostic, a simple rule (e.g., everyone over 65 is at high risk), the prediction of a machine learning algorithm, etc.

If the test were perfect, then given a risk score of $\mathcal{HIGH}$, the probability of actually belonging to the high risk tier[9] would be 1.0. Similarly, for a risk score of $\mathcal{LOW}$, the probability of belonging to the low risk tier[10] would be 1.0 as well. In this idyllic setting, there would be no model error and *every* high risk individual tested would be stratified into $\mathcal{HIGH}$ category and *every* low risk individual tested would be stratified into $\mathcal{LOW}$.

However, if the risk scoring system were not perfect, then the probability of being at high risk, given a score of $\mathcal{HIGH}$, would not be 1.0 and/or the probability of being low risk, given a score of $\mathcal{LOW}$, would not be 1.0. Furthermore, for an imperfect scoring system, the probability of getting a score of $\mathcal{HIGH}$ if the disease were actually present may also not be 1.0, so some results would not correctly stratify individuals into their appropriate risk tier.

The probability of getting a *correct* test result of $\mathcal{HIGH}$, when the subject is indeed at high risk is the **true positive rate**. The probability of getting an *incorrect* test result of $\mathcal{HIGH}$, when the subject is really at low risk, is the **false positive rate** (or the **Type I error rate**). We can define the *true negative* rate and **false negative rate** (or the **Type II error rate**) in the same way. In what follows, we label these $\mathcal{TP}$, $\mathcal{FP}$, $\mathcal{TN}$ and $\mathcal{FN}$, respectively.[11] The next example shows how these rates affect risk stratification.

[9] i.e., $p(\mathbb{H} \mid \mathcal{HIGH})$.

[10] i.e., $p(\mathbb{L} \mid \mathcal{LOW})$.

[11] Note also that it is not necessary to explicitly specify all of the accuracy and error rates, because of the following identities:

$$\mathcal{TP} = 1 - \mathcal{FN} \quad (1)$$
$$\mathcal{TN} = 1 - \mathcal{FP} \quad (2)$$
$$\mathcal{FP} = 1 - \mathcal{TN} \quad (3)$$
$$\mathcal{FN} = 1 - \mathcal{TP} \quad (4)$$

Example 3. How many infected reporters get into a press briefing? Imagine that a policymaker were worried about infected reporters spreading the novel coronavirus at a press conference, but she also worried about restricting reporters' access to the press conference.

The policymaker might consider screening each reporter as they enter the briefing room. In this case, we would be concerned about infected individuals entering the briefing room and possibly infecting others.

Using either Equation (B.1) or Equation (C.8), we can calculate the expected number of infected reporters who will enter the briefing room after being screened.

Assume for purposes of this example that

- the test is 97% effective at detecting COVID-19 when an individual is infected;
- the test is 90% effective at detecting the absence of an infection when an individual is not infected; and
- the prevalence of active infections is 20% at the time of the briefing.

If the press corps for the briefing includes 300 reporters then we have:

$$\begin{aligned} n_T &= 300 && \textit{(the number screened)} \\ \mathcal{TP} &= 0.97 && \textit{(true positive rate)} \\ \mathcal{TN} &= 0.90 && \textit{(true negative rate)} \\ \mathcal{FP} &= 0.10 && \textit{(false positive rate)} \\ \mathcal{FN} &= 0.03 && \textit{(false negative rate)} \\ p(\texttt{COVID-19}) &= 0.20 && \textit{(prevalence of COVID-19)} \end{aligned}$$

With this information, we can calculate the expected outcomes of the testing using only counting and basic arithmetic:

$$\begin{aligned} \boldsymbol{n_{\mathbb{C}}} &= \mathbf{60} && \textit{(\# with COVID-19)} && \equiv p(\texttt{COVID-19}) \times n_T && = 0.2 \times 300 \\ \boldsymbol{n_{\mathbb{W}}} &= \mathbf{240} && \textit{(\# who are well)} && = n_T - n_{\mathbb{C}} && = 300 - 60 \\ \boldsymbol{n_{\mathcal{N}}} &= \mathbf{218} && \textit{(\# negative results)} && = \mathcal{FN}(n_{\mathbb{C}}) + \mathcal{TN}(n_{\mathbb{W}}) && = 0.03(60) + 0.9(240) \\ \boldsymbol{n_{\mathcal{P}}} &= \mathbf{82} && \textit{(\# positive results)} && = n_T - n_{\mathcal{N}} && = 300 - 218 \\ \\ \boldsymbol{n_{\mathbb{W}\mathcal{N}}} &= \mathbf{216} && \textit{(\# true negatives)} && = \mathcal{TN} \times n_{\mathcal{N}} && = 0.90 \times 240 \\ \boldsymbol{n_{\mathbb{C}\mathcal{P}}} &\approx \mathbf{58} && \textit{(\# true positives)} && \approx \mathcal{TP} \times n_{\mathbb{C}} && = 0.97 \times 60 \\ \boldsymbol{n_{\mathbb{C}\mathcal{N}}} &\approx \mathbf{2} && \textit{(\# false negatives)} && \approx n_{\mathbb{C}} - n_{\mathbb{C}\mathcal{P}} && \approx 60 - 58 \\ \boldsymbol{n_{\mathbb{W}\mathcal{P}}} &\approx \mathbf{24} && \textit{(\# false positives)} && \approx n_{\mathcal{N}} - n_{\mathbb{C}\mathcal{P}} && \approx 240 - 216. \end{aligned}$$

Thus, a total of 240 reporters would test negative and be admitted to the press briefing ($n_{\mathcal{N}}$). Within this group, about 2 reporters would be infected with novel coronavirus and presumably contagious (false negatives, ($n_{\mathbb{C}\mathcal{P}}$). In addition, 24 reporters, who should have been permitted to participate, would be excluded (false positives, $n_{\mathbb{W}\mathcal{P}}$).

To rely on this test, given the conditions at the time, the administration would need to be comfortable trading off the "cost" both of allowing 2 infected reporters to circulate among the press corps and leadership, while also excluding 24 reporters who should have been allowed to participate, in exchange for the benefit of preventing another 58 infected reporters from participating.[12] ♢

Application Template - Estimating the number of infected people erroneously cleared using a specific test for screening

All calculations may be done on a hand calculator or simple spreadsheet. Example 3 can be used as a template.

Input: Test accuracy
prevalence of COVID-19

Result: Expected number of infected individuals missed
Expected number of uninfected denied

2.1 Model error, data error, and risky risk-stratification policy

Up until this point, we have assumed that the population prevalence was known. However, it could be the case that prevalence rate is different from the estimate that the administration used for screening. This would be the case if, for example, the incidence rate were estimated using only patients that presented at hospitals and requested a test, but who were only tested if they were symptomatic.

In the case of COVID-19, it has been suggested that asymptomatic patients may also be contagious. However, due to shortages in testing, these asymptomatic patients would not have been unlikely to present at hospitals for testing, and their cases would not be counted in calculating the prevalence, so the population estimate of the prevalence of COVID-19 would be understated.[13]

[12] It could be the case, however, that clinicians are more concerned about COVID-19 cases that will become severe and require hospitalization than about COVID-19 cases that are likely to be mild. In this case, we may apply Equation (B.2 to combine the probability of contracting COVID-19 with the probability of requiring hospitalization given contracting COVID-19, and proceed as before, but now including the error rates for the model (or other stratification approach) used to stratify the risk that a COVID-19 patient will require hospitalization.

[13] This would also affect probability estimates from statistical models, such as those developed by applying machine learning algorithms to data collected on *recorded* COVID0-19 cases or on patients hospitalized for COVID-19. The next example demonstrates this.

2.2 What does a positive test result really mean?

To motivate this next section, imagine that there is a condition that has an incidence rate of zero among some subset of the population. For our purposes, we consider prostate cancer as the condition, and female individuals as the subset.[14]

Imagine further that there is a diagnostic blood test for this condition that is very, very accurate at detecting the disease, i.e., has a very low false positive rate (e.g., a Type II error rate of 0.01% or less than one false positive per 10,000 tests. Thus, the probability a subject with prostate cancer of receiving a positive result is 1-0.01% = 99.99%).

If the test were given to a female individual, even considering the very low Type II error, there is still a small chance that this particular test will return a result that is positive positive.

In this case, even the subject received a positive result, and even though the tests is 99.99% "accurate," we would still know that the probability that the subject in fact had prostate cancer, would be zero. In fact, regardless of how many female individuals tested positive, we would *always* know that their probability of having prostate cancer is zero because the overall population prevalence among the female population for that condition is zero. Female individuals do not have prostates.

If instead there were another condition that were slightly more common among the female population (e.g. the condition was very rare, but still occurred occasionally, say 1 case per 1000 individuals), a positive test result for a female individual would still be surprising, and although we would no longer think that the probability of the individual having this condition were zero, we might suspect that it was still very low.

Finally, if the incidence rate across all sexes were equally likely, and a very accurate test returned a positive result, we might become more concerned.

Note that in each of the preceding cases, the error rate of the test remained the same. However, our intuition about what a positive result implied changed based on our sense of the probability of an individual *ever* having the disease. The less likely this were, the less likely we were to accept a negative result as truth.

On the other hand, imagine that we were considering a test result for a very common disease, but that the test itself had a very high Type II error ($\mathcal{FP}$ is large) and frequently produced many alarms. In this case, again, we might be cautious in interpreting a positive test result, given the high levels of test error.

[14] Some of the particulars of hypothetical narrative, not necessarily accord with the pathology of prostate cancer. We ignore this for exposition.

The observation that the probability of an individual acquiring a disease impacts the interpretation of test results error rates explains why it is so challenging for clinicians to *develop policies* for public health without adequate prevalence data, on the one hand, or to *implement policies* without accurate tests, on the other.

2.3 Adjusting test results as better data comes in

The actual probability, given a positive test result, that a patient has contracted a disease like COVID-19 is a simple form of risk stratification.

We argued informally in the last section that even with a positive test result, the reliability of the stratification is still a function both the error of the test (the and the prevalence (or incidence rate, depending on context) of COVID-19.

However, it is generally understood that, for a variety of reasons (including shortages in test kits and selection bias in testing), data on the true prevalence of COVID-19 is of poor quality and incomplete, likely understating the true prevalence of the virus among the population.[15] In cases where the data used to estimate the prevalence of a disease is of poor quality, this uncertainty may critically affect the effectiveness of risk stratification.

We can make our earlier intuition more formal (see, Appendix C). Doing so results in a simple adjustment that we can make to a test result, as new data comes in about infection rates, etc.

The probability transformation converts a "bad" estimate of the probability of having COVID-19 given a positive test result (such as an estimate that was made using a "bad" estimate of the COVID-19 prevalence) into a more accurate real-world estimate of having COVID-19. This allows clinicians to begin risk stratification using the best estimate available of the real-world prevalence, but to continually update their policy guidelines in a consistent fashion as new information comes in.[16]

[15] A recent study reported by Andrew Cuomo, the Governor of New York, supports this understanding. The study randomly tested 3,000 New Yorkers across the state, and found that 13.9% of them had signs of the virus, which is about ten times higher than prior estimates, which were based primarily on patients presenting with symptoms (LaVito et al., 2020).

[16] While this specific adjustment works in some settings, it is not appropriate in many others. See important caveats discussed in Appendix C.

Example 4. How many infected reporters really get into the press briefing?

Using Equation (C.2) (or Equation Equation (C.8)), we return to Example 3, which involved using a test with known accuracy to screen reporters for COVID-19 prior to allowing them to enter an official press briefing.

Recall that the hypothetical the press corps for the briefing includes 300 reporters and the test accuracy is:

$$\begin{aligned} n_T &= 300 \\ \mathcal{TP} &= 0.97 \\ \mathcal{TN} &= 0.90 \\ \mathcal{FP} &= 0.10 \\ \mathcal{FN} &= 0.03. \end{aligned}$$

If the true prevalence were →	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**
	(A)	(B)	(C)	(D)	(E)	(F)	(G)
Individuals allowed in/barred							
Total admitted	217.8	191.7	165.6	139.5	113.4	87.3	61.2
Total barred	82.2	108.3	134.4	160.5	186.6	212.7	238.8
Wrongly barred	24.0	21.0	18.0	15.0	12.0	9.0	6.0
Contagious people allowed in	***1.8***	***2.7***	***3.6***	***4.5***	***5.4***	***6.3***	***7.2***
True risk after screening							
Chance of COVID-19 if admitted	**0.8%**	**1.4%**	**2.2%**	**3.2%**	**4.8%**	**7.2%**	**11.8%**
Chance of no COVID-19 if barred	11.0%	11.0%	10.9%	10.8%	10.6%	10.3%	9.8%

Table 1: The impact of test error varies greatly with the assumptions/data on the true prevalence

This table compares various measures of risk in Example 4, given different COVID-19 prevalence. For this example, we assume that 300 reporters are seeking entry into an official press briefing. In order to enter the press room, a reporter must be screened using a test with the following performance: of $\mathcal{TP} = 0.97, \mathcal{TN} = 0.9$. However, we assume that the observed prevalence is unknown, due to a lack of testing kits.

This time, however, we assume that the policymaker believes that the data that has been collected on the prevalence of COVID-19 is currently unreliable, due to under-sampling. She she wishes to understand how her policies would will be impacted if the prevalence is higher than the current assumptions.

As before, the administration screens each reporter and, based on the test result, admits or denies access to the reporter based on the outcome of the testing. However, now we examine the realized outcomes, given different levels of the unknown prevalence rate (which actually still remains uncertain as of the date of this writing).

The results of this sensitivity analysis are shown in Table 1.

From the table, it is clear that the increase in prevalence leads to an increase in the number of individuals erroneously admitted to the press conference. For example, recall from Example 3 that screening with the diagnostic test would result in about two infected reporters being admitted into the press room if the real prevalence were 20%. This is roughly equivalent to one infected reporter out of every 121.

In contrast, that would rise to about 7 infected reporters erroneously admitted if the prevalence were 80% (column (G)). Because fewer non-infected reporters are admitted when the prevalence is higher, this would be roughly equivalent to one infected reporter every 9 seats. Furthermore, the realized error rate increases at an increasing rate in the prevalence.

Similar behavior occurs for non-infected population, but inversely. When the prevalence is 20%, 24 out of the 82 reporters denied access should not have been, or about 1 out of every 3.5 or so barred reporters (column (A)). This compares to only about 6 out of about 239, or only about 1 out of every 40 reporters barred wrongly (column (G)) when prevalence is 80%.

Figure 2 shows these relationships graphically. ♢

In evaluating Table 1 and Figure 2 from Example 4, we observed that given a fixed test error rate, the number of infected individuals admitted is proportional to the real prevalence in the population, while the number of individuals wrongly barred is inversely proportional.

There are two ways that we can think about the results in Example 4. On the one hand, Table 1 emphasizes how important accurate data on prevalence is for making sound decisions. In the example, using the same test, say the real prevalence were 50% rather than the 20% we assumed in the first example. In that case, rather than

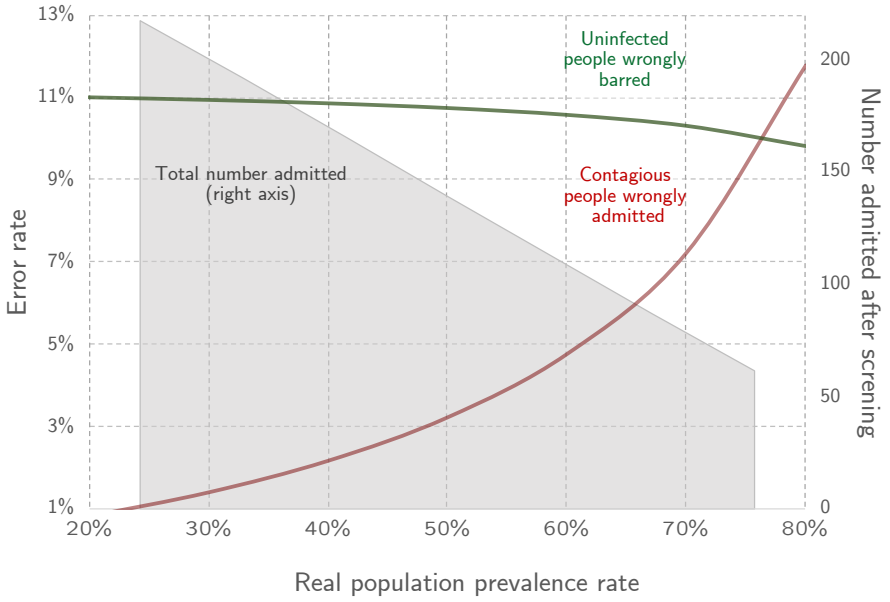


Figure 2: The impact of test error varies greatly with the assumptions/data on the true prevalence

This figure shows the impact of differing prevalences (x-axis) on the percentage of infected individuals allowed to enter and the percentage of uninfected individuals wrongly prevented from entering (y-axis), using a test with fixed error rate. For this example, we assume that 300 reporters are seeking entry into a press briefing. In order to enter the press room, a reporter must be screened using a test performance of $\mathcal{TP} = 0.97, \mathcal{TN} = 0.9$. However, we assume that the observed prevalence is unknown, due to a lack of testing kits. (See: Table 1).

erroneously admitting one infected individual for every 121 or so reporters, the policy would allow one infected individual in for every 30.

Another way to think about these results is to use these insights, and the recognition that information on prevalence will be updated from time to time as data collection and testing become more common. In this case, test results for individuals, or test thresholds for communities or sub-populations, could be updated dynamically to reflect the most current information available.

Application Template - Updating individual/group risk or updating policy thresholds based on new or localized data

All calculations may be done on a hand calculator or simple spreadsheet. Example 4 can be used as a template.

Example applications:

- Updating individuals' test results as new prevalence (or incidence) information becomes available in real-time.
- Use their policymaker's specific knowledge or assumptions about local prevalence to interpret general test results for policy purposes.
- Adjust risk cutoffs dynamically for communities or sub-populations as new data arrives.

Input: Test result (described as a probability)
Initial estimate of COVID-19 prevalence
(or incidence rate, depending on application)
Updated (or local) COVID-19 prevalence or incidence rate

Result: Updated estimate of probability of having the condition

3 Setting policy cutoffs to minimize total harm

Although clinicians can update their assumptions about or information on prevalence or incidence rates, they have little control over the actual levels. This is also true of the accuracy of a given test.

However, clinicians do have control over how tests are applied and test results are interpreted for policy purposes.

For example, a very basic risk stratification approach might be to simply use a person's age as an indicator of risk. Age is continuous, and there is no natural point at which crossing over a given age threshold makes someone suddenly much older (or riskier) than they were the day before.[17]

In such settings, the policymaker may choose a *cutoff score*, above which a test subject would be considered to be "high risk" and below which she would be considered "low risk."

For a test with given performance, it is this cutoff that determines the false negative and false positive error rates, $\mathcal{FN}$ and $\mathcal{FP}$.

As another example, consider that if *all* scores were deemed by the policymaker to be "high risk" (the "EVERYONE-IS-HIGH-RISK" policy) there could be no false negatives ($\mathcal{FN} = 0$), since there would be *no* negatives. Similarly if all scores were considered by the policymaker to be "low risk" (the EVERYONE-IS-LOW-RISK policy), there could be no false positives ($\mathcal{FP} = 0$), since there would be no positives, and everyone would pass the screening.

Were a policymaker to believe that even a single false negative (an infected individual admitted into the public) were untenable, he would be forced to adopt the EVERYONE-IS-HIGH-RISK policy. Similarly, were the policymaker to believe that even a single unfairly barred individual would be unacceptable, he be forced to choose the EVERYONE-IS-LOW-RISK policy.

In general, there are costs to mistakes, both false positives and false negatives.

- Most clinicians would reject the EVERYONE-IS-HIGH-RISK policy, due to the unacceptably high cost to the health and economics of uninfected people forced to stay home.
- Most clinicians would similarly reject the EVERYONE-IS-LOW-RISK policy because the cost to public health of widespread COVID-19 infections would be too

[17] The current discussion of age buckets such as "65 and over" is more reflective of the manner in which data is collected and reported, than of some natural measure of overall health or susceptibility.

high, both in terms of lives lost (especially given finite health system capacity) and in terms of the social and economic disruption that would result.

This means that most policies will not be all-or-nothing strategies, and will necessarily involve a trade-off between

(a) **the benefits** of allowing some lower risk individuals to return to the workforce and general public, while preventing some higher risk ones from doing so, and

(b) **the costs** of erroneously allowing some infected individuals to enter the workforce and general public, while also erroneously preventing some uninfected members from entering.

We generalize this notion in what we call:

The Law of Risk-Based Decision Making

Any policy based on less than perfect risk stratification, will involve trading off competing objectives involving harm in one form or another.

Although this "law" seems obvious, a surprising number of policies avoid any discussion of these trade-offs explicitly. In the particular case of COVID-19, the Law of Risk-Based Decision Making implies:

Until an effective COVID-19 vaccine is available widely, any policy for returning workers to the workforce that is based on risk stratification must involve trading off the cost of infecting more individuals, some of whom will die, against the costs of lost economic activity, food insecurity, and health insecurity, and unemployment.

This does not imply that policy cannot proceed unless models are perfect. Rather, it implies that clinicians can explicitly develop policies that cause the least harm by determining how best to stratify risk, and how to subsequently select cutoffs for risk tiers. Given the imperfect nature of the information and tests that form the basis of policy decisions, if a PM does not choose a policy that minimizes harm, conversely, any alternative policy is causing unnecessary harm.

3.1 Setting a policy cutoff

In its most fundamental form, harm "minimization" involves setting cutoffs for defining the risk tiers that achieve the best possible trade-offs, given the current error rates of

diagnostic tools and models and data available, and the expected costs and benefits of errors and good risk stratifications, respectively.

We will assume that there is a set of benefits and costs that a policymaker can calculate, at least approximately. The costs represent the real-world harm that errors in assessing an individual's risk incorrectly, while the benefits represent the positive outcomes associated with correct assessments.

Imagine that a policy there is a test that produces a score for each individual ranging from s_1 ("lowest risk") through s_K ("highest risk"). This score could be the propensity score of a machine learning algorithm or the concentration level of SARS-CoV-2 RNA (cp/μL) detected using an assay or a medical authority's assessment of at what age morbidity increases significantly for those exposed to the novel coronavirus, etc.

The abstract goal of the policymaker is to determine which score cutoff to use for stratifying the risk of the individuals into "high risk" (scores worse than the cutoff) and "low risk" (scores equal to or better than than the cutoff). Because the test is not perfect, any cutoff selected other than will result in some errors. If a policymaker chooses the k^{th} score as the cutoff, then the corresponding performance measures would be $\mathcal{TN}_k, \mathcal{TP}_k, \mathcal{FN}_k$ and $\mathcal{FP}_k$.

If the policymaker has an estimate of the prevalence, then, given this information (performance at the cutoff, costs, benefits and prevalence), we can calculate the total value, V_k, of a policy that uses this test and sets the cutoff at score s_k. (see: Appendix A).

It turns out that V_k is simply the weighted net value of two trade-offs: a correct vs. an incorrect classification of a test result as low risk; and the weighted value of a correct vs. incorrect classification of an individual as high risk. The weighting is done based on the prevalence and the probability of being correct or incorrect, which depends on the model accuracy ($\mathcal{TP}_k$, etc.).

The goal of the policymaker can now be seen more concretely as to maximize the value of the policy, which will minimize unnecessary harm. This is done by selecting a cutoff that results in the largest value of V_k.

In practice, there are a number of approaches for maximizing V_k.[18]

For many real world problems involving risk stratification objectives that may have a variety of attributes, it is sometimes convenient to simply calculate $\mathcal{TP}_k, \mathcal{FP}_k$, etc. for a range of k values, and to then use the cutoff that results in the maximum value of V_k.[19]

[18] See, Stein (2005) for details.

[19] The Risk Stratification Workbench app does this type of calculation automatically. See Appendix E

The following example shows an application of this approach.

Example 5. Which screening threshold to use for entry to the press briefing?

Imagine now that a new COVID-19 test is developed and has the performance shown in Table 2, and that the administration now wishes to use this new test to determine which reporters are permitted to enter a press briefing.

Risk level	s_κ	$\mathcal{TP}$	$\mathcal{FP}$
Lowest Risk	R1	99%	62%
·	R2	99%	49%
·	R3	98%	20%
·	R4	96%	14%
·	R5	91%	10%
·	R6	85%	8%
Highest Risk	R7	65%	4%

Table 2: Performance of hypothetical COVID-19 diagnostic tool

This table shows the hypothetical performance of a new test (e.g, an assay, a data-driven predictive model, etc.) for screening individuals in order to stratify them into risk tiers, based on the likelihood that they are infected. $\mathcal{TP}, \mathcal{FP}$ are true positive (correct high risk) and false positive (low risk that is erroneously stratified as high risk) rates, respectively. s_κ is the candidate cutoff.

The test results take the form of one of seven risk scores R1 through R7 with R1 being the lowest risk and R7 being the highest risk. The last two columns of the table show the accuracy of the test at identifying infected and non-infected individuals, $\mathcal{TP}$ and $\mathcal{FP}$, respectively, if the individuals were split into two risk tiers using that risk score.

This information can also be presented in the form of an **ROC curve**, as shown in Figure 3. An ROC (receiver operator characteristic) curve is an analytic tool for evaluating diagnostic and decision tools (e.g., Green and Sweats, 1966; Provost and Fawcett, 2001). An ROC curve plots a diagnostic tool's Type II error (false positive rate) on the x-axis error against one minus the Type I error (1-$\mathcal{FN}$= $\mathcal{TP}$, the true positive rate). In the case of novel coronavirus detection, an ROC curve describes the percentage of non-infected individuals tested that will be inadvertently stratified as high risk, given each possible cutoff k, in order to correctly screen out a specific percentage of infected individuals when using a specific test.

In general, the better a test is at differentiating between infected and virus-free individuals, the farther up and to the left the curve will be placed. If one test's ROC is above that of another test's at every point, then there are no situations in which using the diagnostic with the lower ROC will result in better decisions than using the diagnostic with the higher one ROC. The 45° line represents the ROC curve for a random decision tool. since the tool is effectively not providing any information (to eliminate y% of the COVID-19 cases, one must also eliminate y of the non-COVID-19

cases; i.e., using the random test, the only way to block y of the infected individuals is to just eliminate y of *all* individuals.

We have annotated Figure 3, to identify where on the ROC curve for the new COVID-19 test, each of the risk scores falls. For example, we can see from the figure that setting a cutoff at risk score R7 would screen out 65% of the infected individuals tested, but at the cost of erroneously mis-classifying 4% of the virus-free individuals, and calling them high risk.

In order to choose a cutoff for admitting or barring individuals, the policymaker must determine how best to trade off the effects of the errors. For example, if the policymaker believed that it was much, much worse to have an infected individual enter the press conference than to wrongly deny a virus-free reporter access, she might choose to use risk class R1 as a cutoff, since doing so would ensure that almost all infected individuals would be screened out, albeit at the cost of unfairly denying more than 60% of all reporters access.

In order to determine the "best" cutoff the policymaker explicitly assigns costs and benefits to the errors and correct stratifications, respectively:

$$
\begin{aligned}
c_{\mathcal{FN}} &= 20 \equiv \text{cost of false negative (infected allowed in)} \\
b_{\mathcal{TP}} &= 0 \equiv \text{benefit of true positive (correctly stopped infected)} \\
c_{\mathcal{FP}} &= 4 \equiv \text{cost of false positive (non-infected is kept out)} \\
b_{\mathcal{TN}} &= 3 \equiv \text{benefit of true negative (non-infected allowed in)} \\
p_{\mathbb{C}} &= 20\% \equiv \text{probability of COVID-19 (prevalence)\%}
\end{aligned}
$$

Using this set of assumptions, the policymaker can now apply Equation (A.1) to the compute the value of of a policy that would use that each. Table 3 shows the results of this analysis. In the table, V_k represents the value of a policy that allows all individuals with risk scores equal to or better (lower risk) than the corresponding score to return to work, while barring all of those with worse scores than the corresponding score.

	Score (s_κ)	V_k	Decision
Lowest Risk	**R1**	-1.11	OK
·	**R2**	-0.40	OK
·	**R3**	1.20	OK
·	**R4**	1.46	OK
·	**R5**	**1.48**	OK
·	**R6**	1.35	NOT CLEARED
Highest Risk	**R7**	0.78	NOT CLEARED

Table 3: Risk stratification by a hypothetical test and associated "values" (Equation (A.1))

Based on this analysis, the policymaker would choose to use risk score **R5** as a cutoff since this score maximizes the benefit with the lowest costs.

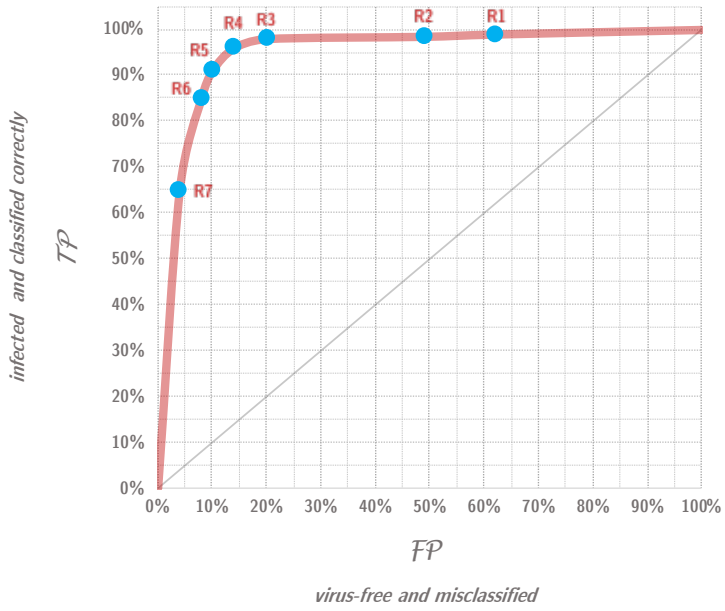


Figure 3: ROC curve for test performance in Table 2

This figure an provides an example of an ROC plot. The x-axis is the Type II error ($\mathcal{FP}$) as a percentage of all non-infected individuals. The higher the error rate, the larger percentage of non-infected individuals is excluded unintentionally. The y-axis shows the $\mathcal{TP}$: the percentage of infected individuals that is correctly screened out at a given cutoff. The points on the ROC are labeled to correspond to the error rates associated with each risk stratification level from the least risky (the smallest percentage of infected individuals missed) at R1 to the most risky, R7, which screens out the smallest percentage of infected individuals out (See: Table 2).

Said differently, using this test, there is no other risk score that would have as high a benefit (allowing reporters into the press conference) at a lower risk or, conversely, that would minimize the risk more for the same number of reporters admitted. ♢

Application Template - Determining a risk stratification policy (Basic)

All calculations may be done on a hand calculator or simple spreadsheet. Example 5 can be used as a template.

The basic strategy assumes a single cutoff for the entire population that is chosen to minimize total harm to the community (maximize the value of the policy), given the community priorities, needs, costs and objectives.

Input: Test accuracy
prevalence of COVID-19
Cost of $\mathcal{TN}$ and $\mathcal{FP}$ based on community objectives, economy and needs
Benefit of $\mathcal{TP}$ based on community objectives, economy and needs

Result: Risk stratification threshold (cutoff) that optimizes community objectives

3.2 Costs, benefits, and the Law of Risk-Based Decision Making

In setting up Example 5, we implicitly assumed that a policymaker would be comfortable making trade-offs among competing outcomes and the errors that accompany them. Is this reasonable?

The short answer in many cases is "No." Most people are, in fact, extremely *uncomfortable* considering trade-offs such as *"How many long machinists will I bring back to work if the expected harm is that an additional 55 year old individual will need to be admitted to ICU for a month?"*

However, whether a policymaker is comfortable with such a trade-off or not, The Law of Risk-Based Decision Making ensure that any policy that relies on risk stratification and screening using a test (assay, algorithm, age-based heuristic, etc.) that has error rates greater than zero, *is* making trade-offs, regardless of whether the PM evaluates these trade-offs explicitly or not.

Importantly, in addition to depending on the accuracy of the diagnostic, the risk stratification cutoff will also depend on the prevalence in the target population, and the various costs and benefits assigned to different errors and correct predictions. Table 4 demonstrates this by calculating the best cutoff using the same model as described in Examples 3 - 5, but then varying the assumptions about prevalence and costs.

In this Table 4, we allow the policymaker to make different assumptions about the prevalence of COVID-19, and the costs and benefits of different errors and accurate predictions. As can be seen from the "Risk cutoff" row, the range of selected cutoffs

	Base Case	Prevalence ▼	Prevalence ▲	Prevalence ▲▲	Cost of $\mathcal{FP}$ ▼	Cost of $\mathcal{FP}$ ▲	Benefit of $\mathcal{TN}$ ▼	Benefit of $\mathcal{TN}$ ▲	Cost of $\mathcal{FN}$ ▼	Cost of $\mathcal{FN}$ ▲
Best risk cutoff	***R5***	↕ ***R6***	↕ ***R4***	↕ ***R3***	↕ ***R4***	↕ ***R6***	↕ ***R4***	↕ ***R6***	↕ ***R7***	↕ ***R4***
	← ..... V_k ..... →									
R1	-1.11	-1.23	-0.88	-0.4	0.4	-9.05	-1.72	4.1	-1.08	-1.2
R2	-0.40	-0.42	-0.38	-0.3	0.8	-6.68	-1.22	6.5	-0.36	-0.5
R3	1.20	1.40	0.80	**0.0**	1.7	-1.36	-0.08	12.1	1.26	1.0
R4	1.46	1.74	**0.89**	-0.2	**1.8**	-0.34	**0.08**	13.2	1.58	**1.1**
$\mathcal{S}_\kappa$ **R5**	**1.48**	1.89	0.66	-1.0	1.7	0.26	13.7	13.7	1.75	0.8
R6	1.35	**1.90**	0.26	-1.9	1.5	**0.33**	-0.12	**13.9**	1.80	0.2
R7	0.78	1.75	-1.17	-5.1	0.9	0.26	-0.76	12.0	**1.83**	-2.0
	← ..... *Cost and Benefit assumptions* ..... →									
$c_{\mathcal{FN}}$	20	20	20	20	20	20	20	20	**5**	**60**
$b_{\mathcal{TP}}$	0	0	0	0	0	0	0	0	0	0
$c_{\mathcal{FP}}$	4	4	4	4	**1**	**20**	4	4	4	4
$b_{\mathcal{TN}}$	3	3	3	3	3	3	**1**	**20**	3	3
$p_{\mathbb{C}}$	20%	**5%**	**40%**	**80%**	20%	20%	20%	20%	20%	20%

Table 4: Risk stratification cutoffs depend on prevalence, costs and benefits

This table shows the differences in lowest permissible risk score policy, given changes in prevalence, costs and benefits. *Best risk cutoff* is the cutoff that minimizes harm. Below that score, individuals will be considered "high risk" so the cutoff represents the worst acceptable risk score for an individual to still be considered low risk, given the costs and benefits (i.e., risk scores from that risk score up to **R1** would be labeled "low risk.")

KEY:

▲, ▼ : a negative upward or downward change in the parameter

▲, ▼ : a positive upwards or downwards change

↕, ↕ : standards to became looser or tighter relative to the baseline.

ranges from a fairly conservative R4, to the most aggressive R7 depending on the assumptions made by the policymaker.

This variability may be seen as both a bug *and* a feature for clinicians.

On the one hand, Table 4 highlights the importance of having good information on which to base policy. For example, if the *true* prevalence rate is higher than a policymaker believes, but she has no data on this, she will make policy decisions based on the "bad" prevalence rate, but the policy will be executed in the world of the true prevalence rate. In many cases, that difference will make a material differences in risk cutoffs which in turn will affect which and how many workers are scheduled to return to work at different points in time.

On the other hand, the Table also highlights the great flexibility that clinicians can use in designing different risk stratification cutoffs for different demographics and job types, since different cutoffs may be determined for each segment using different costs and benefits in each case. We discuss this point in more detail in Section 3.4.

We have developed enough of the machinery on risk stratification to design a basic risk stratification system and to decide which risk tiers are cleared to return and which are not.

The basic algorithm is as follows:

Selecting a Risk Stratification Policy

1. Identify a test that can be practically administered in the target setting (e.g., that has costs and throughput speed that meet the needs of the application).
2. Determine $\mathcal{TP}_\kappa$, etc., for each of the possible results of the test, making use of the identities in Eqs. (1) - (4) as needed.[a]
3. Determine the appropriate cost and benefit functions for the community.[b]
4. Using the performance information from (2) and the cost and benefit information from (3), calculate V_κ for each candidate stratification level, κ.
5. Use the cutoff risk level, κ_{best}, that has the highest value of V_κ, for the test, given the prevalence information and costs and benefits.

[a] Most FDA approved tests, for example, publish this information, while most good quality quantitative models also disclose similar information.

[b] This is typically non-trivial and will form the core of a clinicians expertise. We discuss this in more detail in Section 4.

3.3 Are more expensive tests worth the expense?

In addition to getting better information, and making more rational policies based on that information, a policymaker may also, in some cases, have access to better quality tests. However, given the large number of tests that a community may be required to administer, it is reasonable for clinicians to analyze the value of more expensive vs. less expensive tests.

Using the machinery of the previous section, a PM can evaluate this value explicitly. This is done by calculating the best cutoff to use for each model, and then comparing the values of V_k.

If the difference in the value of using the more expensive tests exceeds the additional costs of those tests, a policymaker may be more likely to opt for the more expensive tests.

Example 6. Deciding whether to purchase more expensive tests

Assume that a policymaker has access to two different COVID-19 tests with the performance shown in Table 5. The policymaker needs to decide whether to mandate use of the less expensive, less accurate test, or the more expensive, more accurate test.

		Lower cost test		Higher cost test	
Risk level	s_κ	$\mathcal{TP}$	$\mathcal{FP}$	$\mathcal{TP}$	$\mathcal{FP}$
Lowest Risk	R1	99%	62%	99%	60%
·	R2	99%	49%	99%	50%
·	R3	98%	20%	98%	19%
·	R4	96%	14%	97%	10%
·	R5	91%	10%	92%	9%
·	R6	85%	8%	87%	6%
Highest Risk	R7	65%	4%	70%	4%

Table 5: Performance of two different hypothetical COVID-19 diagnostic tools

This table shows the hypothetical performance of two tests for screening and risk stratifying individuals, based on the likelihood that they are infected with COVID-19. For each test, $\mathcal{TP}$ and $\mathcal{FP}$ are the true positive rates (correct high risk detection) and false positive rates (low risk erroneously stratified as high risk), respectively. s_κ is the candidate cutoff.

We assume that the costs, benefits, and prevalence are:

$$
\begin{aligned}
c_{\mathcal{FN}} &= 20 \equiv \text{cost of false negative (infected allowed in)} \\
b_{\mathcal{TP}} &= 0 \equiv \text{benefit of true positive (correctly stopped infected)} \\
c_{\mathcal{FP}} &= 4 \equiv \text{cost of false positive (non-infected is kept out)} \\
b_{\mathcal{TN}} &= 3 \equiv \text{benefit of true negative (non-infected allowed in)} \\
p_{\mathbb{C}} &= 20\% \equiv \text{probability of COVID-19 (prevalence)\%}
\end{aligned}
$$

All else equal, we can see by examining Table 5 that there is no cutoff that can be chosen for which the less expensive test will perform better than the more expensive one, since, at each cutoff, the $\mathcal{TP}$ rate is higher and the $\mathcal{FP}$ rate lower for the more expensive test.

However to determine the whether the value of this difference is worth the expense, the policymaker calculates the best cutoff for each test and compares these:

Less expensive test:	$V_k = 1.48$	at cutoff **R5**
Less expensive test:	$V_k = 1.72$	at cutoff **R4**

As expected, the more expensive test is more valuable ($1.48 > 1.72$). However, the determination of whether or not the difference in value of 0.24 is substantial enough to justify the added cost will depend on the policymaker's assessment of the difference in value and the difference in cost. (Note that here we deliberately avoid using specific units to denominate the value, as each policymaker may use different calculations in determining value. See Section 4.) $\diamondsuit$

Application Template - Deciding whether to invest in better tests and data

All calculations may be done on a hand calculator or simple spreadsheet. Example 6 can be used as a template.

To determine whether a proposed investment in additional resources is cost effective, clinicians can use the templates developed so far, along with information on the expected performance of each model to calculate value that would be realized if the new investment were made.

In principle, this may be done using any scale for costs and benefits, however, it is most easily done if the scale of the expense of the new test, and the scale of the benefits and costs, are the same.

Input: Accuracy of current risk stratification approach
Accuracy of current risk stratification approach if investment made
prevalence of COVID-19
Cost under current approach of $\mathcal{FN}$ and $\mathcal{FP}$
Cost under proposed new approach of $\mathcal{FN}$ and $\mathcal{FP}$ (if different)
Benefit under current approach of $\mathcal{TN}$ and $\mathcal{TP}$
Benefit under proposed new approach of $\mathcal{TN}$ and $\mathcal{TP}$ (if different)

Result: "Optimal" risk stratification thresholds (cutoffs) for each approach
Relative (or absolute) value of new and current approaches

3.4 Flexible risk stratification policies for different needs across communities

Recently there has been heightened interest in formulating strategies and policies that would permit some "low risk" workers to return to work sooner than other "high risk" workers. For example, (Stein et al., 2020) proposes a *Total Harm Minimization* framework that prioritizes the timing of returning workers by both economic and health-risk factors, as well as by the degree to which a worker may productively work remotely, and thus continue to self-isolate if needed, during the first phases of the process.

Such strategies explicitly recognize that the risks faced by different segments of the population vary considerably. For example, while two individuals may have similar age and health profiles, if one of them has already been exposed to the novel coronavirus and the other has not, the exposed individual may be at lower risk than the one who has not been exposed.

Table 4 provided a preview of how clinicians can tailor risk stratification cutoffs to different situations. In the next example, we allow a policymaker to assume that for certain functions returning to work urgent than it is for others, and that keeping certain

employees away from their place of employment may be more or less costly than it is for others.

Example 7. Fact-based, flexible risk stratification for returning to the workforce

In this example, we consider a policymaker who is working towards bringing as many workers back into the workforce as possible, without undertaking unreasonable risk to the public.

The policymaker has consulted his economic advisors, and together the team has determined which industries are more critical to the region than others and which workers may work from home for a longer period with less loss to the economy and low health risks.

Furthermore, new serology tests now permit the identification of individuals who are expected to have acquired immunity because they have either been confirmed to have had COVID-19 and have recovered, or because they tested positive for protective antibodies. (We assume here that the scientific community has guided the policymaker in this regard.)

The group also has access to a fairly accurate test that can be used assign initial risk grades to individual workers. The performance of the test is known. as well as the effective accuracy and error rates for that procedure. (For convenience in this example, we assume that this test has the same accuracy, etc. as the one we have been using in the last example.

What remains is for the policy group to draft the costs and benefits associated with different policies. However, rather than draft a one-size-fits-all cost function, the group takes advantage of the work it has done with economic advisors and medical professionals and creates a set of cost functions that can be used with different segments of the workforce, depending on their characteristics. Table 6 shows the result.

The table shows how different segments of the population could be triaged for returning to work. As the risks and costs go down, or the benefits go up for a worker's return, the worker is given a more lenient threshold for returning to the workforce.

For example, the cost of keeping a worker who can work from home is much lower than the cost of keeping a worker at home who cannot work from home. In this case, the first worker who can work remotely would be permitted to return to work only with a relatively good test score of (low risk tier) **R4** or better (i.e., **R1** - **R4**. In contrast, the policy would take more risk in the case of the worker who can only work at his place of business. This worker would be allowed back with a much poorer risk score of **R7** (i.e., any risk score from **R1**-**R7**). ♢

	Base Case	Recent exposure		Can work remotely		Industry		Immunity	
		No	Yes	No	Yes	Critical	Non-critical	Yes	No
Best risk cutoff	***R5***	↕ ***R6***	↕ ***R3***	↕ ***R6***	↕ ***R4***	↕ ***R6***	↕***R4***	↕ ***R7***	***R5***
	← ..... V_k .....→								
strict ← POLICY → loose									
R1	-1.11	-1.23	-0.66	-9.05	0.97	2.54	-1.72	-1.33	-1.11
R2	-0.40	-0.42	-0.35	-6.68	1.36	4.49	-1.22	-0.43	-0.40
R3	1.20	1.40	**0.40**	-1.36	2.27	8.88	-0.08	1.59	1.20
R4	1.46	1.74	0.33	-0.34	**2.37**	9.71	**0.08**	2.01	1.46
R5	**1.48**	1.89	-0.16	0.20	2.27	10.12	0.04	2.29	**1.48**
R6	1.35	**1.90**	-0.82	**0.33**	2.05	**10.18**	-0.12	2.43	1.35
R7	0.78	1.75	-3.11	0.26	1.26	9.99	-0.76	**2.71**	0.78
	← ..... *Cost and Benefit assumptions* .....→								
$c_{\mathcal{FN}}$	20	20	20	20	20	20	20	▼**0**	20
$b_{\mathcal{TP}}$	0	0	0	0	3	0	0	0	0
$c_{\mathcal{FP}}$	4	4	4	▲**20**	▼**1**	4	4	4	4
$b_{\mathcal{TN}}$	3	3	3	3	3	▲**15**	▼**1**	3	4
$p_{\mathcal{C}}$	20%	▼**10%**	▲**60%**	20%	20%	20%	20%	▼**0%**	20%

Table 6: A flexible back-to-work policy based on information on economic and health risks

This table shows how the same test may be used to risk stratify different groups of workers for prioritizing back-to-work stages for different health and economic profiles. This is done by adjusting the costs, benefits and prevalence. The hypothetical factors are only considered in isolation here, but a more robust approach would consider combinations. *Best risk cutoff* is the cutoff that minimizes harm. Below that score, individuals will be considered "high risk" so the cutoff represents the worst acceptable risk score for an individual to still be considered low risk, given the costs and benefits (i.e., risk scores from that risk score up to **R1** would be labeled "low risk.")

KEY:

▲, ▼ : a negative upward or downward change in the parameter

▲, ▼ : a positive upwards or downwards change

↕, ↕ : standards to became looser or tighter relative to the baseline.

Application Template - Determining a flexible risk stratification policy (Intermediate)

All calculations may be done on a hand calculator or simple spreadsheet. Example 7 can be used as a template.

To develop a more flexible risk stratification policy that permits different segments to be evaluated differentially using segment-specific criteria, the basic approach discussed in Section 3.1 may be used with parameters tailored to each subgroup's specific roles and health characteristics.

Input: Accuracy of risk stratification approach
Prevalence of COVID-19 for each subgroup
Cost for each subgroup of $\mathcal{TN}$ and $\mathcal{FP}$ in context of community
Cost for each subgroup of of $\mathcal{TN}$ and $\mathcal{FP}$...
Benefit for each subgroup of $\mathcal{TP}$ and $\mathcal{FP}$...

Result: Risk stratification policy for each subgroup enabling looser/tighter risk as warranted by the community's needs

3.5 Unified risk stratification: incorporating disparate types of tests

Example 8. Incorporating different kinds of tests into a single policy

In this final example, consider a policymaker who has decided on the appropriate costs and benefits for each community but must rely on a variety of tests, depending on the test availability and capabilities at different testing facilities.

Imagine that there is a virology test that has fairly high accuracy, but which is only available in limited supply. Because of the limited supply, only about 10% of the community will be able to be tested using it. A second virology test, which is not as accurate, will be available for about a third of the remaining community. a small percentage of the of the community can be screened with a predictive model developed with machine learning techniques that has good accuracy, but which is only applicable to small subsets of the general population. For the remainder, a less accurate simple heuristic test will be administered (e.g., based on an individual's age and whether they live alone or with others).

For simplicity, we assume that the costs and benefits are the same in all regions of the community:

$$
\begin{aligned}
c_{\mathcal{FN}} &= 20 \equiv \text{cost of false negative (infected allowed in)} \\
b_{\mathcal{TP}} &= 0 \equiv \text{benefit of true positive (correctly stopped infected)} \\
c_{\mathcal{FP}} &= 4 \equiv \text{cost of false positive (non-infected is kept out)} \\
b_{\mathcal{TN}} &= 3 \equiv \text{benefit of true negative (non-infected allowed in)} \\
p_{\mathbb{C}} &= 20\% \equiv \text{probability of COVID-19 (prevalence)\%}
\end{aligned}
$$

The tests' results use different scales, which range from having seven risk tiers to having just two, and the quality is variable across the tests and between grades. Table 7 displays the performance for each test.

		Higher accuracy viral test		**Lower accuracy viral test**		**Low accuracy heuristic test**		**Machine Learning predictions**	
Risk level	s_κ	$\mathcal{TP}$	$\mathcal{FP}$	$\mathcal{TP}$	$\mathcal{FP}$	$\mathcal{TP}$	$\mathcal{FP}$	$\mathcal{TP}$	$\mathcal{FP}$
Low Risk	R1	99%	60%	95%	18%	**60%**	**11%**	93%	24%
·	R2	99%	50%	**85%**	**14%**	-	-	90%	18%
·	R3	98%	19%	-	-	-	-	**85%**	**14%**
·	R4	**97%**	**10%**	-	-	-	-	73%	10%
·	R5	92%	9%	-	-	-	-	65%	8%
·	R6	87%	6%	-	-	-	-	-	-
High Risk	R7	70%	4%	-	-	-	-	-	-

Table 7: Performance of two different hypothetical COVID-19 diagnostic tools

This table shows the performance of four different hypothetical tests for stratifying individuals into risk tiers, based on the likelihood that they are infected. $\mathcal{TP}$ and $\mathcal{FP}$ are true positive rates (correct high risk) and false positive rates (low risk erroneously stratified as high risk), respectively. The rows in bold indicate the test's performance at their best cutoffs, given the costs and benefits described in the example. This is the cutoff that minimizes harm. Below that score, individuals will be considered "high risk" so the cutoff represents the worst acceptable risk score for an individual to still be considered low risk, given the costs and benefits (i.e., risk scores from that risk score up to **R1** would be labeled "low risk.").

The policymaker would like screen as many individuals as possible using whatever tests are available. However, they are concerned that the test all have different performance and that each also use a different number of risk tiers to report test results. Table 7 makes it clear that an "R1" from one test does not communicate the same information about the probability of an individual being infected with COVID-19 as another, and thus two "R1" risk tiers cannot be compared directly.

However, given the scarcity of tests, the policymaker must find a way to use whatever

tests they can offer to community. Said differently, they need to *integrate* all of the disparate tests into a single scale.

To do this, the policymaker can first determine the best cutoff for each test, and then determine the probability that an individual who is screened has COVID-19.

This transformation may be done using the techniques developed earlier. In particular, using either Equation (C.2)) or Equation (C.8), each tests' performance at its best cutoff can be mapped to a probability, and the probabilities can then be directly compared.

Table 8 shows the result of performing these calculations for each test.

	Probability of COVID-19 infection if...		*Healthy but...*		
	...test result is negative	*...test result is positive*	*...test result is positive*	$\mathcal{TP}$	$\mathcal{FP}$
Viral test - H (R4)	**0.8%**	70.8%	**11.0%**	97%	10%
Viral test - L (R2)	**4.2%**	60.3%	**15.6%**	85%	14%
ML test (R3)	**4.2%**	60.3%	**15.6%**	85%	14%
Heuristic test (R1)	**10.1%**	57.7%	**11.1**	60%	11%

Table 8: Converting to probabilities to compare individuals tested with different tests [Equation (**??**), (C.8)]

This table shows how individuals tested using four different types of tests, each with different performance and each with a different scale, can be compared and integrated into a single coherent policy for risk stratification. The tests are a higher accuracy virology test, a lower accuracy virology test, a heuristic rule, and predictive algorithm developed using machine learning (ML) techniques; the best cutoff for each test is shown in parentheses after the test name. $\mathcal{TP}$ and $\mathcal{FP}$ are true positive rates (correct high risk) and false positive rates (low risk erroneously stratified as high risk) , respectively. The columns in bold show the percentage of negative (positive) results for which a the subject was actually COVID-19 infected (healthy), and thus increased the total harm for the community.

The results in Table 8 illustrate a number of useful points. First, note that the hypothetical high accuracy viral test dominates all other tests in terms of stratification, with uniformly lower percentages of erroneous risk scores for both healthy and infected individuals. If it were possible to use *only* that test, the reduction of harm would be substantial.

Second, notice that two of the tests, the lower accuracy viral test and the test developed using machine learning, produced identical performance, even though they used different methods and different scales. Thus, under theses assumptions about costs and incidence, the policy should treat results of the the tests as equivalent, when their best cutoffs are used.[20]

[20] This is not strictly true for a variety of reasons. For example, the variance of the estimates may be different, etc.

Finally, the "mislabeled" rates in the two bold columns appear to show three distinct risk stratifications: high (around 10% risk of COVID-19 infection), medium (about 4%) and low (about 1%).

Furthermore, the actual risk grades can now be used to calculate other quantities of interest, such as the expected number of new cases, under the policy and so forth.

Note also that, were the policymaker to move to an entirely probability-based scoring system, there would no longer be a need to create cutoffs at all: each risk grade in each test's scale can be mapped directly to a probability and these probabilities can be used to manage the staging of the return more effectively (see, Section 4).

Application Template - Integrating disparate risk stratification methods in a single policy

All calculations may be done on a hand calculator or simple spreadsheet. Example 8 can be used as a template.

Example applications:

- Map all testing protocols and tests to a unified risk stratification scale
- Map test results to probabilities of exposure, morbidity, mortality, etc.
- Calculate, e.g., expected number of new cases under policy, etc.
- Create a single risk stratification policy that can be used irrespective of any current or new tests that are selected by the policymaker.
- Use as input to more advanced risk management approaches that consider clustering of incidence (not discussed here)

Input: Accuracy of each risk stratification approach
Prevalence of COVID-19 for the community
Cost of $\mathcal{TN}$ and $\mathcal{FP}$ in context of community
Cost of $\mathcal{TN}$ and $\mathcal{FP}$...
Benefit of $\mathcal{TP}$...

Result: A common probability scale for directly comparing and ordering the results of all testing approaches that may be in use; quantitative information that can be used to manage and monitor staging.

4 Discussion

There are a number of open issues and extensions that may be considered to the basic framework we outlined in Section A. Here we discuss two of these: coming to terms with how best to intentionally assign cost and benefit values (rather than doing so unintentionally by implication), and potential improvements that allow more precise and customized risk stratifications.

4.1 The difficulty in assessing costs and benefits

In this paper, we have assumed that clinicians are able to explicitly state approximate values for cost and benefits. This is extremely challenging and, realistically, may only be possible very coarsely. In the case of novel coronavirus, this assessment is made more complicated by several factors, including:

- generally poor data on true prevalence, contagion rates, mortality rates, hospital admission rates (given infection), ICU admission rates (given hospitalization), etc. Without better data on these parameters and others, it will be exceedingly difficult to properly form policy.;
- complexity in both the pathology and evolution of the novel coronavirus, and how these are impacted by interventions, prior medical histories, genotype and phenotype information, etc.
- complexity in relating epidemiological phenomena to economic phenomena;
- complexity in understanding supply chain dynamics, and the many interrelationships between different economic sectors (e.g., retail and trucking), as well as the degree of modularity in any particular business or industry, and how these relate back to individuals.;
- the need to execute short-term policy within the context of long-term economic and health planning objectives

It is both outside of our knowledge and beyond the scope of the current paper to undertake this analysis in any detail for any specific case. Indeed such decisions require information on the economic and health impacts of different outcomes within a community, as well as information on the demographics of the population. These analyses also require a normalization and prioritization of outcomes that are measured on different scales, such as the health costs of self-isolation vs. the risk of overloading hospitals vs. the cost to the community and individuals of lost wages and economic output.

However, these need not be exceedingly refined or comprehensive in some cases. It may be the case, for example, that clinicians, economists, and health officials can easily

agree on the extreme cases, and then work backwards using, e.g., a point system for different costs and benefits.

Examples of costs and benefits could include a wide variety of global, national and local factors. Table 9 provides examples of some of the hypothetical factors that might go into a discussion with health and economic policy advisors while formulating a policy.

In some settings, decision makers can agree more easily on relative value than on absolute levels. In such cases, valuing factors in terms of points, rather than dollars and health outcomes may be more tractable. This is particularly so given the complexity, co-dependence, and tight coupling of the systems and actors in question.

Hypothetical examples of...

...potential costs of erroneously keeping uninfected people isolated ($c_{\mathcal{FP}}$)	**...potential benefits of getting uninfected people out of isolation ($b_{\mathcal{TP}}$)**	**...potential costs of erroneously returning infected people to population ($c_{\mathcal{FN}}$)**
▲ psychological illness	▲ economic output	▲ infection rate
▼ economic participation	▲ ability to care for others	▲ hospital loads
▲ expenses	▲ national security	▲ spread of COVID-19
▲ food/medical insecurity	▲ services and product availability	▲ mortality due to COVID-19
▲ domestic abuse	▲ resources for community support	▲ reduced health svcs. access
▲ disruption of supply chain	▲ quality of life	▼ resilience of community
▲ lifestyle-related morbidity	▲ lifestyle-related health	▼ prolonged population lock-down
etc.	etc.	etc.

Table 9: Some hypothetical factors for discussing a return-to-work policy. (*Example only*)

This table shows some of the dimensions that clinicians might consider in a hypothetical discussion of return-to-work policy construction. It is intended as an example only. The actual dimensions will differ and be weighted differently in various real settings. (▲, ▼indicate a negative increase or decrease as the result of an error; ▲ indicates a positive increase as the result of a benefit.)

4.2 Risk stratification without cutoffs: More refined tests minimize harm

While cutoffs can be chosen by clinicians to maximize overall benefits to a community these are only "best" when the options are restricted to designating a single cutoff. If the clinicians and the community have access to more refined tests, these may be used to create much more flexible screening procedures.

In such cases, particularly if the test is calibrated to produce a continuous scale and can be calibrated to produce probability estimates, a policy can be designed to accommodate a wide range of individual situations. This permits highly customized assessments of specific sub-populations, and even of individuals.

Furthermore, such tools can be explicitly designed to incorporate multiple dimensions of risk (e.g., age AND health status AND recent exposure, etc.) rather than assessing each independently, as in Example 7. Doing this allows risk stratification to be informed by a fuller picture of the individual or sub-population, and also reduces double-counting substantially.

It can be shown (Stein, 2005) that such approaches are generally superior to simple cutoff-based approaches. Furthermore, as better empirical data on COVID-19 becomes available, this type of more refined analysis and prediction is likely to become increasingly available.

The question of how to best determine a scale and to then calibrate involves a number of dimensions and may be done in several ways, each offering different advantages and limitations, depending on the applications (see, (Bohn and Stein, 2009, , Chapter 4) for a discussion.

5 Conclusion

In this note we have tried to provide guidance for clinicians on ways to make more informed decisions as part when developing risk stratification approaches for bringing individuals back to work and the general public prior to the introduction of an effective vaccine for COVID-19.

Our discussion was framed in this context and our examples focused on COVID-19 entirely. However, the techniques we used in this context are general, as are their mathematical formulations.

These general techniques can provide a grounding in facts and science for clinicians to draw on in evaluating the challenging policy decisions that now face many national and local governments.

These methods perform better with more accurate data, and as a result, they can also help clinicians understand how best to collect this information, and up until what point the benefits of data collection continue to outweigh the costs.

Finally, the methods are flexible and can be readily adapted to local needs, using whatever information is available.

Many discussions of policy were initially focused on emergency tactics to prevent healthcare systems and communities from being overrun with COVID-19 outbreaks. As these discussions now turn towards longer-term strategies to manage how communities might start to return to more "normal" activities, when the threats of COVID-19 begin to abate, we hope that this type of analysis will inform better, more effective, and less costly policies for total harm minimization.

References

AHQR. AHRQ National Scorecard on Hospital-Acquired Conditions Updated Baseline Rates and Preliminary Results 2014–2017. Report, The Agency For Heathcare Research and Quality, 2019.

Christopher R Bilder and Joshua M Tebbs. Pooled-testing procedures for screening high volume clinical specimens in heterogeneous populations. *Statistics in medicine*, 31(27):3261–3268, 11 2012. doi: 10.1002/sim.5334. URL https://pubmed.ncbi.nlm.nih.gov/22415972.

Jeffrey R. Bohn and Roger M. Stein. Approaches to improving a bank's share value using credit-portfolio management and credit-transfer pricing. *Journal Of Investment Management*, 11 (2):47–72, 2013.

J.R. Bohn and R.M. Stein. *Active credit portfolio management in practice*. Wiley Finance, 2009.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, volume 2, pages 973–8, San Francisco, CA, USA, August 2001. IJCAI, Morgan Kaufmann Publishers Inc.

David M. Green and John A. Sweats. *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York, 1966.

David L. Katz, Roger M. Stein, Wesley Pegdan, Maria Chikina, and Dina Aronson. COVID-19 risk modeling options, conclusions & concerns to date. White paper, True Health Initiative, 2020. URL (https://www.truehealthinitiative.org/wp-content/uploads/2020/04/COVID-19-Risk-Modeling-Options-Conclusions-and-Concerns-to-Date-2020-04-08.pdf).

Sean Khozin and Roger M. Stein. Explanations, intuition and adjustments for low external validity using simple mathematics. *forthcoming*, 2020.

Angelica LaVito, Kristen V Brown, and Keshia Clukey. New york finds virus marker in 13.9%, suggesting wide spread. *Bloomberg*, April 23 2020.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(2), 2001.

P. Rui and K. Kang. National Hospital Ambulatory Medical Care Survey: 2017 emergency department summary tables. Report Table 11, National Center for Health Statistics, 2018.

Roger M. Stein. The relationship between default prediction and lending profits: Integrating roc analysis and loan pricing. *Journal of Banking and Finance*, 29:1213–1236., 2005.

Roger M. Stein, Daniel J. Arbess, Michael Kanef, David L. Katz, and Timothy S. Walsh. A total-harm-minimization framework for developing expedient and low-risk return-to-the-workforce policies during the COVID-19 pandemic. White paper, True Health Initiative, 2020. URL https://www.truehealthinitiative.org/wp-content/uploads/2020/04/A-Total-Harm-Minimization-Framework-for-Developing-Expedient-and-Low-Risk-Return-to-the-Workforce-Policies-During-the-COVID-19-Pandemic-2020-04-08.pdf.

A Appendix: Setting a policy cutoff

Assume that there is a set of benefits and costs that we can calculate for a correctly and incorrectly stratifying an individual.[21] Denote these costs and benefits as follows:

$$
\begin{aligned}
b_{\mathcal{TN}} &\equiv \text{benefit of correctly stratifying a disease-free individual as low-risk} \\
b_{\mathcal{TP}} &\equiv \text{benefit of correctly stratifying of an infected individual as high-risk} \\
c_{\mathcal{FN}} &\equiv \text{cost of incorrectly stratifying of an infected individual as low-risk} \\
c_{\mathcal{FP}} &\equiv \text{cost of incorrectly stratifying of an disease-free individual as high-risk}
\end{aligned}
$$

Imagine further that the diagnostic produces a score, s_i for each individual, that ranges from s_1 ("lowest risk") through s_K highest risk. This score could be the propensity score of a machine learning algorithm, the concentration level (cp/μL) of SARS-CoV-2 detected using an assay, a person's age, etc.

The goal of the policymaker is to determine the score cutoff κ to use for stratifying the risk of the population into "high risk" (scores higher than κ) and "low risk" (scores lower than κ). Without a loss of generality, let k be the percentile of s_κ in the population. [22] Because the diagnostic is not perfect, any κ selected other than s_1 or s_K will produce accuracy and error rates $\mathcal{TN}_k, \mathcal{TP}_k, \mathcal{FN}_k$ and $\mathcal{FP}_k$.

Given this information, the total value of a policy, V_k, that uses this diagnostic with the cutoff k is:

$$
V_k = p^*_{\mathbb{F}} \left[b_{\mathcal{TN}} \times \mathcal{TN}_k - c_{\mathcal{FP}} \times \mathcal{FP}_k \right] + p^*_{\mathbb{C}} \left[b_{\mathcal{TP}} \times \mathcal{TP}_k - c_{\mathcal{FN}} \times \mathcal{FN}_k \right], \tag{A.1}
$$

where, as before, $p^*_{\mathbb{C}}$ and $p^*_{\mathbb{F}}$ are the true prevalence and one minus the true prevalence, respectively.

This is simply the weighted net value of a correct vs. incorrect classification of an individual as low risk plus the weighted value of a correct vs. incorrect classification of an individual as high risk. The the weighting is done based on the probability of being virus-free or infected and the probability of being correct or incorrect depends on the model accuracy ($\mathcal{TP}_k$, etc.).

In order to maximize the value (equivalent to minimizing the negative value) we can V_k to zero, differentiating with respect to k and rearrange terms. This gives the slope of a line with marginal cost equal to zero.

This gives:

[21] It is beyond the scope of this note to discuss the many factors that must go into such determinations. However, regardless of whether clinicians explicitly make these calculations or not, by the Law of Risk-Based Decision Making the costs and benefits will be implied in any risk stratification policy.

[22] For completeness, $k = \#\{s_i < \kappa\}/N,\ i = 1...N$, where N is the size of the total population and $\#\{x\}$ is the count of the number of cases in which x is true.

$$S = \frac{(1 - p_{\mathbb{C}}) \times [c_{\mathcal{FP}} + b_{\mathcal{TN}}]}{p_{\mathbb{C}} \times [c_{\mathcal{FN}} + b_{\mathcal{TP}}]}.$$

The point at which a line with slope S, as defined above, forms a tangent to the ROC curve for a stratification method, defines optimal cutoff, given a particular set of costs and benefits: this point will be the one at which marginal payoffs (costs) are zero. (Green and Sweats, 1966) provides a discussion of this approach and an analytic formulation of the problem as applied to ROC analysis.

The goal of the policymaker is to maximize the value of the policy, which is done by selecting a value of k that maximizes V_k. This may be done graphically using ROC curves or analytically, as above. (see: (e.g., Stein, 2005) for details and exertions).

As noted earlier, for real-world problems involving risk stratification systems that may have a variety of attributes, and for which there may only be a small number of potential cutoff values, it may be convenient to calculate $\mathcal{TP}_k, \mathcal{FP}_k$, etc. for the range of k values, and to then use the cutoff that results in the maximum value of V_k.

B Appendix: A general analytic result using Bayes' Rule

Bayes' Theorem gives:

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}, \tag{B.1}$$

where, as usual,

$p(x)$: the probability of event x occurring; and
$p(x|y)$: the probability of event x occurring given that y is observed.

For more involved calculations, we note that the chain rule permits calculation of joint probabilities based on their conditional products:

$$p(A_n, \ldots, A_1) = p(A_n|A_{n-1}, \ldots, A_1) \cdot p(A_{n-1}, \ldots, A_1). \tag{B.2}$$

For example, if we were worried about spreading the novel coronavirus at a press conference, but also worried about restricting reporters' access to the press conference, we might consider screening each reporter as they enter briefing room. In this case, we would be concerned about infected individuals entering the briefing room and possibly infecting others. If we wished to determine the probability that an individual had COVID19, given that she got a negative test result, $\mathcal{N}$, Equation (B.1) becomes:

$$p(\texttt{COVID-19} \mid \mathcal{N}) = \frac{p(\mathcal{N} \mid \texttt{COVID-19}) \times p(\texttt{COVID-19})}{p(\mathcal{N})}.$$

If there were n_T total test given, (negative test result, $\mathcal{N}$ indicating no COVID19 and positive result $\mathcal{P}$ indicating COVID-19 present), Equation (B.1) becomes:

$$p(\texttt{COVID-19} \mid \mathcal{N}) = \frac{(n_{\mathbb{C}\mathcal{N}}/n_{\mathbb{C}}) \times (n_{\mathbb{C}}/n_T)}{(n_{\mathcal{N}}/n_T)}. \tag{B.3}$$

where

$$\begin{aligned}
n_{\mathbb{C}} &\equiv \text{total number of COVID-19 infected individuals} \\
n_{\mathbb{W}} &\equiv \text{total number of individuals tested who are well (no COVID19)} \\
n_T &\equiv \text{total number of individuals tested} = n_{\mathbb{C}} + n_{\mathbb{W}} \\
n_{\mathcal{P}} &\equiv \text{total number of individuals who test positive} \\
n_{\mathcal{N}} &\equiv \text{total number of individuals who test negative} \\
n_{\mathbb{C}\mathcal{P}} &\equiv \text{total number of individuals with COVID-19 who test positive} \\
n_{\mathbb{C}\mathcal{N}} &\equiv \text{total number of individuals with COVID-19 who test negative}
\end{aligned}$$

which (of course) simplifies to

$$p(COVID19 \mid \mathcal{N}) = \frac{n_{\mathbb{C}\mathcal{N}}}{n_{\mathcal{N}}} \equiv \mathcal{FN}. \tag{B.4}$$

The false positive rate may be calculated similarly.

It is useful at times to recall that $\mathcal{TP}, \mathcal{FP}$, etc. do not depend on $p_{\mathbb{C}}$, the incidence rate. (See Appendix D for a proof.)

C Appendix: Calibrating the probability of a COVID-19 infection as new information arrives

In some cases, we can make simple adjustments, using the Bayesian framing, in order to adjust for differences between the noisy population prevalence estimates and (perhaps new) true population prevalence to better characterize the effectiveness of a diagnostic or model. In the simplest case, a straightforward adjustment to the "bad" probability, using a new "better" estimate of prevalence may suffice (see, e.g., Bohn and Stein, 2013; Khozin and Stein, 2020, for details).[23]

It can be shown, for example, that absent systematic selection bias (see: footnote 23).

One such fix-up is:

$$p(\mathbb{S} \mid \mathcal{R})^{tru} \equiv p^{tru}_{\mathbb{S}\mathcal{R}} = p^{tru}_{\mathbb{S}} \left[\frac{p^{bad}_{\mathbb{S}\mathcal{R}} \times \left[1 - p^{bad}_{\mathbb{S}}\right]}{p^{bad}_{\mathbb{S}} - (p^{bad}_{\mathbb{S}} \times p^{bad}_{\mathbb{S}\mathcal{R}}) + (p^{tru}_{\mathbb{S}} \times p^{bad}_{\mathbb{S}\mathcal{R}}) - (p^{tru}_{\mathbb{S}} \times p^{bad}_{\mathbb{S}})} \right]. \tag{C.1}$$

where

$$\begin{aligned}
\mathbb{S} &\equiv \text{the true disease status (} \texttt{COVID-19} \text{ or } \texttt{well}\text{)} \\
\mathcal{R} &\equiv \text{the result of the diagnostic test } \mathcal{P} \text{ or } \mathcal{N} \\
p_{\mathbb{S}} &\equiv \text{the true prevalence of } \mathbb{S} \\
p^{tru}_{x} &\equiv \text{is the correct estimate of the probability of } x \text{ if correct data were used} \\
p^{bad}_{x} &\equiv \text{is the bad estimate of the probability of } x \text{ using unreliable data, and} \\
p_{\mathbb{S}\mathcal{R}} &\equiv p(\mathbb{S} \mid \mathcal{R}).
\end{aligned}$$

So, for example, the true probability of a person who tested positive actually having COVID-19, can be computed from an estimate, $p^{bad}_{\mathbb{C}\mathcal{P}}$, that was originally made based on unreliable data:

$$p(\texttt{COVID-19} \mid \mathcal{P})^{tru} \equiv p^{tru}_{\mathbb{C}\mathcal{P}} = p^{tru}_{\mathbb{C}} \left[\frac{p^{bad}_{\mathbb{C}\mathcal{P}} \times \left[1 - p^{bad}_{\mathbb{C}}\right]}{p^{bad}_{\mathbb{C}} - (p^{bad}_{\mathbb{C}} \times p^{bad}_{\mathbb{C}\mathcal{P}}) + (p^{tru}_{\mathbb{C}} \times p^{bad}_{\mathbb{C}\mathcal{P}}) - (p^{tru}_{\mathbb{C}} \times p^{bad}_{\mathbb{C}})} \right]. \tag{C.2}$$

The above equations **can only be used in certain special settings**, we introduce them here only to demonstrate the magnitude of the impact that unreliable data may have on a risk stratification, and because certain intermediate results of the proof[24] are useful for calibration.

[23] The adjustment shown here assumes that there is no systematic sampling bias, i.e., that the actual missing cases are MCAR. Clearly, in the case of the COVID-19 testing discussed here, this is not the case. However, in such settings, more detailed analysis may also be used to adjust probabilities (see, e.g., Khozin and Stein, 2020, and references therein).

[24] Much of the proof of this result follows closely a version given in (Elkan, 2001), though the one given here contains more of the details, and contains extensions.

Proof of Eq. (C.2)

We begin with Equation (B.1):

$$p(A \mid B) = \frac{p(B \mid A) \times p(A)}{p(B)},$$

from which it also follows that,

$$p(B \mid A) = \frac{p(A \mid B) \times p(B)}{p(A)}. \tag{C.3}$$

For convenience, let

$$\begin{aligned} p(A \mid B) &= p(\texttt{COVID}-\texttt{19} \mid \mathcal{P}) \equiv p_{\mathbb{C}\mathcal{P}}, \text{ and} \\ p(B \mid A) &= p(\mathcal{P} \mid \texttt{COVID}-\texttt{19}) \equiv p_{\mathcal{P}\mathbb{C}}. \end{aligned}$$

$$p_{\mathbb{C}\mathcal{P}} = \frac{p_{\mathcal{P}\mathbb{C}} \times p_{\mathbb{C}}}{p_{\mathcal{P}}}$$

and let $p_{\mathbb{W}} = 1 - p_{\mathbb{C}}$ be the probability of not having COVID-19 (being *well*), and

$$\begin{aligned} p_{\mathcal{P}} &= p_{\mathbb{C}} \times p_{\mathcal{P}\mathbb{C}} + p_{\mathbb{W}} \times p_{\mathcal{P}\mathbb{W}} \\ &= p_{\mathbb{C}} \times p_{\mathcal{P}\mathbb{C}} + (1 - p_{\mathbb{C}}) \times p_{\mathcal{P}\mathbb{W}} \end{aligned} \tag{C.4}$$

If our estimate of $p_{\mathbb{C}}$ is based on a "bad" estimate, then, $p_{\mathbb{C}\mathcal{P}}$ will also be "bad." Using the notation above, the erroneous ("bad") estimate of $p_{\mathbb{C}\mathcal{P}}$, $p^{bad}_{\mathbb{C}\mathcal{P}}$, is then:

$$\begin{aligned} p^{bad}_{\mathbb{C}\mathcal{P}} &= \frac{p^{bad}_{\mathcal{P}\mathbb{C}} \times p^{bad}_{\mathbb{C}}}{p^{bad}_{\mathcal{P}}} \\ &= \frac{p^{bad}_{\mathcal{P}\mathbb{C}} \times p^{bad}_{\mathbb{C}}}{p^{bad}_{\mathbb{C}} \times p^{bad}_{\mathcal{P}\mathbb{C}} + (1 - p^{bad}_{\mathbb{C}}) \times p^{bad}_{\mathcal{P}\mathbb{W}}} \\ &= \frac{p^{bad}_{\mathbb{C}}}{p^{bad}_{\mathbb{C}} + (1 - p^{bad}_{\mathbb{C}}) \times \theta} \end{aligned} \tag{C.5}$$

where $\theta^{bad} = p^{bad}_{\mathcal{P}\mathbb{W}}/p^{bad}_{\mathcal{P}\mathbb{C}}$. Note that because θ is equivalent to $\mathcal{FP}/\mathcal{TP}$, it does not depend on $p_{\mathbb{C}}$ (see, Appendix D), and therefore

$$\theta^{bad} = \theta^{tru} = \theta,$$

so we also have

$$p^{tru}_{\mathbb{C}\mathcal{P}} = \frac{p^{tru}_{\mathbb{C}}}{p^{tru}_{\mathbb{C}} + (1 - p^{tru}_{\mathbb{C}}) \times \theta}, \tag{C.6}$$

where the superscript tru indicates the true value of the parameter, which is the equivalent to Equation (C.5), but using the good estimate $p^{tru}_{\mathbb{C}}$.

Although theta does not depend on $p_{\mathbb{C}}$, $p^{bad}_{\mathbb{C}\mathcal{P}}$ and $p^{tru}_{\mathbb{C}\mathcal{P}}$ do depend on $p^{bad}_{\mathbb{C}}$ and $p^{tru}_{\mathbb{C}}$, respectively. Thus, in order to accomplish the adjustment, we can solve for θ in terms of $p^{bad}_{\mathbb{C}\mathcal{P}}$. This gives

$$\begin{aligned} \theta &= \frac{p^{bad}_{\mathbb{C}} \times (1 - p^{bad}_{\mathbb{C}\mathcal{P}})}{p^{bad}_{\mathbb{C}\mathcal{P}} \times (1 - p^{bad}_{\mathbb{C}})} \\ &= \frac{p^{bad}_{\mathbb{C}} - (p^{bad}_{\mathbb{C}} \times p^{bad}_{\mathbb{C}\mathcal{P}})}{p^{bad}_{\mathbb{C}\mathcal{P}} - (p^{bad}_{\mathbb{C}\mathcal{P}} \times p^{bad}_{\mathbb{C}})} \end{aligned} \tag{C.7}$$

Plugging Equation (C.7) back into Equation (C.6) gives

$$\begin{aligned}
p_{\mathbb{C}\mathcal{P}}^{tru} &= \frac{p_{\mathbb{C}}^{tru}}{p_{\mathbb{C}}^{tru} + (1 - p_{\mathbb{C}}^{tru}) \times \frac{p_{\mathbb{C}}^{bad} - (p_{\mathbb{C}}^{bad} \times p_{\mathbb{C}\mathcal{P}}^{bad})}{p_{\mathbb{C}\mathcal{P}}^{bad} - (p_{\mathbb{C}\mathcal{P}}^{bad} \times p_{\mathbb{C}}^{bad})}}, \\
&= \frac{p_{\mathbb{C}}^{tru}}{p_{\mathbb{C}}^{tru} + (1 - p_{\mathbb{C}}^{tru}) \times \frac{p_{\mathbb{C}}^{bad} \times (1 - p_{\mathbb{C}\mathcal{P}}^{bad})}{p_{\mathbb{C}\mathcal{P}}^{bad} \times (1 - p_{\mathbb{C}}^{bad})}}, \\
&= \frac{p_{\mathbb{C}}^{tru} \times p_{\mathbb{C}\mathcal{P}}^{bad} \times (1 - p_{\mathbb{C}}^{bad})}{p_{\mathbb{C}}^{bad} + (p_{\mathbb{C}\mathcal{P}}^{bad} \times p_{\mathbb{C}}^{tru}) - (p_{\mathbb{C}}^{tru} \times p_{\mathbb{C}}^{bad}) - (p_{\mathbb{C}}^{bad} \times p_{\mathbb{C}\mathcal{P}}^{bad})}, \\
&= p_{\mathbb{C}}^{tru} \left[\frac{p_{\mathbb{C}\mathcal{P}}^{bad} \times [1 - p_{\mathbb{C}}^{bad}]}{p_{\mathbb{C}}^{bad} + -(p_{\mathbb{C}\mathcal{P}}^{bad} \times p_{\mathbb{C}}^{bad}) + (p_{\mathbb{C}}^{tru} \times p_{\mathbb{C}}^{bad}) - (p_{\mathbb{C}}^{tru} \times p_{\mathbb{C}}^{bad})} \right]. \blacksquare
\end{aligned}$$

Importantly, because θ is a the ratio $\mathcal{FP}/\mathcal{TP}$ does not involve $p_{\mathbb{C}}$, Equation (C.2) may be simplified substantially if we know the performance ($\mathcal{FP}$ and $\mathcal{TP}$) of the test, which can be convenient:

$$p_{\mathbb{C}\mathcal{P}}^{tru} = \frac{p_{\mathbb{C}}^{tru}}{p_{\mathbb{C}}^{tru} + (1 - p_{\mathbb{C}}^{tru}) \times \frac{\mathcal{FP}}{\mathcal{TP}}}, \tag{C.8}$$

since any specific test result will be associated with a specific $\mathcal{FP}$ and $\mathcal{TP}$, which, in turn, will result in a specific probability estimate(see, Appendix sec:bayes). For tests that produce a range of test values, any specific test value will still have a specific $\mathcal{FP}$ and $\mathcal{TP}$ and resulting probability associated. (See Figure 3).

D Appendix: Proof that $\mathcal{TP}$ (and $\mathcal{FP}, \mathcal{FN}, \mathcal{TN}$) do not depend on the prevalence $p_{\mathbb{C}}$

We wish prove that $\mathcal{TP}$ and $\mathcal{TN}$ do not depend on $p_{\mathbb{C}}$. It is sufficient to prove that $\mathcal{TP}$.

We begin by proving that $\mathcal{TP} = 1 - \mathcal{FN}$.

$$
\begin{aligned}
\mathcal{TP} \;=\; p(\mathcal{P} \mid \mathbb{C}) \quad &= \quad \frac{p(\mathbb{C} \mid \mathcal{P}) \times p_{\mathcal{P}}}{p_{\mathbb{C}}} \\
&= \quad \frac{\frac{n_{\mathbb{C}\mathcal{P}}}{n_{\mathcal{P}}} \times \frac{n_{\mathcal{P}}}{n_T}}{\frac{n_{\mathbb{C}}}{n_T}} \\
&= \quad \frac{n_{\mathbb{C}\mathcal{P}}}{n_{\mathcal{P}}} \times \frac{n_{\mathcal{P}}}{n_T} \times \frac{n_T}{n_{\mathbb{C}}} \\
&= \quad \frac{n_{\mathbb{C}\mathcal{P}}}{n_{\mathbb{C}}} \\
&= \quad \frac{n_{\mathbb{C}} - n_{\mathbb{C}\mathcal{N}}}{n_{\mathbb{C}}} \\
&= \quad \frac{n_{\mathbb{C}}}{n_{\mathbb{C}}} - \frac{n_{\mathbb{C}\mathcal{N}}}{n_{\mathbb{C}}} \\
&= \quad 1 - \mathcal{FN}. \;\blacksquare
\end{aligned}
$$

It can similarly be shown that

$$\mathcal{TN} = 1 - \mathcal{FP}.$$

It now remains to show that that $\mathcal{TP}$ does not depend on $p_{\mathbb{C}}$, which, by exertion of the first proof, establishes the result for $\mathcal{FN}, \mathcal{TP}$ and $\mathcal{FN}$ as well.

$$\begin{aligned} \mathcal{TP} &= \frac{n_{\mathbb{C}\mathcal{P}}}{n_{\mathbb{C}}}, \\ n_{\mathbb{C}\mathcal{P}} &= n_{\mathbb{C}} \times n_T \\ \\ n_{\mathbb{C}\mathcal{P}} &= (p_{\mathbb{C}} \times n_T) \times \mathcal{TP}. \end{aligned}$$

Now let $p_{\mathbb{C}}^{new}$ be an alternative value of $p_{\mathbb{C}}$ such that

$$\begin{aligned} \frac{p_{\mathbb{C}}}{p_{\mathbb{C}}^{new}} &= \lambda. \\ p_{\mathbb{C}}^{new} &= \lambda \times p_{\mathbb{C}} \qquad \text{(D.1)} \\ n_{\mathbb{C}}^{new} &= \lambda \times p_{\mathbb{C}} \times n_T \\ &= \lambda \times n_{\mathbb{C}} \qquad \text{(D.2)} \end{aligned}$$

Now

$$\begin{aligned} n_{\mathbb{C}\mathcal{P}}^{new} &= p_{\mathbb{C}}^{new} \times n_T \times \mathcal{TP}. \\ &= \lambda \times p_{\mathbb{C}} \times n_T \times \mathcal{TP}. \\ n_{\mathbb{C}\mathcal{P}}^{new} &= \lambda \times n_{\mathbb{C}\mathcal{P}} \qquad \text{(D.3)} \end{aligned}$$

The new value of $\mathcal{TP}$, $\mathcal{TP}^{new}$ is the value of $\mathcal{TP}$ based on the new prevalence rate $p_{\mathbb{C}}^{new}$:

$$\mathcal{TP}^{new} = \frac{n_{\mathbb{C}\mathcal{P}}^{new}}{n_{\mathbb{C}}^{new}}, \qquad \text{(D.4)}$$

Substituting Equation (D.2) and (D.3) into Equation (D.4) gives

$$\begin{aligned} \mathcal{TP}^{new} &= \frac{n_{\mathbb{C}\mathcal{P}}^{new}}{n_{\mathbb{C}}^{new}} \\ \\ \mathcal{TP}^{new} &= \frac{\lambda \times n_{\mathbb{C}\mathcal{P}}}{\lambda \times n_{\mathbb{C}}} \\ &= \mathcal{TP}. \ \blacksquare \end{aligned}$$

E Risk Stratification Workbench: Staging, cutoff, and policy evaluation

In order to facilitate education regarding the techniques discussed in this article, we have implemented a web-based tool called *Risk Stratification Workbench* for performing these analyses.[25]

Risk Stratification Workbench is publicly available at no charge, and can be found at:

http://www.rogermstein.com/covid-19-resources/.

Risk Stratification Workbench is intended to be used in conjunction with this article, however, the following two pages provide a "Cheat Sheet" for quick reference.

[25] Although the software is available at no cost, some caveats are in order: This software is only intended for educational purposes and there is no express or implied warranty regarding the use of this tool or its outputs. Although we believe the code to be free of errors, the user should note that the accuracy of the software and results are not guaranteed, and the applicability of the analysis may or may not be appropriate for any specific purpose. In light of this, any decisions that may or may not result from the use of this software are the responsibility of the user. (Also note, that the author has no medical training.)

Risk Stratification Workbench Cheat Sheet

Enter information about community priorities on a scale of 1-100 (see, e.g. Section 3.1)

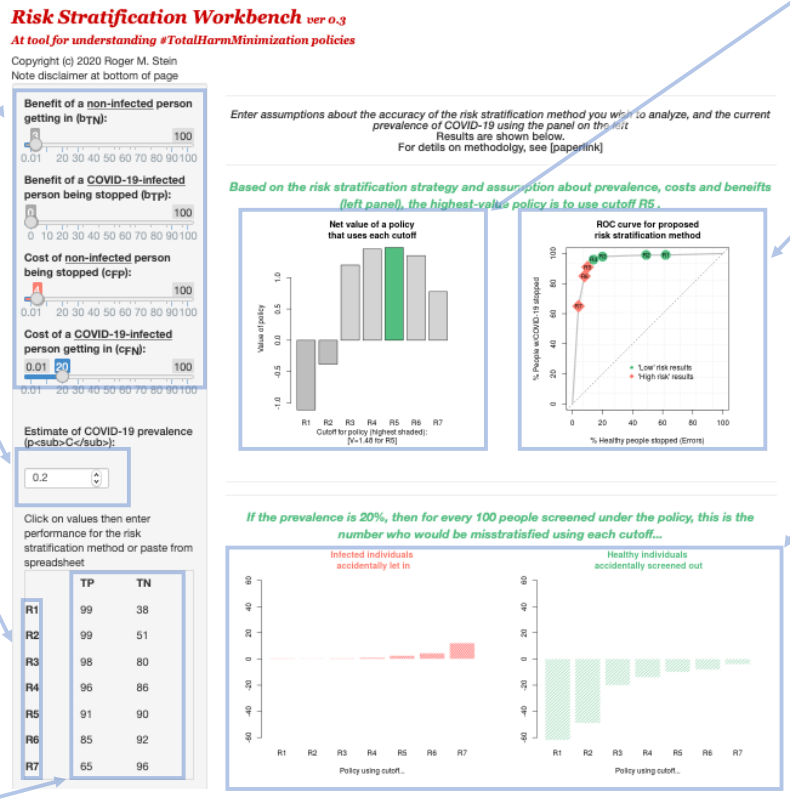
Enter the prevalence of COVID-19 based on the best information available using a probability scale [0.0 - 1.0] (see, e.g. Section 2.3)

Note the scale for risk stratification system; K1 is lowest risk…K7 is riskiest. (See Section 2.1)

If there are fewer than seven levels in the risk stratification system, start with the riskiest levels (bottom to top) until all levels are entered; then fill the rest of the table by repeating performance for the least risky.

Enter data on the accuracy of the risk stratification method. (see, e.g. Section 3.1)

(e.g. TP = % of COVID-19 infected person is classified as being "high risk"



View plot showing the relative value of a policy using each risk stratification as a cutoff (see, Section 3.1)

View plot showing the ROC curve for the stratification method (see, Section 3.1)

View plots showing how policies based on each possible risk cutoff level in the system would perform. (see, Section 2).

The height of the bars indicates the number infected individuals who should not have been approved but were allowed to "enter" (left) and the number of uninfected individuals who should have been approved but were not allowed to "enter" (right).

For details, see: Stein, R. M (2020), "Where to draw the line for risk stratifications: Designing return-to-work policies that consider diagnostic error, costs, benefits and COVID-19."

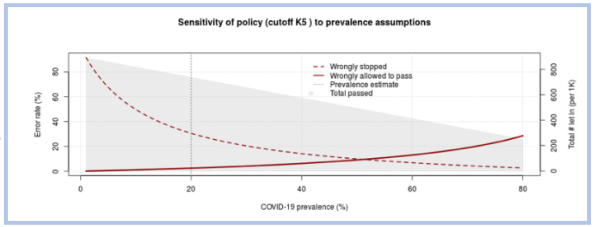
Risk Stratification Workbench Cheat Sheet

View the graph showing the how the true probability of having (not having) COVID-19 varies for the risk stratification system at the best cutoff would change if updated (or more accurate) information on the prevalence becomes available. (see, Section 2.3)

The graph shows the percentage of individuals who tested negative ("low risk") would actually be COVID-19 infected and allowed to "enter" and the number of individuals who tested positive ("high risk") who would actually be non-infected, but prevented from "entering."

In addition, the total number of individuals (per 1,000 screened) who would be allowed to "enter is shown in gray.

Using the cutoff K5 (that maximizes value under the assumptions in the left panel), for every 100 people screened under the policy, this is the number who would be misstratisfied at different prevalences...



(Each row shows a different assumption about the population prevalence)

prevalence	passed.with.COVID.19	stopped.but.healthy
1.00	0.10	91.58
10.00	1.10	49.72
20.00	2.44	30.53
30.00	4.11	20.41
40.00	6.25	14.15
50.00	9.09	9.90
60.00	13.04	6.83
70.00	18.92	4.50
80.00	28.57	2.67

View the table showing the numerical values implied by the figures in the graph, per 1,000 individuals screened. (see, e.g. Section 2.3)

All values are calculated based on the risk stratification profile entered, and assuming that the policy continued to use the best cutoff as determined by the current prevalence assumption.

Note that were the prevalence rate to be different than the current assumption, different cutoff might be implied.

For details, see: Stein, R. M (2020), "Where to draw the line for risk stratifications: Designing return-to-work policies that consider diagnostic error, costs, benefits and COVID-19."