Quasi-experimental designs for learning systems

Valerie Odeh-Couvertier[1], Gabriel Zayas-Cabán[1,3], Juan Camilo David Gomez[1], Kenneth Nieser[2], Brian Patterson[3], and Amy Cochran[2,4]

[1]Department of Industrial & Systems Engineering
[2]Department of Population Health Sciences
[3]BerbeeWalsh Department of Emergency Medicine
[4]Department of Mathematics

Abstract

The integration of risk prediction models with targeted interventions are playing a pivotal role in shaping how many service systems, such as healthcare systems, operate. In these and other settings, risk prediction algorithms may prevent bad outcomes by identifying individuals or processes that may benefit from specific interventions. However, the effectiveness of this integration is often poorly investigated, limiting its practical application. A main hurdle is that traditional research designs, like randomized controlled trials, face substantial challenges when evaluating interventions based on risk prediction models, making quasi-experimental designs a valuable alternative.

We propose a causal inference framework, based on quasi-experimental designs known as regression discontinuity (RD) designs, that is specially crafted for systems that use risk prediction algorithms to intervene. The proposed framework allows for the identification and estimation of the local average treatment effect (LATE) while addressing interference – a core violation of standard RD designs - that occurs because past interventions influence future ones. We identify and propose an estimator for the LATE, which we then apply to data from the 2017-2018 cycle of the National Health and Nutrition Examination Survey (NHANES), targeting the 8-year risk of diabetes incidence in middle-aged adults. We simulated a wide range of scenarios, including the recalibration and revision of the risk prediction model, adjustment of treatment assignment thresholds, and the examination of various clinical outcomes, ensuring comprehensive evaluation of the estimator's robustness and applicability. We find that our framework can effectively estimate the LATE under different simulated scenarios with low bias and variance, minimal mean square error, and high coverage. In short, the proposed framework aims to empower systems to adapt risk prediction models and their implementation based on data collected by the system, streamlining the evaluation of these models in real-world practice.

1 Introduction

While the past three decades have seen a nearly four-fold increase in publications on risk prediction models, very few are evaluated on their clinical benefit [1, 2]. Yet as Atkins *et al* note, the usefulness of risk prediction lies in its pairing with evidence-based interventions that can improve outcomes [2]. Governing bodies are actively working on standards for translating risk prediction models into clinical care, and their recommendations are likely to become more influential as the use of these machine learning models become widespread. In pursuit of a solution, this paper proposes a quasi-experimental designs that can independently evaluate the effectiveness of pairing a risk prediction model with a well-defined intervention in a modern healthcare system. This study is motivated directly from our collaborative work with two recent examples serving as compelling illustrations for the necessity of the proposed methods.

The first is on preventing at-home falls for older adults. We are collaborating with a research team whose focus is on an electronic tool that assists clinical decisions in the Emergency Department (ED) [3]. The tool utilizes a machine learning algorithm developed by the team to identify older adults in the ED who are at risk of experiencing a fall at home [4]. This algorithm has been incorporated into the electronic health software used in the ED of a large tertiary teaching hospital in the Midwest, enabling automatic flagging of at-risk patients based on their electronic health records. ED providers are notified of flagged patients and are encouraged to refer them to a specialized outpatient clinic for preventing at-home falls as part of their workflow.

Their goal is to evaluate the effectiveness of their electronic tool [5], but they face several challenges in doing so. First, conducting a randomized trial is not feasible due to ethical and implementation concerns. As a result, it is difficult to directly compare flagged and non-flagged individuals. Second, the research team has made updates to the algorithm in order to improve its face validity from the perspective of ED providers. Third, the risk cutoff at which patients are flagged has been lowered in response to few patients being referred to the outpatient clinic. Fourth, ED providers have undergone additional training to ensure they follow the recommendation of referring flagged patients to the outpatient clinic, in an another effort to increase referrals. Fifth, the team plans to implement the tool in a different healthcare system, which may require further updates to the algorithm, risk cutoff, and provider compliance. It is worth noting that these updates are occurring at unscheduled and unpredictable times. All of these changes – to the algorithm, to the threshold for intervening, and to provider compliance — demonstrate the need for frameworks that can rigorously evaluate the impact of intervening in a modern health system according to risk calculator.

The second example is on screening for social determinants of health. We are collaborating with a working group at a large tertiary teaching hospital that has developed cutting-edge risk calculators that have been integrated into clinical care and made publicly available on the HIPx-Change [6]. As part of an affordable care organization, this health system has recently introduced a comprehensive intervention to improve the health outcomes of individuals who have unfavorable social determinants of health. However, with around a million individuals in the patient population, it is impossible to screen everyone for possible benefit from the intervention. To address this challenge, the working group is developing a risk calculator to identify those at high risk for poor social determinants of health. By targeting the high-risk population, they can more feasibly screen individuals in-person for the intervention. As the team works towards implementing this tool, they expect to face similar implementation challenges as as the ones highlighted in the first motivating example above. Thus, the team will need to evaluate the effectiveness of their screening tool in a manner that allows for updates to the risk calculator and risk cutoffs and for streamlining the process from screening to intervention uptake.

These are but a few examples of this burgeoning practice. Researchers around the world have developed risk prediction models that are now primed for real-world implementation, but they need flexibility to adapt to issues that may arise. The present study aims to provide the necessary support to ensure these models are optimized for success. However, to be widely applicable, such a framework should strike a balance between upholding the scientific rigor of a randomized control trial and meeting the requirements of modern healthcare systems. Healthcare systems recognize that randomized controlled trials can face significant ethical and implementation challenges when evaluating interventions based on risk prediction models. This is because modern healthcare systems seek to *rapidly learn* from their own internal data. Therefore, evaluations should be able to handle the rapid adaptation that ensues. Quasi-experimental designs can provide nearly the same level of evidence as a randomized controlled trial without facing the same barriers.

Quasi-experimental designs provide a natural starting point to think of how to assess the impact of intervening on a patient based on a risk prediction model. A quasi-experimental design is an alternative to the experimental design, or randomized control trial, frequently with fewer ethical, regulatory, cost, and implementation barriers. By definition, these designs strive to find settings that can provide the same level of causal evidence of an experimental study without actually randomizing the intervention assignment. We consider the specific quasi-experimental design known as the regression discontinuity (RD) design [7], focusing on settings when patients are intervened upon when their predicted risk exceeds a threshold. A RD design would then take advantage of the resemblance of the intervention assignment to a randomized assignment for those patients whose predictions are near the threshold. Indeed, a RD design has been used to evaluate intervening on older adults in New Zealand according to a risk prediction [8]. However, quasi-experimental designs like RD are not optimized to accommodate the rapid adaptation to internal data.

As an illustration, suppose a cardiovascular risk prediction model has been published in the literature. In order to assess the effectiveness of prescribing statins to patients whose calculated risk exceeds a certain threshold, a healthcare system may wish to utilize electronic health record (EHR) software. This software could be customized to automatically identify at-risk patients and recommend to the care team that a statin be prescribed. The quasi-experimental design known as a regression discontinuity (RD) design could then be used to estimate the local average treatment effect (LATE), contrasting the observed risk of patients just above the threshold versus those just below the threshold. However, using the standard RD design would prohibit the intervention or risk prediction model from adapting to accommodate new internal data. This adaptation would violate the *no interference* assumption, which says that one person's intervention should not affect another person's potential outcomes.

By contrast, our collaborating research team has found adaptation to be indispensable for integrating a risk prediction model for at-home falls into Emergency Department (ED) care.[4, 3] Based on their experiences, they have found adaptation useful to re-calibrate algorithms (e.g., to generalize to new populations); modify alert triggers (e.g., to target patients with modifiable risk); and re-train care teams (e.g., to improve translation of alerts into recommended interventions). Moreover, their efforts have occurred at unscheduled and unpredictable times.

This paper offers quasi-experimental designs that permits adaptation, within the unifying causal inference framework of Richardson and Robins [9]. We do so by introducing a new RD design and proposing accompanying estimators of the LATE that allow for adaptation. We consider an analogue of intent-to-treat. In particular, we establish new RD estimators for the LATE of *recommending* a well-defined intervention according to a risk prediction model. To do this, we first characterize interference in this setting and provide sufficient conditions for identification. We then introduce estimators and evaluate their properties, such as robustness and bias, through a simulation study using a stratified sample from the 2017-2018 NHANES dataset, focusing on 1,592

middle-aged, non-diabetic adults to assess the impact of interventions based on a diabetes risk prediction model.

Organization of the paper. The remainder of this paper is organized as follows. First, we provide an overview in section 2 on research designs for causal inference, with an emphasis on quasi-experimental designs (e.g., RD design) and interference. In section 3 we conceptualize the ways in which continued learning of risk prediction can occur within a learning system. To motivate our work, we consider a risk prediction model developed at the University of Wisconsin (UW) Health to flag patients who are at risk of a fall after leaving the Emergency Department. Then, we formalize the problem of interest within the language of causal inference. In section 4 and 5, we present strategies based on the RD design for identifying and estimating the average treatment effect, respectively. We next provide details of our simulation study in section 6 and their results in section 7. Broader implications and concluding remarks on the subject matter are given in section 8.

2 Literature Review

This study represents a synthesis of three areas of the academic literature. The first area is the use of risk prediction to inform clinical care, with an emphasis on how a risk prediction model uses prior information, such as historical patient data or a prior risk prediction model, to update their prediction model for future prediction. This work serves to highlight the types of learning we are interested in evaluating. The second area is the learning system, in which we describe the current vision for how risk prediction models could be integrated and updated within a learning system. The third area is RD design, and extensions thereof upon which this study builds, such as RD designs with multiple risk thresholds and RD designs with interference between units. For each area, we highlight the existing gaps that our approach aims to fill.

Clinical risk prediction

According to the Institute of Medicine, a learning system is designed to "*generate* and *apply* the best evidence for the collaborative healthcare choices of each patient and provider" [10]. Clinical risk prediction is part of the vision for how a learning system "generates" evidence [11, 12]. Consider the following historical example of clinical risk prediction. The Framingham Heart Study was a long-term, population-based study that shaped how we understand and treat heart disease. A key outgrowth of this study were risk prediction models for predicting whether a patient would develop heart disease. For example, the risk of hard coronary heart disease can be calculated from a risk prediction model for non-diabetic individuals aged 30 to 79. Nowadays, a patient or care provider can enter a person's age, cholesterol, blood pressure, smoking status, and blood pressure mediation into the model, which then returns the patient's likelihood of developing hard coronary heart disease [13, 14, 15]. Clinical risk prediction have since become an important endeavor for identifying patients at risk for other adverse outcomes, such as falls [4], suicide [16], and hospital readmission [17], to name a few.

A critical first step towards clinical risk prediction is the validation of risk prediction models [11, 12]. Validation involves showing a risk prediction model is accurate and interpretable, provides meaningful clinical information, and can generalize beyond the data used to build the model [18, 19, 2]. During validation, however, the dataset is usually *fixed*, i.e. unchanging over time. This has recognized downsides, which a learning system is meant to overcome. First, data is not collected in a fixed way, but rather dynamically over several years (or even decades in the case of the Framingham Heart Study) and in waves and cohorts. Second, the accuracy of predictions can deteriorate over time [20, 11, 21], due to changes in healthcare practices, variations in data quality, and the emergence of unmeasured risk factors that affect treatment outcomes [22]. Third,

a risk prediction model may not generalize well when moving from data collected in one setting or population to another [21].

Thus to remain relevant over time and in new settings or populations, emerging work has outlined ways a risk prediction model can be updated. These include recalibration (adjusting a risk model for new data setting by updating some or all coefficients) [23], revision (re-estimation of some or all coefficients) [24], and extension (adjusting a risk model with the addition of a new risk markers) [20]. More recent ideas include meta-analytical approaches and dynamic updating [25, 26]. Meta-analytical approaches are used in situations where multiple clinical risk models have been developed for the same or similar outcomes and populations [20, 27]. The idea is to combine models into one single "meta-model" that is expected to have better predictive performance and generalizability than a single model. Meanwhile, dynamic approaches focus on constantly updating models at periodic or continuous time points. Traditional methods such as recalibration and revision may be used for periodic updating [20], while Bayesian dynamic modeling and dynamic model averaging are suggested for continuous modeling updating [27].

Learning systems

While clinical risk prediction, and the strategies outlined above, are useful for determining how a learning system can "generate" evidence, it provides limited information on how to "apply" this evidence. Yet, applying evidence is essential to our definition of learning:

> *Learning*, as in learning system, is the phenomenon by which a system or organization adapts its knowledge or behavior, usually more than once, according to the history of observed events, thereby altering the likelihood of future events.

Put differently, the learning considered in the present paper requires the adaptation of the system, which distinguishes it from the learning considered in machine learning or statistical learning. From this perspective, a learning system is *learning* if it uses prior data to enhance how future patients are cared for [28, 29, 30], and clinical risk prediction can contribute to the learning of a learning system if the models can both account for prior data *and* inform the care of future patients.

The importance of learning, in the context of the definition above, is echoed by Atkins and co-authors: "Risk prediction is only useful to the extent that it is paired with evidence-based interventions that can improve outcome" [2]. Indeed, the past three decades have seen a nearly four-fold increase in publications on risk prediction models; yet, very few models demonstrate clear clinical benefit [1, 2]. In general terms, just because a risk prediction model is valid does not necessarily translate into an intervention that is effective [31]. For example, health care utilization may be difficult to modify among individuals at highest risk for hospitalization, which includes, for instance, individuals with terminal illnesses [32]. Alternatively, there may be issues with intervention design or implementation, competing organizational priorities, staff reluctance to an intervention, and resource constraints [33, 34]. Even worse, an intervention paired with clinical risk prediction might result in worse outcomes or higher costs, making it an imperative to demonstrate the benefits outweigh the risks. As a result, few health systems have achieved learning, even though the necessary steps are often in place: risk predictions models could inform current clinical care, and risk prediction models could be updated to account for prior data.

The pressing question has become how to — ethically, rapidly, and rigorously — evaluate the consequences of the learning of a health system. Larger health systems are experimenting with answers to this question. At NYU Langone Health, patients were identified with unmet medical care needs (e.g., vaccines, yearly checkups), and interventions were developed for these patients (e.g., mailed remainders, telephone follow-up) [33]. Notably, intervention delivery was then randomized to evaluate the effectiveness of their interventions in improving care needs and compliance. The

Veterans Health Administration experimented with pushing the output of risk prediction models to clinician alerts and dashboards to guide care for individuals at risk for adverse events such as suicide, opioid use, and high healthcare utilization [2]. Like NYU Langone Health, they also explored research designs that involved randomization (in this case, step-wedged designs). Another example can be found in [35]. Here, the authors used risk prediction models targeting common newborn disorders, such as sepsis, neonatal encephalopathy, respiratory distress, and thermoregulation, to assist with postnatal care in low resource settings. Initial implementation of this learning system indicated that it was well-received and deemed highly usable by healthcare professionals.

The limited examples above highlight that, on one hand, experimental research designs, in which intervention delivery is randomized, can help evaluate the consequences of integrating clinical risk models into a learning system [33, 2]. This is not a surprise, since these designs have been the standard for evaluating effectiveness of clinical interventions. Yet, at the same time, these designs are prohibitively slow and costly, diminishing the potential for rapid response of learning systems. On the other hand, a non-experimental research design, or observational study, can provide more timely evidence, but at the expense of misattributing differences in outcomes to an intervention. In the case of clinical risk prediction, for example, individuals predicted to be at high risk may trend towards better outcomes over time, regardless of intervention, a phenomenon referred to as "reverting to the mean". Time and again, researchers have emphasized the need for specific tools, guidelines, and frameworks that are solely dedicated to evaluating effectiveness of clinical risk prediction models within a learning system [2, 36, 37].

Regression discontinuity design

A compromise between an experimental research design and an observational study is the *quasi-experimental* design. These are observational studies meant to resemble an experimental design. One such quasi-experimental design is the regression discontinuity (RD) design, first introduced by Thistlewaite and Campbell [38]. A RD design captures the following scenario: units have a score and are assigned to an intervention according to a function of the score. For example, the original paper by Thistlewaite and Campbell considered students (units) who received a Certificate of Merit (intervention) if they received performed sufficiently well on a test (score). This scenario can be mapped onto our context if patients (units) are assigned to different intervention groups (intervention) according to their predicted risk (score) of experiencing some adverse event. A RD design resembles an experimental design for units with a score near the cut-off, in that, prior to intervention delivery, units with a score right above the cut-off should be similar to units right below the cut-off.

Hahn [7] formalized a RD design within the potential outcomes framework of Neyman and Rubin [39, 40]. Each unit is associated with two potential outcomes capturing the hypothetical scenarios *were* a unit assigned to the intervention and *were* they not assigned to the intervention. The goal is to estimate the average difference in these potential outcomes conditional on the score being equal to the cut-off. This quantity is called the local average treatment effect (LATE). Hahn, Todd, and Van der Klaauw showed that the LATE can be re-written equivalently in terms of observed variables (i.e. identified) under assumptions about the continuity of mean potential outcomes and the discontinuity of mean intervention assignment around the cut-off [7]. RD designs can also be thought of as experiments that are randomized locally, with units near a cutoff score effectively randomized to intervention groups, and continuity assumptions replaced with alternative assumptions that are easier to evaluate.[41, 42] This framework also allows for randomization inference, which uses simulated intervention assignments to construct confidence intervals and perform hypothesis tests, and can be coupled with Rosenbaum sensitivity analyses to systematically explore assumption violations [43, 44].

Robinson and co-authors have explored a RD design in two different settings involving clinical risk prediction [8, 45]. In [8], the risk of cardiovascular disease was predicted for primary care patients in New Zealand using a Framingham risk prediction model modified to be relevant to New Zealand [46]. They tested for discontinuity in the probability of statin dispensation for predicted risks of 15% and 20%, representing recommended values for guiding treatment. Although unsuccessful, if they found discontinuity, then a RD design could have been used to evaluate the impact of statin dispensation on downstream outcomes. In [45], the risk of readmissions was predicted for older adults (65+ years) who had an acute hospitalization in Auckland, New Zealand [46]. Persons whose predicted risk exceeded 20% were enrolled into a special program called the Integrated Transition of Care, implemented to reduce hospital readmissions. Using a RD design, they did found that the LATE of enrollment into the program on the probability of readmission did not differ significantly from zero.

While these two examples illustrate how an RD design can evaluate consequences of clinical risk prediction, the risk prediction model was fixed throughout the study period [8, 45]. Therefore, the risk prediction model was not updating and the system was not learning from prior patients. To illustrate, determining the appropriate cutoff or threshold for risk interventions is a crucial aspect of the learning and adaptive process that this work seeks to address. If the risk threshold is set too high, only a small number of individuals may be targeted, which may not have a significant impact on the population outcomes. On the other hand, setting the threshold too low can put immense pressure on finite hospital staff and resources that support the intervention, which may ultimately undermine its effectiveness. RD designs have also been extended to handle multiple cutoffs [47, 48, 49]. One potential solution is to normalize the cutoffs to zero and estimate a single pooled effect, which Cattaneo and colleagues have formally analyzed [47]. At first glance, this appears relevant because, in practice, healthcare systems may change the risk cutoff used to determine intervention groups based on internal data. However, this approach assumes cutoffs are independent of the data generation process, meaning it does not solve the problem we are currently facing.

One violation called *interference* is at the heart of this work. This happens when one person's intervention can affect another's potential outcomes, which goes against Rubin's stable unit treatment value assumption (SUTVA) [50]. We want to allow for interference, as we think a patient's experiences can inform future patients' care. Rosenbaum proposed a way to handle arbitrary interference using randomization inference strategies [51], which were later adopted into an RD design.[44] These ideas were then expanded upon by Aronow and colleagues [52], who reinterpreted the typical estimator as the local average direct effect of Hudgens and Halloran [53]. This latter effect is a function of the assignment mechanism, but does permit arbitrary interference.

Although the methods proposed in those studies are similar to ours, none of them explicitly consider an RD design that allows for updates to the risk prediction model. As we will see, we do not need to consider arbitrary interference, which enables us to propose more effective estimators.

3 Research design

Setting

We propose a research design to evaluate the impact of intervening on individuals (e.g., patients) based on a clinical risk prediction model in a system (e.g., health system such as an ED). The design involves a cohort of patients with common characteristics and/or settings, such as older adult patients presenting to the ED with chest pain chief complaints, who interact with the system. The system collects data on each patient, including demographic information, clinical history, and presentation, which is used as input for a clinical risk prediction model. If the predicted risk of an outcome relevant to the cohort exceeds a certain threshold, the system triggers a well-defined

intervention, such as a referral to a falls clinic. The patient is then monitored by the system for a subsequent outcome, such as a hospitalization or completed referral. Patient and system perspectives of the proposed design are illustrated in Figure 1.

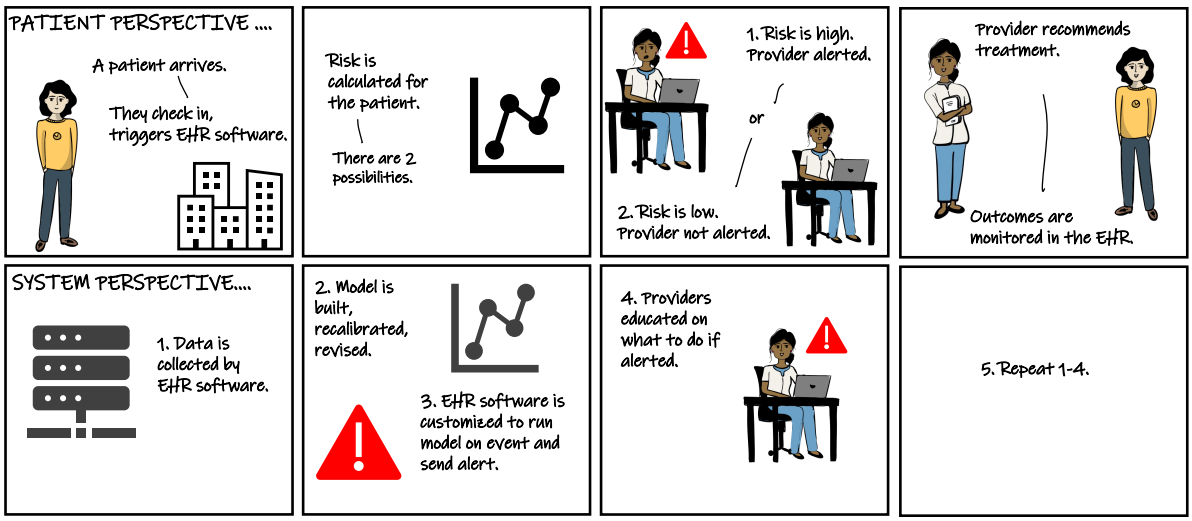

Fig 1. Patient and system perspectives of proposed RD design

Critically, two things can be repeatedly updated by the system: the risk prediction model or the threshold. In either case, we impose no restrictions on when these updates occur, nor do we consider how the update might occur. This latter flexibility ensures that our design can accommodate any strategy for updating the model and threshold. It goes without saying that only data collected by the time an update occurs can be used for the update. The consequence of this is that patients can be ordered according to which risk prediction model and threshold is being used, so that earlier patients use early instances of the risk prediction model and threshold. This ensures that any patient is using a risk prediction model and threshold that depends on data from only those patients earlier in the sequence.

Notation

We use the following notation:

- Capital letters for random variables (e.g., X).
- Greek letters for parameters (e.g., $\theta, \xi, \eta, \alpha, \beta$).
- Lower case letters for observations of a random variable (e.g., $x \in \mathcal{Z}^{(X)}$) where $\mathcal{Z}^{(C)}$ for the target space of a random variable X)
- Subscripts for vectors or sequences, e.g., $X_{1:k}$ is $(X_1, \ldots, X_k)$ and X_B is $(X_{i_1}, \ldots, X_{i_k})$ when $B = \{i_1, \ldots, i_k\} \subseteq \mathbb{N}$.
- Apostrophe for transpose, e.g. Z' is the transpose of Z.
- $\|Z\|$ for the L_2-norm of Z.

In addition, we adopt the convention that $X_{1:k}$ and $\mathcal{Z}^{X_{1:k}}$ are empty when $k = 0$ and that $\mathcal{Z}^{(X_{1:2})}$ represents the product space $\mathcal{Z}^{(X_1)} \times \mathcal{Z}^{(X_2)}$ associated with $Z_{1:2}$ (similarly for sequences).

Variables

The research design is formalized as follows. Let i index a patient within a learning system with $i = 1, \ldots, n$. Patient i is associated with random variables:

- $X_i \in \mathcal{X}$ is a vector of patient characteristics (model input);
- $R_i \in \mathbb{R}$ is the predicted risk (model outcome);
- $A_i \in \{0, 1\}$ is the assignment to intervention (immediate consequence of model outcome);
- $Y_i \in \mathbb{R}$ is the observed outcome of interest (downstream consequence of model outcome).

Several assumptions are made on these variables. Patients characteristics X_i is taken to be the input into the risk prediction model used for person i, and R_i is defined to be the predicted risk that is outputted from said model. Intervention assignment is determined based on whether the predicted risk R_i exceeds some patient-specific threshold C_i: patients with $R_i \geq C_i$ are assigned to receive the intervention, and patients with $R_i < C_i$ are assigned to not receive the intervention. This leads to $A_i = \mathbb{I}(R_i \geq C_i)$. Since R_i is sufficiently general that we could always redefine R_i to be $R_i - C_i$, it will be without any loss of generality that we assume $C_i = 0$ for all i.

Finally, we assume an ordering on patients such that the predicted risk R_i can only be informed by the random variables of patients with a smaller index $j < i$ and X_i. Formally, we take this latter assumption to mean that the σ-algebra generated by $\{X_k, R_k, A_k, Y_k\}_{k=1}^{j}$ is a subset of the σ-algebra generated by $\{X_k, R_k, A_k, Y_k\}_{k=1}^{i}$ whenever $j < i$. Relative to person i, we refer patients with a smaller index as "prior" patients and patients with a larger index as "future" patients, though strictly speaking it is not necessary that patients arrive to the system in the specified order. In addition, upon conditioning $\{X_k, R_k, A_k, Y_k\}_{k=1}^{i-1}$, we assume the risk prediction model for person i is a deterministic function, which allows us to view R_i as a function from $\mathcal{X}$ to $\mathbb{R}$. This last observation will often be made explicit with the notation $R_i(X_i) := R_i$.

Along with the original random variables, potential outcomes will be defined following the generalized po-calculus framework described by Richardson and Robins [9] and later by Malinsky, Shpitser, and Richardson [54]. To simplify notation, let x_B for $B = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ and $i_1 \leq \ldots \leq i_k$ be the shortened notation for the vector $(x_{i_1}, \ldots, x_{i_k})$. Then in broad terms, random variables

$$X_i(a_{1:n}), A_i(a_{1:n}), R_i(a_{1:n}), Y_i(a_{1:n})$$

are introduced to model the random variables associated with the hypothetical scenario were we to assign patient j to intervention $a_j \in \{0, 1\}$ for $j = 1, \ldots, n$. Put differently, each vector of hypothetical assignments $a_{1:n}$ will be mapped to random variables denoted above and respectively taking values in the target space of X_i, R_i, A_i, and Y_i. In the case of no interference, the only assignment that would influence these variables for person i would be person i's hypothetical assignment a_i. As such, we could drop the irrelevant arguments to simplify notation to $X_i(a_i)$, $A_i(a_i)$, $R_i(a_i)$, $Y_i(a_i)$. In the case of full interference, i.e. every assignment could potentially influence every random variable, then we could not drop any argument. Our problem will be shown to be somewhere in between these two cases.

Causal model

Potential outcomes are defined more formally upon assuming the following structural causal model for the original random variables X_i, R_i, A_i, Y_i:

Assumption 1 (Structural causal model)**.** *There exists independent random variables* $\{N_{x_i}, N_{y_i}\}_{i=1}^{n}$ *with* $\{N_{x_i}\}_{i=1}^{n}$ *identically distributed and similarly for* $\{N_{y_i}\}_{i=1}^{n}$*. Further, there exists deterministic*

assignment functions $\{f_{r_i}\}_{i=1}^n$ *and* f_y *so that the following random variables are assigned iteratively from* $i = 1$ *to* $i = n$ *according to:*

$$
\begin{aligned}
X_i &:= N_{x_i} \\
R_i &:= f_{r_i}(X_i, X_{1:i-1}, A_{1:i-1}, Y_{1:i-1}) \\
A_i &:= \mathbb{I}(R_i \geq 0) \\
Y_i &:= f_y(X_i, A_i, N_{y_i})
\end{aligned}
$$

It is important to recognize what is implied by the assumed structural causal model. The first is that the X_i are independent and identically-distributed (iid). The second is that each R_i is a deterministic function of the characteristics X_i of person i and all variables associated with persons j whose $j < i$, which by construction are exactly those persons whose data could contribute to the current risk prediction model used for person i. A person-specific assignment f_{r_i} is sufficiently general to allow for any possible way in which the model could be updated based on data from persons $j < i$. Consequently, the R_i are generally not iid. The third is that intervention assignment A_i is a deterministic function of R_i, taking a value of 1 or 0 depending on whether R_i exceeds zero. As such, R_i is not needed in any subsequent assignments. With R_i not iid, neither are A_i. The last is that the outcome Y_i for person i depends, up to some independent noise term N_{Y_i}, only on the intervention assignment A_i and characteristics X_i of person i. This, together with assuming the same assignment f_y for each person, means that Y_i conditional on $A_{1:n}$ are iid. However, the unconditional Y_i are generally not iid.

We also remark on what scenarios are excluded by the assumed structural causal model. We are not allowing the distribution of the X_i to change between persons. This might occur were there changes to how X_i is measured, e.g., a more precise way to measure blood pressure, or changes to the sample population over time, e.g., an aging population. In addition, we are not allowing the outcome distribution of Y_i conditional on the intervention A_i to change between persons. This might occur were prior patients to inform changes to how the intervention is delivered.

The assumed structural causal model induces a directed acyclic graph (DAG). Nodes are the variables X_i, R_i, A_i, and Y_i, and a directed edge exists from one variable to another if and only if the latter is an argument in the assignment of the former. The DAG is depicted in the simple case $n = 3$ (Figure 2A). For example, the predicted risk R_3 for person 3 could be influenced by the data from persons 1 and 2: characteristics X_1 and X_2, intervention assignments A_1 and A_2, and outcomes of Y_1 and Y_2.

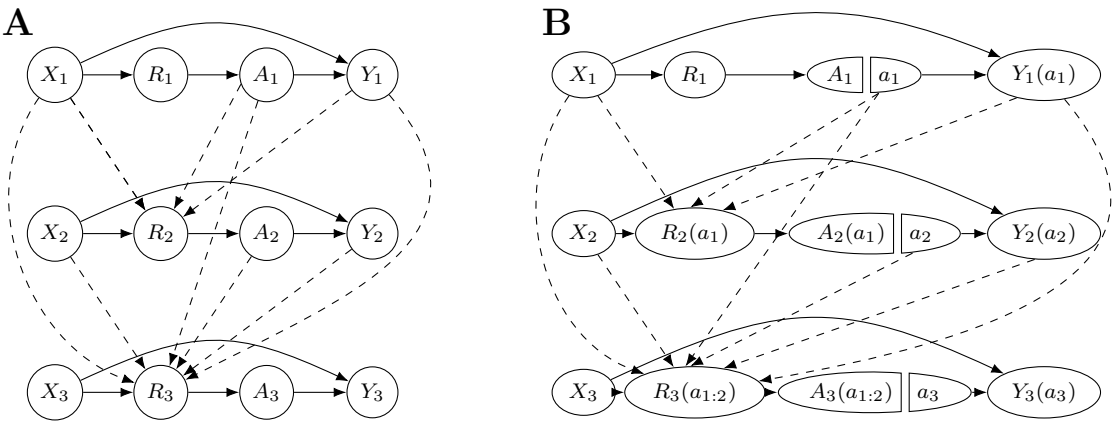

Fig 2. A) DAG induced by assumed causal model and **B)** corresponding SWIG when $n = 3$.

Potential outcomes

With the structural causal model, we can now define potential outcomes:

Definition 1 (Potential outcomes)**.** *Assume Assumption 1 holds. Let* $\{N_{x_i}, N_{y_i}\}_{i=1}^n$, $\{f_{r_i}\}_{i=1}^n$, *and* f_y *be the objects invoked by this assumption. For* $B \subseteq \{1, \ldots, n\}$ *and* $a_{1:n} \in \{0,1\}^n$, *define potential outcomes*

$$X_i(a_B), R_i(a_B), A_i(a_B), Y_i(a_B)$$

iteratively from $i = 1$ *to* $i = n$ *as follows:*

$$\begin{aligned}
X_i(a_B) &:= N_{x_i} \\
R_i(a_B) &:= f_{r_i}(X_i(a_B), X_{1:i-1}(a_B), \bar{A}_{1:i-1}(a_B), Y_{1:i-1}(a_B)) \\
A_i(a_B) &:= \mathbb{I}(R_i(a_B) \geq 0) \\
Y_i(a_B) &:= f_y(X_i(a_B), \bar{A}_i(a_B), N_{y_i})
\end{aligned}$$

where, for ease of notation, we introduce

$$\bar{A}_i(a_B) := \begin{cases} a_i & i \in B \\ A_i(a_B) & i \notin B \end{cases}.$$

We set B equal to $\{1, \ldots, n\}$ to get the specific potential outcomes of interest. The above construction is compatible with Neyman-Rubin potential outcomes and is based on the idea of Richardson and Robins [9]. The intervened variables, which in this case are the intervention assignments A_B, have their corresponding potential outcome $A_B(a_B)$ constructed. Yet, where A_B would have been passed forward to construct subsequent variables in the structural causal model, the intervened values a_B are now passed forward. This construction benefits from an induced DAG for the potential outcomes and the intervened values a_B called the single-world intervention graph (SWIG)[9]. Alternatively, the SWIG can be constructed directly from the original DAG by splitting each intervened variable into two: the rightmost variable takes the corresponding intervened value and inherits the original node's outgoing edges; the leftmost variable inherits the original node's incoming edges. All random variables in the SWIG are expressed as potential outcomes. We do not observe any of the original variables in the SWIG. This obscures how realizations of the original variables could ever be used to learn about the potential outcomes. To remedy this, we note that the three rules of Pearl's do-calculus [55] are compatible with the three rules of po-calculus, as described by Malinksy, Schpitser, and Richardson [54]. Further, the third rule of po-calculus can be recast as a property called causal irrelevance. This property can be described concisely in the context of the SWIG: i.e. only descendants of the intervened values a_i are influenced by a_i. Alternatively, for those familiar with the graphical concept of d-separation, a descendent of any a_i in the SWIG are exactly those nodes that are not d-separated from a_i. For completeness, we re-state Rule 3 in the context of our notation and problem, but without SWIGs, as was presented in Malinksy, Schpitser, and Richardson [54]:

Theorem 1 (Malinsky et al. (2019), Rule 3)**.** *Fix disjoint* $B, C \subseteq \{1, \cdots, n\}$ *and* $a_{1:n} \in \{0,1\}^n$. *Take* Z *to be one of the variables* X_i, R_i, A_i, Y_i *and* $Z(a_{B\cup C})$, $Z(a_B)$, $Z(a_C)$, *to be that variable's corresponding potential outcomes. If every directed path from* A_i *with* $i \in \{B \cup C\}$ *to* Z *in the original DAG visits some* A_j *with* $j \in B$ *after visiting an* A_k *with* $k \in C$ *before reaching* Z, *then* $Z(a_{B\cup C}) = Z(a_B)$.

Rule 3 allows us to drop (causally) irrelevant arguments in the potential outcomes. For $n = 3$, the resulting SWIG after dropping irrelevant arguments through Rule 3 is shown in Figure 2B. This SWIG captures both realized variables and potential outcomes and is able to more precisely characterize interference in our model. For example, intervention assignment of the previous $i-1$ patients is irrelevant to the potential outcome of patient i. We formalize these observations in Proposition 1.

Proposition 1 (Causal Irrelevance)**.** *Assume Assumption 1 holds. The following equalities hold:*

$$\begin{aligned} X_i(a_{1:n}) &= X_i \\ R_i(a_{1:n}) &= R_i(a_{1:i-1}) \\ A_i(a_{1:n}) &= A_i(a_{1:i-1}) \\ Y_i(a_{1:n}) &= Y_i(a_i). \end{aligned}$$

In addition, potential outcomes $Y_i(a_i)$ are iid for $i = 1, \ldots, n$.

Proposition 1 is fundamental to subsequent results. In particular, patient assignments do not interfere with the potential outcomes $Y_i(a_{1:n}) = Y_i(a_i)$ of successive patients. Patient assignments do, however, interfere with the potential outcomes associated with the risk predictions and with the intervention assignments, but in a specific way. Only assignments of prior patients ($j < i$) interfere with $R_i(a_{1:n}) = R_i(a_{1:i-1})$ and $A_i(a_{1:n}) = A_i(a_{1:i-1})$. The proof for Proposition 1 is presented in Appendix A.

Local average treatment effect

Our characterization of interference, in particular the observation that $Y_i(a_{1:n}) = Y_i(a_i)$, allows us to define a LATE that is similar to what is used in a traditional RD designs:

Definition 2 (Local average treatment effect)**.** *Assume Assumption 1 holds. Let β_i denote the local average treatment effect for person i with predicted risk $R_i = 0$:*

$$\beta_i := \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| R_i = 0\right].$$

The main distinction between the causal effect above and the traditional LATE is that the R_i are not iid. Consequently, it is generally the case that $\beta_i \neq \beta_j$ whenever $i \neq j$, meaning we have a LATE that is specific to each person. Later, we will see how this observation affects the use of data from persons $i \neq j$ to estimate β_i.

4 Identification

Non-parametric identification is the process by which we re-express the causal effect — equivalently — in terms of the distribution for the original random variables (X_j, R_i, A_j, Y_j) $(i = 1, \ldots, n)$. Given that data is collected on the original random variables (X_j, R_i, A_j, Y_j), and not the potential outcomes, this step is critical if we are to use the data to estimate β_i. Here, we derive conditions that allow us to identify β_i.

Consistency

We start with one of the familiar assumptions of causal inference, called *consistency*. In words, consistency refers to the assumption that the observed outcomes Y_i are exactly the potential outcomes $Y_i(a)$ were the hypothetical intervention assignment be the actually realized intervention assignments, i.e. $A_i = a$. What it means to be the potential outcome under the realized assignment, often denoted by $Y_i(A_i)$ and considered to be a random variable, is often left vague. Confusion can arise since the notation $Y_i(A_i)$ makes it seem like it is a composition of functions. This is not

the case. The variable A_i is a function from the underlying probability space to $\{0,1\}$, whereas the mapping $a_i \mapsto Y_i(a_i)$ is a function from $\{0,1\}$ to a random variable. The composition of these two functions would not be random variable as desired, but rather a function from the underlying probability space to a random variable. We get around this issue by defining the random variable $Y_i(A_i)$ sample-wise:

Definition 3 (Potential outcomes under realized assignments)**.** *Assume Assumptions 1 holds. Let Ω denote the probability space associated with the invoked random variables $\{N_{x_i}, N_{y_i}\}_{i=1}^n$. Fix $B \subseteq \{1, \cdots, n\}$. Take Z to be one of the variables X_i, R_i, A_i, Y_i and $Z(a_B)$ to be that variable's corresponding potential outcomes for any $a_{1:n} \in \{0,1\}^n$. Define $Z(A_B)$ sample-wise to be the random variable that returns, for each $\omega \in \Omega$:*

$$Z(A_B)[\omega] := Z(a_B)[\omega]$$

whenever $A_B[\omega] = a_B$.

The importance of this definition is that consistency is now an immediate consequence, *and no longer an assumption.* We also point out that, in addition to the last definition, the reason that consistency is not an assumption is because it is embedded in the assumed structural causal model (Assumption 1) and in the subsequent construction of the potential outcomes according to the framework of Richardson and Robins [9]. This connection is made apparent in Proposition 2 (see proof of consistency in Appendix B).

Proposition 2 (Consistency)**.** *Assume Assumptions 1 holds. Take Z to be one of the variables X_i, R_i, A_i, Y_i and $Z(A_{1:n})$ to be that variable's corresponding potential outcomes under the realized intervention. Then, consistency holds, i.e.,*

$$Z(A_{1:n})[\omega] = Z[\omega]$$

for all $\omega \in \Omega$.

Local randomization

With consistency, and one last assumption, we can identify causal effects β_i. We can do this in two ways, depending on the last assumption we want to make. On way is to lean on the continuity assumption of potential outcomes as stated in Hahn, Todd, and Van der Klaauw [7]:

Assumption 2 (Continuity of potential outcomes)**.** *Assume the function*

$$r \mapsto \mathbb{E}[Y_i(a) | R_i = r]$$

is continuous at $r = 0$.

This assumption is perhaps easy to understand mathematically, but difficult to reason about in practice. An alternative way is to lean on an assumption akin to an assumption about local randomization [41]:

Assumption 3 (Local randomization)**.** *Patient characteristics X_i can be decomposed like*

$$U_i := g(X_i)$$

for some deterministic g such that $Y_i(a_i)$ is independent from X_i conditional on U_i and the conditional cdf $P(R_i \leq r \,|\, U_i = u)$ is conditionally differentiable at $r = 0$ and satisfies:

$$0 < P(R_i \leq 0 \,|\, U_i = u) < 1$$

for every u in the image of U_i.

This local randomization assumption has several pieces, making it perhaps more difficult to understand mathematically than the continuity of potential outcomes assumption (Assumption 2). Yet, once one can wrap their head around this assumption, local randomization is perhaps easier to reason about in practice.

This reasoning goes like this. In addition to X_i, every person is described by U_i. Any information in X_i about potential outcomes $Y_i(a_i)$ is captured entirely by U_i; yet knowledge of U_i never perfectly predicts the risk R_i. Consider, for example, cardiovascular disease. A patient X_i might be described by observables like age and noisy measurements of blood pressure and unobservables like biological age and true blood pressure. The risk prediction R_i is never be fully determined by the unobservables. Yet, these unobserved variables could contain all the information in X_i needed to best predict the potential $Y(a_i)$ for cardiovascular disease. This assumption allows us to view individuals as being locally randomized near the cutoff. Under either assumption, we conjecture that we can identify the LATE, viz,

Theorem 2. *Under the assumptions of either continuity (Assumption 2) or local randomization (Assumption 3), and considering R_i as the prediction model for individual i that maps $x \in \mathcal{X}$ to $\mathbb{R}$, we have*

$$\beta_i = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| R_i = 0\right] = \lim_{r \to 0} \left(\mathbb{E}\left[Y_i \middle| A_i = 1, X_i \in W_{r,i}\right] - \mathbb{E}\left[Y_i \middle| A_i = 0, X_i \in W_{r,i}\right]\right)$$

where $W_{r,i} = \{x \in \mathcal{X} : \exists\, x' \in \mathcal{X}, \left\|x - x'\right\| < r, R_i(x') = 0\}$.

This conjecture differs from prior RD identification methods by capitalizing on our precise knowledge of the mapping of the risk scores. We can then redefine the concept of local randomization to incorporate the input variables used in risk calculation. This is reflected in $W_{r,i}$ (See lemma in Appendix C.1), which captures individuals whose characteristics, with slight variations, would yield a predicted risk of zero. In addition, identification is currently specific to each individual. However, in the estimation process, we aim to make use of the iid assumptions of the relevant variables. The proof for Theorem 2 is presented in Appendix C.2.

5 Estimation of average causal effects

Our estimation strategy builds on Theorem 2; namely, on the relationship between causal effects and observed variables. We propose estimators that approximates the limit of conditional expectations in Theorem 2, akin to estimators from traditional RD designs. To address interference, the proposed estimator leverages iid assumptions of the random variables so as to borrow information from prior patients to estimate β_i.

For a fixed patient i, the estimator uses a piecewise polynomial to model the expected values of the outcomes:

$$\mathbb{E}\left[Y_i \middle| A_i, R_i, X_i \in W_{r,i}\right] \approx \alpha_0 + (1 - A_i) f_0(R_i) + A_i(f_1(R_i) + \alpha_1) \tag{5.1}$$

for some polynomial functions $f_0(\cdot)$ and $f_1(\cdot)$ with $f_0(0) = f_1(0) = 0$. To estimate the above regression coefficients, we use data collected on patient characteristics, intervention assignments, and outcomes up to patient i:

- $x_1, \ldots, x_i$ (observed characteristics)
- $a_1, \ldots, a_i$ (observed assignments)
- $y_1, \ldots, y_i$ (observed outcomes)

We also use the current prediction model, denoted as R_i, to retroactively estimate the risk for prior patients:

- $\hat{r}_1, \ldots, \hat{r}_i$ (restroactive risk)

Next, we apply a bandwidth, denoted as r, to select only those patients whose characteristics yield a predicted risk close to zero if slightly varied. The indices of these selected patients are captured in the following set:

$$\mathcal{J} := \{j \in \{1, \ldots, i\} : x \in \mathcal{X} : \exists\, x' \in \mathcal{X}, \left\|x_j - x'\right\| < r, R_i(x') = 0\}.$$

Among these selected patients, we then regress observed outcomes y_j onto observed characteristics x_j, observed intervention assignment a_j, polynomial term $a_j f_1(\hat{r}_j)$, and polynomial term $(1 - a_j) f_0(\hat{r}_j)$ $(j = 1, \ldots, i)$. This results in an estimate of α_1.

The estimator proposed use two novel concepts. The first concept involves using the current prediction model R_i to retroactively generate risks $\hat{r}_j$ for previous patients. Our key insight is that although the risk predictions themselves are not generally iid, the retroactive risks can be treated as iid samples from the distribution of R_i. Then when incorporating the iid assumptions from our causal model (Assumption 1), we expect that our estimators approximate our causal effects. The second concept revolves around restricting the data based on the model input x_i rather than the predicted risk. While two individuals may have vastly different predicted risks, if their characteristics used for the prediction are similar, we consider them as exchangeable.

6 Simulation study: Effect of intervening according to a diabetes mellitus risk prediction model

Our numerical study affords us the opportunity to simulate a wide range of realistic scenarios for how risk prediction models may be used to intervene in a healthcare setting. The goal is not simply to capture these scenarios, but rather, to apply and thus evaluate our proposed framework, specifically the RD estimator (as referenced in 5.1), in each such scenario to estimate the effect of sequentially intervening on individuals or units based on risk predictions that may be dynamically adapted via sequentially updating a model or a model's risk threshold. In particular, we had three primary objectives: (i) simulate a range of realistic scenarios for how risk prediction models may be adaptively used to intervene (ii) applying the proposed RD estimator 5.1 for each scenario thereby identifying situations where the estimators may or may not work well; and (iii) determine effective strategies for using the proposed RD estimator to guide the adaptation of risk models. We will do this within the context of the 8-year incident diabetes mellitus risk prediction model, for middle-aged adults, based on the Framingham offspring study [56]. For clarity and reproducibility of our simulation study, we used a publicly available dataset from which to independently sample X_i $(i = 1, \ldots, n)$. Specifically, we used the 2017-2018 cycle of the National Health and Nutrition Examination Survey (NHANES) [57], a large ongoing survey that collects information on the health and nutritional status of the US household population through interviews, medical exams, and laboratory tests.

Simulation scenarios

We proposed and analyzed four ways in which healthcare systems such as EDs or outpatient clinics may implement and assess the use of a risk prediction model. The first is *model re-calibration*, which refers to when models are adjusted to accommodate new data or different populations. The second is *model revision*, which refers to when models are comprehensively re-estimated and/or predictors may be removed or added in response to changes in EHR instances. The third is *cutoff*

updating, which refers to when cutoffs are modified to, say, align with capacity constraints or specific target groups. Fourth is *alternative outcomes*, which refers to when outcomes differ from the predicted outcome of the risk prediction model. Each scenario is adjusted by parameters summarized in Table 1, which are described in detail below. We proposed five main simulation scenarios, where for each scenario a dynamic cutoff updating approach is implemented along with one other of the four aforementioned risk prediction model implementation and evaluations.

Simulation 1 Risk prediction model remains static, with no updates or learning.

Simulation 2 Risk prediction model is recalibrated with shrinkage term of 0.5.

Simulation 3 Risk prediction model is recalibrated with shrinkage term of 1.

Simulation 4 Risk prediction model is comprehensively revised with a shrinkage term of 0.5.

Simulation 5 Risk prediction model remains static, with alternative outcome.

Our first simulation scenario is our baseline scenario, wherein the risk prediction model remains static, with no recalibration or revision. The second and third scenarios simulated the recalibration of the risk prediction model. Such recalibrations may be important, given that empirical evidence suggests there may be a potential decline in predictive accuracy as more data is incorporated into the system over time [23]. The fourth and fifth simulations simulated the revision of the risk prediction model by refitting the model completely to re-estimate all regression coefficients. To mitigate the risks associated with overfitting in the revision of the model, a shrinkage factor of 0.5 was applied in the direction of the original regression coefficients. While the first four simulation scenarios are based on risk prediction models of a binary outcome, the fifth simulation scenario predicts changes in weight, a continuous variable.

Parameter	Scenario	Description
n	All	Sample size of simulated population
n_{batch}	All	Sample size before recalibration & revision
ζ_k	All	Original regression coefficients
α_k	Model recalibration	Recalibrated regression coefficients
λ_k	Model revision	Revised regression coefficients
s_i	Model recalibration & revision	Shrinkage factor to control extent of model update
c_i	All	Person-specific risk cutoff
p_{assign}	All	Target proportion of individuals assigned to intervention
γ_k	Alternative outcomes	Parameters for success probability of Y_i

Table 1: Parameters adjusted to capture various simulation scenarios.

Distribution for X_i

Following the recommended guidelines for the use of the diabetes mellitus risk model, the 2017-2018 NHANES dataset was filtered from the original sample of 9254 persons to 1592 persons *without* diabetes from the ages of 45 to 65 years, so that we could study the impact of intervening on those middle-aged patients at risk for diabetes. Ten predictors were used to measure the predicted 8-year risk of incident diabetes mellitus for each individual in our dataset. These included age, sex, parental history of diabetes, body mass index (BMI), systolic blood pressure, diastolic blood pressure, treatment for hypertension, fasting glucose, HDL cholesterol, and triglyceride. Missing continuous and categorical variables were imputed by chain equations with plausible values drawn

from a distribution specifically designed for each missing datapoint under certain assumptions about the data missingness mechanism [58, 59].

Distribution for R_i

We iteratively sampled R_i from $i = 1$ to $i = n$ to simulate the scenario of a dynamic risk calculator for diabetes. To model recalibration and revision, we inputted the characteristics X_i to obtain a prediction R_i for the 8-year risk of incident diabetes mellitus in an initial batch of individuals ($i \leq n_{\text{batch}} \leq n$). The risk of 8-year diabetes mellitus for each patient in the initial batch was calculated on the log odds scale as:

$$\begin{aligned}
&\log \frac{\text{Risk of diabetes}}{1 - \text{Risk of diabetes}} = \text{ linear predictor} \\
&= \zeta_0 + \zeta_1 \cdot \mathbb{1}_{\{\text{Age}\in[50,65)\}} + \zeta_2 \cdot \mathbb{1}_{\{\text{Age}\in[65,\infty)\}} + \zeta_3 \cdot \mathbb{1}_{\{\text{Male}\}} + \zeta_4 \cdot \mathbb{1}_{\{\text{Parental history}\}} \\
&+ \zeta_5 \cdot \mathbb{1}_{\{\text{BMI}\in[25,30)\}} + \zeta_6 \cdot \mathbb{1}_{\{\text{BMI}\in[30,\infty)\}} + \zeta_7 \cdot \mathbb{1}_{\{\text{HDL-C} < 40 \text{ mg/dL in male or} < 50 \text{ mg/dL in women}\}} \\
&+ \zeta_8 \cdot \mathbb{1}_{\{\text{BP} > 130/85 \text{ mmHG or receiving treatment}\}} + \zeta_9 \cdot \mathbb{1}_{\{\text{Triglyceride} \geq 150 \text{ mg/dL}\}} \\
&+ \zeta_{10} \cdot \mathbb{1}_{\{\text{Fasting glucose}\in[100,126]\}}.
\end{aligned}$$

The values ζ_k for $k = 0, \ldots, 10$ can be found in Wilson et al.[56]. We then transformed this into a risk and subtracted the patient-specific threshold C_i:

$$\frac{1}{1 + \exp\left(-\text{linear predictor}\right)} - C_i.$$

After the initial batch ($n_{\text{batch}} < i \leq n$), we recalibrated (scenarios 2-3) and revised (scenarios 4-5) the risk prediction model based on published recommendations [24, 23]. Given that the risk prediction model's primary purpose is to estimate baseline risk of diabetes (i.e., the risk without treatment), recalibration and revision were done using only untreated individuals. For the recalibration scenarios, we regressed the observed diabetes outcomes on an intercept and the linear predictor of the original model, using the simulated data collected up to individual i. The estimated intercept α_0 and slope α_1 were then adjusted by a shrinkage factor s_i. The risks R_i for individuals i were computed as:

$$\frac{1}{1 + \exp\{-s_i\,\alpha_0 - (1 + s_i\,(\alpha_1 - 1)) \cdot (\text{linear predictor for person } i)\}} - C_i.$$

We considered different choices for s_i to control the extent of recalibration as a function of new data sample size, where $s_i = 0$ represented static risk prediction and $s_i = 1$ represented using the re-calibrated estimates.

For the revision scenarios, we re-estimated the intercept and the regression coefficients of all predictors, using the simulated data collected up to individual i. The re-estimated intercept and regression coefficients were then adjusted by a shrinkage factor s_i using the following formula:

$$\zeta_k + s_i \cdot (\lambda_k - \zeta_k).$$

We considered different choices for s_i to control the extent of revision as a function of new data sample size, where $s_i = 0$ represented the original coefficients and $s_i = 1$ represented using the re-estimated coefficients. The re-estimated coefficients were then used to calculate a revised linear predictor. The risks R_i for individuals i were computed as:

$$\frac{1}{1 + \exp\left(-\text{revised linear predictor}\right)} - C_i.$$

Distribution for A_i

By design, the intervention assignment A_i was $\mathbb{1}_{\{R_i \geq 0\}}$.

Distribution for Y_i

For Simulation 1-4, the binary outcome of diabetes after 8 years, Y_i, was sampled as a Bernoulli random variable with success probability

$$\frac{1}{1 + \exp\{-\gamma_0 - \gamma_1 \cdot (\text{linear predictor for person } i)}} + \gamma_2 A_i.$$

Specifically, when $\gamma_0 = 0$ and $\gamma_1 = 1$, the risk model was correctly specified. When $\gamma_1 = 0$, the prediction is irrelevant. Throughout our simulations, we explored the impact of the risk model on estimation by examining $\gamma_0 = \gamma_1 = 0.5$. Detailed trace plots illustrating how the recalibrated simulations adapted to the incorporated shifts in γ_0 and γ_1 as more patients enter the system can be found in Appendix D.1. Finally, we assigned $\gamma_2 = -0.2$ as the simulated LATE that we want to be able to estimate using our RD estimator.

For Simulation 5, we expanded our analysis beyond the initial prediction of the model with the objective of assessing the model's effectiveness at estimating the LATE on alternative health outcomes. Specifically, we compared weight change in patients who received an intervention against those who did not. For treated patients, change in weight was modeled as a random variable following a normal distribution with a mean (μ) of -5 kg, reflecting the expected average weight loss due to the treatment, and a standard deviation (σ) of 3 kg. The change in weight for untreated patients was also assumed to follow a normal distribution, but with a mean of 0, indicating no expected average change in weight, and a standard deviation of 3.

Simulation Strategy

We conducted 2000 replications of each simulation scenario. In each replication, we used the original 1592 patients in the dataset and bootstrapped 3500 samples with replacement. Across all five simulations, we adopted a patient-specific varying cutoff aimed at achieving a target treatment proportion of 50%. By incorporating this cutoff variability, the simulations offered a more realistic approximation of how risk prediction models might be implemented in practice, adapting to factors such as the number of patients needed for reliable inferences or constraints related to system capacity. Trace plots illustrating the behavior of the cutoff across different simulation scenarios are presented in the Appendix D.2.

We then implemented the proposed estimators to estimate each LATE and calculated its 95% confidence intervals for individual i. To estimate the LATE around the risk cutoff, we first used the risk prediction model of individual i and retroactively estimated the risk for prior patients. We then employed a RD design focusing only on patients who fall inside $W_{r,i}$. All patient characteristics were standardized by centering to their mean and dividing by their standard deviation and these transformed characteristics were then used to find patients inside $W_{r,i}$. The pseudo code for patient selection inside $W_{r,i}$ and estimation can be found in Appendix D.3.

Within $W_{r,i}$, the relationship between the binary outcome (diabetes after 8 years) and the running variable R_i was modeled using a linear probability model (scenarios 1-4). It is important to note that while linear probability models offer intuitive coefficients, they can predict probabilities outside the [0,1] range. Hence, our interpretations are centered around the LATE, especially near the cutoff. Similarly, the relationship between the continuous outcome Y_i (change in weight) and the running variable R was modeled using a linear regression framework for scenario 5. Specifically, the expected outcomes were expressed as:

$$\mathbb{E}\left[Y_i \middle| A_i, R_i, X_i \in W_{r,i}\right] \approx \alpha_0 + \beta_0 R_i + \alpha_1 A_i + \beta_1 (R_i A_i).$$

We implemented sensitivity analyses to assess the robustness of our estimators to three bandwidth sizes (r=0.5, 1, 1.5) and evaluated results with three different individuals serving as the current or final patient in the system (i=1500, 2500, 3500). Note that patient characteristics were standardized and centered, which implies that $r = 1$ is a difference of one standard deviation. By comparing these estimates to the true value of the LATE, we assessed the bias, variance, mean squared error, and confidence interval coverage. These evaluation metrics served as benchmarks for our estimators, helping us gauge their performance. All statistical analyses were performed with R version 4.3.1 (2023-06-16) [60].

7 Results

Sample characteristics

The analyzed sample of patients without diabetes (N=1592) is summarized in Appendix E.1. Briefly, patients were an average of 55.4 years of age and there were slightly more females than males in the sample (52.8%). The majority of patients had no family history of diabetes (75.6%) and most were treated with blood pressure medication (80.3%). Using this sample of patients, we conducted 2000 repetitions of each previously mentioned simulation scenario?. During each repetition, we bootstrapped 3,500 samples with replacement from the initial pool of 1,592 patients. Out of these samples, the goal was to select a subset of patients who were as-if randomly assigned to treatment. The average number of patients ($\bar{n}$) fitting this criterion for each simulation is summarized in Table 2.

Simulation Comparison

Figure 3 presents the estimated LATE of intervening on at risk patients in terms of developing diabetes after 8 years (Simulations 1-4) across all patient and bandwidth sizes. Additional analyses were performed to investigate the LATE of intervening on patients based on diabetes risk predictions on an alternative outcome, in this case the change in weight of a patient, also shown in Figure 3. In general, our causal inference methodology was able to consistently approximate the true value of the simulated effect, $\gamma_2 = -0.2$ across Simulations 1-4 and and $\gamma_2 = -5$ in Simulation 5. Additional details of the analysis can be found in Appendix E.2.

Table 2 summarizes the performance of the model estimates in terms of bias, variance, mean square error (MSE), and coverage across the simulation scenarios. Estimates for Simulation 1, where the risk prediction model remained static, exhibited a range of bias from -0.007 to 0.005 and a range of variance and MSE from 0.011 to 0.056. Model coverage remained high ranging from 94% to 96%. The static simulations presented the lowest overall bias and variance out of all simulations. Bias was lowest in magnitude when $i = 1500$ and $r = 0.5$ (-0.001), while lowest variance and MSE was observed when $i = 3500$ and $r = 1.5$ (0.011).

Estimates of the recalibrated scenarios (Simulations 2 and 3) presented slightly larger ranges of bias and variance. Simulation 2 consistently overestimated the true effect value with a bias ranging from 0.002 to 0.011 and a range of variance and MSE from 0.037 to 0.259, with a slight improvement in coverage that ranged from 95% to 96%. We observed that bias was lowest in magnitude (0.002), accompanied by lowest variance and MSE (0.037) when $i = 3500$ and $r = 0.5$. Similarly, estimates for Simulation 3, where complete recalibration was done, had a bias ranging from -0.006 to 0.005, variance and MSE ranging from 0.019 to 0.29, and coverage ranging from 94% to 96%. In this case, bias was lowest in magnitud when $i = 3500$ and $r = 0.5$ (0.001), while lowest variance and MSE (0.019) was observed when $i = 3500$ and $r = 1.5$.

The revision scenario presented the largest overall bias when compared to Simulations 1-4 (-0.022 to -0.011). Moreover, a decrease in overall variance (0.014 and 0.18) and reduced coverage (92% to 94%) was also observed. Bias was lowest in magnitude when $i = 3500$ and $r = 0.5$ (-0.011), while lowest variance and MSE was found when $i = 3500$ and $r = 1.5$ (0.014). This scenario, however, presented certain estimation challenges where the model failed to converge to an optimal solution, as further detailed in Table 2.

Finally, when considering the effectiveness of the method on the change in weight outcome, we found that Simulation 5 resulted in estimates with a range of bias from -0.044 to 0.028 and a range of variance and MSE from 0.517 to 1.762. Model coverage remained high, ranging from 94% to 96%. While Simulations 1-4 and Simulation 5 are not directly comparable due to different simulated effect sizes, results highlight the usefulness of the framework at identifying the true LATE in all evaluated simulations.

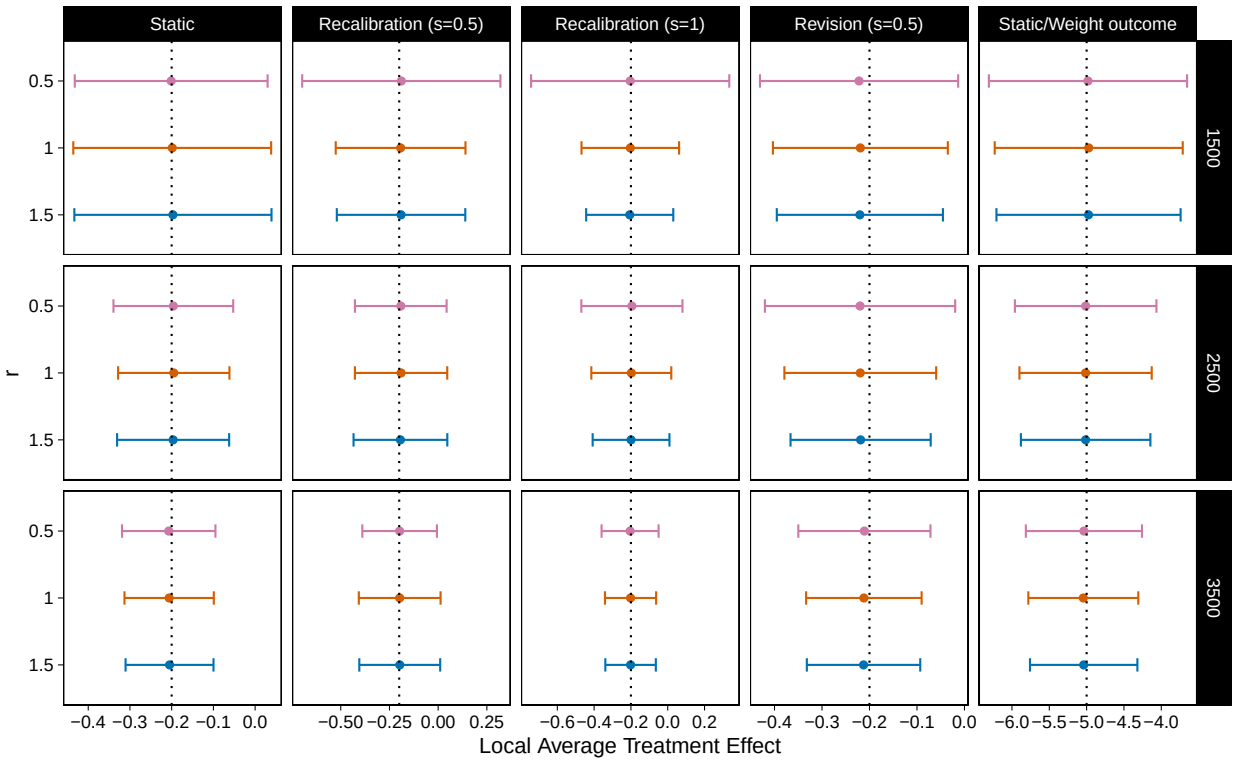

Fig 3. Local average treatment effect of intervening on at risk patients based on predictions from diabetes risk prediction model for each simulation. Dashed line represents the true local average treatment effect of -0.2 for the static, recalibrated and revision simulations (Simulations 1-4), and -5 kg for the static change in weight outcome (Simulation 5).

Table 2: Performance of estimates in terms of bias, variance, mean square error (MSE), and percent coverage of true parameters for 95% confidence intervals at varying sample sizes (i) and bandwidths (r) for each simulation scenario averaged across the 2000 repetitions. True parameter values and average number of patients inside $W_{r,i}$ for each simulation are also reported. Minimum values for bias (magnitude), variance, and MSE for each simulation are in highlighted in bold for ease of identification and comparison.

Simulation*	**i**	**r**	**Value**	**n̄**	**Bias**	**Variance**	**MSE**	**Coverage**
Static	1500	0.5	-0.2	115.7	**-0.001**	0.053	0.053	0.95
Static	2500	0.5	-0.2	192.6	0.004	0.021	0.021	0.95
Static	3500	0.5	-0.2	271.5	-0.007	0.013	0.013	0.95
Static	1500	1	-0.2	141.8	0.001	0.056	0.056	0.95
Static	2500	1	-0.2	235.3	0.005	0.018	0.018	0.95
Static	3500	1	-0.2	331.2	-0.006	0.011	0.012	0.96
Static	1500	1.5	-0.2	144.8	0.003	0.056	0.056	0.94
Static	2500	1.5	-0.2	240.4	0.003	0.018	0.018	0.94
Static	3500	1.5	-0.2	338.7	-0.005	**0.011**	**0.011**	0.96
Recalibration (s=0.5)	1500	0.5	-0.2	114.3	0.011	0.259	0.259	0.96
Recalibration (s=0.5)	2500	0.5	-0.2	188.9	0.008	0.055	0.055	0.96
Recalibration (s=0.5)	3500	0.5	-0.2	263.7	**0.002**	**0.037**	**0.037**	0.95
Recalibration (s=0.5)	1500	1	-0.2	140.3	0.007	0.111	0.111	0.95
Recalibration (s=0.5)	2500	1	-0.2	232.0	0.010	0.056	0.056	0.95
Recalibration (s=0.5)	3500	1	-0.2	323.5	0.003	0.044	0.044	0.95
Recalibration (s=0.5)	1500	1.5	-0.2	143.0	0.010	0.109	0.109	0.95
Recalibration (s=0.5)	2500	1.5	-0.2	236.1	0.006	0.058	0.058	0.95
Recalibration (s=0.5)	3500	1.5	-0.2	329.4	0.003	0.043	0.043	0.96
Recalibration (s=1)	1500	0.5	-0.2	106.0	-0.004	0.290	0.290	0.94
Recalibration (s=1)	2500	0.5	-0.2	175.2	0.005	0.076	0.076	0.96
Recalibration (s=1)	3500	0.5	-0.2	245.4	-0.005	0.024	0.024	0.96
Recalibration (s=1)	1500	1	-0.2	128.5	-0.003	0.070	0.070	0.95
Recalibration (s=1)	2500	1	-0.2	212.2	0.002	0.047	0.047	0.95
Recalibration (s=1)	3500	1	-0.2	296.6	-0.002	0.019	0.019	0.95
Recalibration (s=1)	1500	1.5	-0.2	130.5	-0.006	0.056	0.056	0.95
Recalibration (s=1)	2500	1.5	-0.2	215.7	**0.001**	0.044	0.044	0.95
Recalibration (s=1)	3500	1.5	-0.2	301.7	-0.002	**0.019**	**0.019**	0.95
Revision (s=0.5)**	1500	0.5	-0.2	119.4	-0.022	0.044	0.044	0.94
Revision (s=0.5)	2500	0.5	-0.2	195.8	-0.020	0.040	0.041	0.92
Revision (s=0.5)**	3500	0.5	-0.2	272.7	**-0.011**	0.019	0.019	0.94
Revision (s=0.5)**	1500	1	-0.2	150.9	-0.019	0.034	0.034	0.93
Revision (s=0.5)	2500	1	-0.2	248.7	-0.019	0.026	0.026	0.93
Revision (s=0.5)**	3500	1	-0.2	345.5	-0.012	0.015	0.015	0.94
Revision (s=0.5)**	1500	1.5	-0.2	159.4	-0.020	0.031	0.031	0.93
Revision (s=0.5)	2500	1.5	-0.2	262.7	-0.019	0.022	0.022	0.93
Revision (s=0.5)**	3500	1.5	-0.2	364.4	-0.013	**0.014**	**0.014**	0.94
Static/Weight outcome	1500	0.5	-5	127.4	0.018	1.762	1.762	0.94
Static/Weight outcome	2500	0.5	-5	212.9	**-0.013**	0.899	0.899	0.95
Static/Weight outcome	3500	0.5	-5	301.0	-0.036	0.605	0.606	0.95

Static/Weight outcome	1500	1	-5	148.8	0.028	1.586	1.586	0.95
Static/Weight outcome	2500	1	-5	248.0	-0.014	0.785	0.785	0.95
Static/Weight outcome	3500	1	-5	349.0	-0.044	0.542	0.544	0.95
Static/Weight outcome	1500	1.5	-5	153.8	0.026	1.521	1.522	0.94
Static/Weight outcome	2500	1.5	-5	256.5	-0.013	0.752	0.752	0.95
Static/Weight outcome	3500	1.5	-5	360.9	-0.039	**0.517**	**0.519**	0.95

*In this context, "static" models are those that do not undergo recalibration or revision. Nonetheless, these models do dynamically adjust the treatment cutoff for individual patients to maintain a consistent treatment ratio of 50% across the patient population.

** For these simulations, 1 repetition did not converge and was eliminated from analysis.

Sensitivity to bandwidth and sample size

We assessed the sensitivity of our estimates to variations in bandwidth and sample size. As seen in Figure 3, model performance was robust across bandwidth selections (r) and sample sizes (i). There was a general trend of decreasing bias, variance, and MSE, alongside an increase in coverage, as i increased from 1500 to 3500 (Table 2).

On the other hand, when fixing i, bandwidth selection has a lesser influence on the model's performance. This can be further explained when evaluating how the average sample size ($\bar{n}$) varied across the different bandwidth values, it can be observed that $\bar{n}$ does not significantly increase when fixing i (Table 2). This observation is pivotal, especially when considering the method's need for a large initial sample size. Although we begin with a considerable number of samples, the selection of patients within a specific bandwidth significantly reduces the number of samples that can be effectively considered as quasi-randomly assigned to the treatment group for estimation. This reduction in sample size, as a consequence of patient selection criteria, can substantially impact the robustness of the estimates.

It's crucial to emphasize that this observation is intrinsically linked to the specific estimation strategy employed in our study, which is designed around the configuration of our chosen risk prediction model. The estimation strategy aimed to find patients whose treatment assignment could be considered quasi-random based on small perturbations in their characteristics. The number of patients fitting this criterion is a result of this process, which is tailored to the structure of the risk prediction model used. Hence, this result is particular to our applied risk prediction model, highlighting the need to assess the sensitivity of bandwidth size when applying this optimization strategy to other prediction algorithms.

8 Discussion

Quasi-experimental designs, particularly RD designs, provide a valuable framework for evaluating clinical decision-making aided by risk prediction models. These frameworks maintain the experimental validity of a randomized control trial while eliminating their associated ethical concerns. However, current quasi-experimental designs do not account for the dynamic aspect of learning inherent in modern healthcare systems. To bridge this gap, we introduce an innovative quasi-experimental design that is specifically adapted to the iterative nature of learning environments. Our framework was shown to successfully assess the true average treatment effect of interventions informed by continuously refined risk prediction models while considering the potential effect that individuals may have with each other.

To our knowledge, we are the first study to consider a RD design to estimate the average causal effect of interventions in healthcare that allows for updates to the risk prediction model and adjustments made by the health system. Several scenarios of continuous learning and refining

of a clinical risk prediction model, including recalibration, revision, cutoff updating, and diverse outcomes were evaluated. Consistently low bias, variance, MSE, and high coverage metrics were observed across all tested scenarios. Although this study presents a comparison between static, recalibrated, and revised risk prediction models, it does not advocate for a specific model updating method in the clinical setting. Rather, the focus of this work is to examine the influence of the current patient in system and bandwidth size, as well as dynamic updates, on the frameworks ability to reliably estimate the true LATE. The question of whether using a static, recalibrated or revised model in practice is specific to the application and ideally the decision should be made through ongoing monitoring of the model's performance. Other works have studied the benefits of updating clinical prediction model [23], and have found that simple recalibration updates have improved the model performance to a similar extent as more exhausting revisions methodologies without the added complexity, risk of overfitting, and model instability. Notably, in our simulations, estimating the LATE in the revision scenario exhibited the highest overall bias and encountered convergence challenges, likely a manifestation of model instability due to repeated revisions. Although the revision scenario results improved as more data was available in the system, a simple recalibration in many cases is sufficient and preferable.

Although the motivation for this work stems from learning systems and their related risk prediction models, this framework can be generalized to any other system that uses risk prediction models to improve operations and aid in decision making. We present a general approach for estimating the average treatment effect of intervening based on risk prediction models in any application for any particular outcome of interest, as portrayed through our simulation study. Through this work, we provide algorithms for learning systems to deploy in real-time to guide their use of risk predictions models in day-to-day operations, with the ultimate goal of improving service and reducing costs.

There are several limitations to consider. First, the present study was based on a simulated model and may not accurately reflect the performance of actual risk prediction algorithms when deployed in healthcare systems. Second, the study did not consider the operational aspects of implementing such models in a clinical setting, which could introduce additional sources of bias. Third, the choice of bandwidth and cutoff are context specific highlighting the need to assess sensitivity of these choices when applying this methodology in other settings. These choices are also linked to the available sample size used for estimation which should be robust enough to guarantee reliable estimates. This aspect is especially critical given the framework's requirement for a substantial sample size. Although the study starts with thousands of samples, the process of selecting patients for estimation significantly reduces the number of samples that can be considered as effectively randomized. Consequently, this reduction in sample size can substantially affect the robustness of the estimates.

In addition, the complexity of learning systems extends beyond patient-to-patient interference. While learning systems may help facilitate the identification of high-risk patients, human compliance with algorithmic recommendations has been shown to introduce both automation bias and algorithm aversion into the system [61, 62]. The first term refers to the overreliance of a human to algorithmic decisions while the latter refers to underreliance of algorithmic recommendations. Physicians play a critical role in implementing interventions for patients identified at risk by the system and are ultimately responsible for making treatment decisions. However, care providers might disagree with risk prediction tools based on number of reasons including the patients' particular circumstance, resource constraints or if they suspect bias in the algorithms recommended treatment. Therefore, extensions of this work involve taking into account the human factor in treatment choices, and identifying potential areas for intervention improvement.

In summary, we introduce a quasi-experimental design that leverages the strengths of an RD design and is specially crafted to evaluate the effectiveness of intervening according to a risk predic-

tion model in a modern healthcare system. Our work is both significant and innovative, as it offers rigorous estimation of local average treatment effects while taking into account the interference between patients that occurs as healthcare systems use internal data to improve patient care. This is crucial for the success of risk prediction models in clinical care, as it allows healthcare systems to address implementation issues in real-time and adapt risk prediction models to an ever-evolving healthcare landscape. Our approach is innovative in keying in on the specific type of interference present, allowing us to develop efficient and valid estimators. Unlike other designs that overlook interference or consider it arbitrary, our solution recognizes the importance of interference in evaluating the effectiveness of risk prediction. Ultimately, our approach offers a means for healthcare systems and regulatory bodies to ethically monitor the clinical use of machine learning and artificial intelligence algorithms, without hindering the potential for rapid deployment and improvement. As these algorithms represent the future of healthcare, we believe our work is a critical step towards ensuring their successful integration into clinical care.

References

[1] Douglas W Challener, Larry J Prokop, and Omar Abu-Saleh. The proliferation of reports on clinical scoring systems: issues about uptake and clinical utility. *JAMA*, 321(24):2405–2406, 2019.

[2] David Atkins, Christos A Makridis, Gil Alterovitz, Rachel Ramoni, and Carolyn Clancy. Developing and implementing predictive models in a learning healthcare system: traditional and artificial intelligence approaches in the veterans health administration. *Annual Review of Biomedical Data Science*, 5:393–413, 2022.

[3] Gwen Costa Jacobsohn, Margaret Leaf, Frank Liao, Apoorva P Maru, Collin J Engstrom, Megan E Salwei, Gerald T Pankratz, Alexis Eastman, Pascale Carayon, and Douglas A Wiegmann. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. In *Healthcare*, volume 10, page 100598. Elsevier, 2022.

[4] Brian W Patterson, Collin J Engstrom, Varun Sah, Maureen A Smith, Eneida A Mendonça, Michael S Pulia, Michael D Repplinger, Azita Hamedani, David Page, and Manish N Shah. Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. *Medical care*, 57(7):560, 2019.

[5] Daniel Hekman, Amy Cochran, Apoorva Maru, Manish Shah, Frank Liao, Hanna Barton, Douglas Wiegmann, Maureen Smith, and Brian Patterson. Effectiveness of an emergency department based machine learning clinical decision support tool to prevent outpatient falls among older adults: A protocol paper for a quasi-experimental study (preprint). *JMIR Research Protocols*, 04 2023.

[6] UW Health Innovation Program. Hipxchange.

[7] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

[8] Thomas E Robinson, Lifeng Zhou, Ngaire Kerse, John DR Scott, Jonathan P Christiansen, Karen Holland, Delwyn E Armstrong, and Dale Bramley. Evaluation of a new zealand program to improve transition of care for older high risk adults. *Australasian journal on ageing*, 34(4):269–274, 2015.

[9] Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

[10] J Michael McGinnis, Brian Powers, and Claudia Grossmann. *Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: workshop series summary*. National Academies Press, 2011.

[11] Minoru Nakatsugawa, Zhi Cheng, Ana Kiess, Amanda Choflet, Michael Bowers, Kazuki Utsunomiya, Shinya Sugiyama, John Wong, Harry Quon, and Todd McNutt. The needs and benefits of continuous model updates on the accuracy of rt-induced toxicity prediction models within a learning health system. *International Journal of Radiation Oncology* Biology* Physics*, 103(2):460–467, 2019.

[12] Thomas M Maddox, Nancy M Albert, William B Borden, Lesley H Curtis, T Bruce Ferguson Jr, David P Kao, Gregory M Marcus, Eric D Peterson, Rita Redberg, and John S Rumsfeld. The learning healthcare system and cardiovascular care: a scientific statement from the american heart association. *Circulation*, 135(14):e826–e857, 2017.

[13] Emelia J Benjamin, Daniel Levy, Sonya M Vaziri, Ralph B D'Agostino, Albert J Belanger, and Philip A Wolf. Independent risk factors for atrial fibrillation in a population-based cohort: the framingham heart study. *JAMA*, 271(11):840–844, 1994.

[14] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[15] Pamela A Sytkowski, William B Kannel, and Ralph B D'Agostino. Changes in risk factors and the decline in mortality from cardiovascular disease: the framingham heart study. *New England Journal of Medicine*, 322(23):1635–1641, 1990.

[16] Ronald C Kessler, Irving Hwang, Claire A Hoffmire, John F McCarthy, Maria V Petukhova, Anthony J Rosellini, Nancy A Sampson, Alexandra L Schneider, Paul A Bradley, and Ira R Katz. Developing a practical suicide risk prediction model for targeting high-risk patients in the veterans health administration. *International journal of methods in psychiatric research*, 26(3):e1575, 2017.

[17] Arkaitz Artetxe, Andoni Beristain, and Manuel Grana. Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, 164:49–64, 2018.

[18] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. 306(15):1688–1698.

[19] Stuart W Grant, Gary S Collins, and Samer AM Nashef. Statistical primer: developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery*, 54(2):203–208, 2018.

[20] MAE Binuya, EG Engelhardt, W Schats, MK Schmidt, and EW Steyerberg. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Medical Research Methodology*, 22(1):316, 2022.

[21] Ben Van Calster, Ewout W Steyerberg, Laure Wynants, and Maarten van Smeden. There is no such thing as a validated prediction model. *BMC medicine*, 21(1):1–8, 2023.

[22] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[23] Ewout W Steyerberg, Gerard JJM Borsboom, Hans C van Houwelingen, Marinus JC Eijkemans, and J Dik F Habbema. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in medicine*, 23(16):2567–2586, 2004.

[24] KJM Janssen, KGM Moons, CJ Kalkman, DE Grobbee, and Y Vergouwe. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of clinical epidemiology*, 61(1):76–86, 2008.

[25] Thomas PA Debray, Hendrik Koffijberg, Daan Nieboer, Yvonne Vergouwe, Ewout W Steyerberg, and Karel GM Moons. Meta-analysis and aggregation of multiple published prediction models. *Statistics in medicine*, 33(14):2341–2362, 2014.

[26] David A Jenkins, Matthew Sperrin, Glen P Martin, and Niels Peek. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic and prognostic research*, 2(1):1–9, 2018.

[27] Ting-Li Su, Thomas Jaki, Graeme L Hickey, Iain Buchan, and Matthew Sperrin. A review of statistical updating methods for clinical prediction models. *Statistical methods in medical research*, 27(1):185–197, 2018.

[28] Sarah M Greene, Robert J Reid, and Eric B Larson. Implementing the learning health system: from concept to action. *Annals of internal medicine*, 157(3):207–210, 2012.

[29] Charles P Friedman, Adam K Wong, and David Blumenthal. Achieving a nationwide learning health system. *Science translational medicine*, 2(57):57cm29–57cm29, 2010.

[30] Lynn M Etheredge. A rapid-learning health system: what would a rapid-learning health system look like, and how might we get there? *Health affairs*, 26(Suppl1):w107–w118, 2007.

[31] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

[32] Evelyn T Chang, Jean Yoon, Aryan Esmaeili, Donna M Zulman, Michael K Ong, Susan E Stockdale, Elvira E Jimenez, Karen Chu, David Atkins, and Angela Denietolis. Outcomes of a randomized quality improvement trial for high-risk veterans in year two. *Health Services Research*, 56:1045–1056, 2021.

[33] Leora I Horwitz, Masha Kuznetsova, and Simon A Jones. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med*, 381(12):1175–1179, 2019.

[34] Mona Jabbour, Amanda S Newton, David Johnson, and Janet A Curran. Defining barriers and enablers for clinical pathway implementation in complex clinical settings. *Implementation Science*, 13(1):1–13, 2018.

[35] Michelle Heys, Erin Kesler, Yali Sassoon, Emma Wilson, Felicity Fitzgerald, Hannah Gannon, Tim Hull-Bailey, Gwendoline Chimhini, Nushrat Khan, and Mario Cortina-Borja. Development and implementation experience of a learning healthcare system for facility based newborn care in low resource settings: The neotree. *Learning Health Systems*, 7(1):e10310, 2023.

[36] Ron C Li, Steven M Asch, and Nigam H Shah. Developing a delivery science for artificial intelligence in healthcare. *NPJ digital medicine*, 3(1):107, 2020.

[37] Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, and Zoie Shui-Yee Wong. Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications. *Yearbook of medical informatics*, 28(01):128–134, 2019.

[38] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.

[39] Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Annals of Agricultural Sciences*, 10:1–51, 1923.

[40] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[41] David S Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.

[42] Matias D Cattaneo, Brigham R Frandsen, and Rocio Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, 3(1):1–24, 2015.

[43] Paul R. Rosenbaum. *Observational Studies*. Springer New York, NY, 2002.

[44] Matias D Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. Inference in regression discontinuity designs under local randomization. *The Stata Journal*, 16(2):331–367, 2016.

[45] Thomas Robinson, Rod Jackson, Susan Wells, Andrew Kerr, and Roger Marshall. An observational study of how clinicians use cardiovascular risk assessment to inform statin prescribing decisions. *The New Zealand Medical Journal (Online)*, 130(1463):28–38, 2017.

[46] Sue Wells, Tania Riddell, Andrew Kerr, Romana Pylypchuk, Carol Chelimo, Roger Marshall, Daniel J Exeter, Suneela Mehta, Jeff Harrison, and Cam Kyle. Cohort profile: the predict cardiovascular disease cohort in new zealand primary care (predict-cvd 19). *International journal of epidemiology*, 46(1):22–22, 2017.

[47] Matias D Cattaneo, Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele. Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248, 2016.

[48] Daniel Mark Butler. A regression discontinuity design analysis of the incumbency advantage and tenure in the us house. *Electoral Studies*, 28(1):123–128, 2009.

[49] Luc Behaghel, Bruno Crépon, and Béatrice Sédillot. The perverse effects of partial employment protection reform: The case of french older workers. *Journal of Public Economics*, 92(3-4):696–721, 2008.

[50] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

[51] Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200, 2007.

[52] Peter M Aronow, Nicole E Basta, and M Elizabeth Halloran. The regression discontinuity design under interference: a local randomization-based approach. *Observational Studies*, 3(2):129–133, 2017.

[53] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[54] Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019.

[55] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[56] Peter WF Wilson, James B Meigs, Lisa Sullivan, Caroline S Fox, David M Nathan, and Ralph B D'Agostino. Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study. *Archives of internal medicine*, 167(10):1068–1074, 2007.

[57] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). 2017–2018 national health and nutrition examination survey data.

[58] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[59] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[60] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[61] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648*, 2021.

[62] Riccardo Fogliato, Maria De-Arteaga, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. *Available at SSRN 4050125*, 2022.

Appendix

A Proof of Proposition 1

Proof. Fix $n \in \mathbb{N}$. First note that, in general, $X_i(a_B) = X_i$ for $B \subseteq \{1, \ldots, n\}$. Subsequently, for any $B \subseteq \{1, \ldots, n\} \cup \{i\}$,

$$\begin{aligned}
Y_i(a_B) &= f_y(X_i(a_B), \bar{A}_i(a_B), N_{y_i}) && \text{by Definition 1} \\
&= f_y(X_i, a_i, N_{Y_i}) && \text{since } X_i(a_B) = X_i \text{ and Definition 1} \\
&= f_y(X_i(a_i), \bar{A}_i(a_i), N_{y_i}) && \text{since } X_i(a_i) = X_i \text{ and Definition 1} \\
&= Y_i(a_i) && \text{by Definition 1.}
\end{aligned}$$

Additionally, we prove that the potential outcomes $Y_i(a_i)$ are iid for $i = 1, \ldots, n$. By Definition 1,

$$Y_i(a_i) = f_y(X_i(a_i), \bar{A}_i(a_i), N_{y_i}).$$

By Assumption 1, the random variables $\{N_{x_i}, N_{y_i}\}_{i=1}^n$ are mutually independent with $\{N_{x_i}\}_{i=1}^n$ identically distributed and similarly for $\{N_{y_i}\}_{i=1}^n$. Hence, since $Y_i(a_i)$ is a deterministic function of $\{N_{x_i}, N_{y_i}\}_{i=1}^n$, it is also mutually independent and identically distributed for $i = 1, \ldots, n$. Therefore, potential outcomes $Y_1(a_1), Y_2(a_2), \cdots, Y_n(a_n)$ are iid.

We show the expressions hold for general $i \in \{1, \ldots, n\}$ and any $B \subseteq \{1, \ldots, n\}$. To this end, first observe that from what we showed above

$$\begin{aligned}
X_i(a_{1:n}) &= X_i \\
Y_i(a_{1:n}) &= Y_i(a_{1:i}) = Y_i(a_i).
\end{aligned}$$

Similarly,

$$\begin{aligned}
X_{i-1}(a_{1:n}) &= X_i \\
Y_{i-1}(a_{1:n}) &= Y_{i-1}(a_{1:i-1}) = Y_{i-1}(a_{i-1})
\end{aligned}$$

It follows that,

$$\begin{aligned}
R_i(a_{1:n}) &= f_{r_i}(X_i(a_{1:n}), X_{1:i-1}(a_{1:n}), \bar{A}_{1:i-1}(a_{1:n}), Y_{1:i-1}(a_{1:n})) && \text{by Definition 1} \\
&= f_{r_i}(X_i, X_{1:i-1}, a_{1:i-1}, Y_{1:i-1}(a_{1:n})) && \text{since } X_i(a_B) = X_i \text{ and Definition 1} \\
&= f_{r_i}(X_i, X_{1:i-1}, a_{1:i-1}, Y_1(a_1), \cdots, Y_{i-1}(a_{i-1})) && \text{since } Y_{i-1}(a_{1:n}) = Y_{i-1}(a_{1:i-1}) = Y_{i-1}(a_{i-1}) \\
&= f_{r_i}(X_i, X_{1:i-1}, a_{1:i-1}, Y_{1:i-1}(a_{1:i-1})) \\
&= f_{r_i}(X_i(a_{1:i-1}), X_{1:i-1}(a_{1:i-1}), a_{1:i-1}, Y_{1:i-1}(a_{1:i-1})) && \text{since } X_i(a_{1:i-1}) = X_i \\
&= R_i(a_{1:i-1}) && \text{by Definition 1}
\end{aligned}$$

We can use the last equality to show the remaining expressions:

$$\begin{aligned}
A_i(a_{1:n}) &= \mathbb{I}(R_i(a_{1:n}) \geq 0) && \text{by Definition 1} \\
&= \mathbb{I}(R_i(a_{1:i-1}) \geq 0) && \text{since } R_i(a_{1:n}) = R_i(a_{1:i-1}) \\
&= A_i(a_{1:i-1}) && \text{by Definition 1}
\end{aligned}$$

Thus, Proposition 1 holds for general $i \in \{1, \ldots, n\}$ and any $B \subseteq \{1, \ldots, n\}$, and the proof is complete.

□

B Proof of Proposition 2

Proof. Let $\omega \in \Omega$. Fix disjoint $B \subseteq \{1, \cdots, n\}$. Let $a_B := A_B[w]$. By Definition, $Z(A_B)[\omega] = Z(a_B)[\omega]$. It remains to show that $Z(a_B)[\omega] = Z[\omega]$. We will show this for each of the variables:

$$\begin{aligned}
X_i(a_B)[\omega] &= N_{x_i}[\omega], && \text{by Definition 1}\\
X_i(a_B)[\omega] &= N_{x_i}[\omega] = X_i[\omega], && \text{by Assumption 1}\\
X_i(A_B)[\omega] &= X_i(a_B)[\omega] = X_i[\omega]. && \text{by Definition 3}
\end{aligned}$$

We can use this last equality to show the remaining expressions:

$$\begin{aligned}
Y_i(a_B)[\omega] &= f_y(X_i(a_B)[\omega], \bar{A}_i(a_B)[\omega], N_{Y_i}[\omega])\\
&= f_y(X_i(a_B)[\omega], a_i, N_{Y_i}[\omega]), && \text{by Definition 1}\\
Y_i(A_B)[\omega] &= Y_i(a_B)[\omega] && \text{by Definition 3}\\
&= f_y(X_i(a_B)[\omega], a_i, N_{Y_i}[\omega])\\
&= f_y(X_i[\omega], A_i[\omega], N_{Y_i}[\omega]) && \text{since } X_i(a_B)[\omega] = X_i[\omega] \text{ and } A_i[\omega] = a_i\\
&= Y_i[\omega]. && \text{by Assumption 1}
\end{aligned}$$

It follows that:

$$\begin{aligned}
R_i(a_B)[\omega] &= f_{r_i}(X_i(a_B)[\omega], X_{1:i-1}(a_B)[\omega], \bar{A}_{1:i-1}(a_B)[\omega], Y_{1:i-1}(a_B)[\omega])\\
&= f_{r_i}(X_i(a_B)[\omega], X_{1:i-1}(a_B)[\omega], a_{1:i-1}, Y_{1:i-1}(a_B)[\omega]), && \text{by Definition 1}\\
R_i(A_B)[\omega] &= R_i(a_B)[\omega] && \text{by Definition 3}\\
&= f_{r_i}(X_i(a_B)[\omega], X_{1:i-1}(a_B)[\omega], a_{1:i-1}, Y_{1:i-1}(a_B)[\omega])\\
&= f_{r_i}(X_i[\omega], X_{1:i-1}[\omega], A_{1:i-1}[\omega], Y_{1:i-1}[\omega]) && \text{from equations above}\\
&= R_i[\omega]. && \text{by Assumption 1}
\end{aligned}$$

$$\begin{aligned}
A_i(a_B)[\omega] &= \mathbb{I}(R_i(a_B)[\omega] \geq 0), && \text{by Definition 1}\\
A_i(A_B)[\omega] &= A_i(a_B)[\omega] && \text{by Definition 3}\\
&= \mathbb{I}(R_i(a_B)[\omega] \geq 0)\\
&= \mathbb{I}(R_i[\omega] \geq 0) && \text{from equations above}\\
&= A_i[\omega]. && \text{by Assumption 1}
\end{aligned}$$

□

C Proof of Theorem 2

C.1 Window around cutoff

Lemma 1. *Let $r > 0$ be an arbitrary small number. Consider R_i as the prediction model for individual i that maps characteristics $x \in \mathcal{X}$ to $\mathbb{R}$. Define*

$$W_{r,i} = \{x \in \mathcal{X} : \exists\, x' \in \mathcal{X}, \left\|x - x'\right\| < r, R_i(x') = 0\}$$

and assume $\mathbb{P}[A_i = a, R_i = r] > 0$ *for all* $a \in \{0, 1\}$ *in a small of neighborhood around* 0*. Then*

$$\lim_{r\to 0} \mathbb{E}[Y|A_i = 1, R_i = r] = \lim_{r\to 0} \mathbb{E}[Y|A_i = 1, X_i \in W_{r,i}], \text{ and}$$
$$\lim_{r\to 0} \mathbb{E}[Y|A_i = 0, R_i = -r] = \lim_{r\to 0} \mathbb{E}[Y|A_i = 0, X_i \in W_{r,i}].$$

Proof. Conditioning on $X_i \in W_{r,i}$ we obtain

$$\begin{aligned}\mathbb{E}[Y|A_i = 1, R_i = r] = \mathbb{E}[Y|A_i = 1, R_i = r, X_i \in W_{r,i}]\,\mathbb{P}[X_i \in W_{r,i}|A_i = 1, R_i = r] + \\ \mathbb{E}[Y|A_i = 1, R_i = r, X_i \in W_{r,i}]\,\mathbb{P}[X_i \in W^c_{r,i}|A_i = 1, R_i = r].\end{aligned}$$

where $W^c_{r,i}$ is the complement of $W_{r,i}$ in Ω.

Taking the limit as $r \to 0$ in the expression above, the proof is completed once we demonstrate the following:

$$\lim_{r\to 0} \mathbb{P}[X_i \in W_{r,i}|A_i = 1, R_i = r] = 1$$

First, we show the following equality of events

$$\bigcap_{n\in\mathbb{N}} \{X_i \in W_{\frac{1}{n},i}, A_i = 1, R_i \geq 0\} = \{A_i = 1, R_i = 0\}. \tag{C.1}$$

Let ω be in the left-hand side event above, where ω is such that $X_i(\omega) \in W_{\frac{1}{n},i}$ for all $n \in \mathbb{N}$. The definition of $W_{\frac{1}{n},i}$ implies that $X_i(\omega) = x'$ for some $x' \in \mathcal{X}$ with $R_i(x') = 0$. In other words, $R_i(X_i(\omega)) = 0$.

For the other inclusion, let $n \in \mathbb{N}$ be arbitrary. If $R_i(\omega) = 0$. Consider $x' = X_i(\omega)$, clearly $X_i(\omega) \in W_{\frac{1}{n},i}$.

By Assumption 1, $A_i = \mathbb{1}(R_i \geq 0)$, which implies

$$\{X_i \in W_{r,i}, A_i = 1, R_i = r\} = \{X_i \in W_{r,i}, A_i = 1, R_i \geq 0\}.$$

Therefore by Equation (C.1) we have

$$\lim_{r\to 0} \mathbb{P}[X_i \in W_{r,i}|A_i = 1, R_i = r] = \lim_{r\to 0} \frac{\mathbb{P}[X_i \in W_{r,i}, A_i = 1, R_i = r]}{\mathbb{P}[A_i = 1, R_i = r]} = 1$$

□

C.2 Proof of LATE

Proof. Let $r > 0$ be an arbitrary small number. Define $\alpha_i = Y_i(0)$ and $\gamma_i = Y_i(1) - Y_i(0)$.

By causal irrelevance (Proposition 1) and consistency (Proposition 2), we can write:

$$Y_i = Y_i(A_i) = Y_i(0) + A_i(Y_i(1) - Y_i(0)) = \alpha_i + A_i\gamma_i,$$

Replacing the expression for Y_i in the calculation of the average difference-in-means for units with predicted risk right above/below the cutoff, we arrive at the following conclusion:

$$\begin{aligned}\mathbb{E}[Y_i|R_i = r] - \mathbb{E}[Y_i|R_i = -r] = \{\mathbb{E}[\alpha_i|R_i = r] - \mathbb{E}[\alpha_i|R_i = -r]\} + \\ \{\mathbb{E}[A_i\gamma_i|R_i = r] - \mathbb{E}[A_i\gamma_i|R_i = -r]\}.\end{aligned} \tag{C.2}$$

As $r \to 0$, the first term of Equation (C.2) tends to zero by Assumption 2. The remaining proof focuses on the second term in the limit, considering either continuity or local randomization assumptions.

Identification under continuity: Adding and subtracting $\mathbb{E}[A_i(0)\gamma_i|R_i = r]$ to the second term of Equation (C.2), we can rewrite it as:

$$\mathbb{E}[(A_i(1) - A_i(0))\gamma_i|R_i = r] + \mathbb{E}[A_i(0)\gamma_i|R_i = r] - \mathbb{E}[A_i(0)\gamma_i|R_i = -r],$$

and by Assumption 2, this expression tends to

$$\mathbb{E}[(A_i(1) - A_i(0))\gamma_i|R_i = 0] \text{ as } r \to 0$$

Combining these results, and using the fact that $A_i = \mathbb{1}(R_i \geq 0)$ (Assumption 1), we get an expression for the difference-in-means at the limit, namely:

$$\begin{aligned}\lim_{r\to 0^+}(\mathbb{E}[Y_i|A_i = 1, R_i = r] - \mathbb{E}[Y_i|A_i = 0, R_i = -r]) &= \mathbb{E}[(A_i(1) - A_i(0))\gamma_i|R_i = 0] \\ &= \mathbb{E}\left[Y_i(1) - Y_i(0)\middle|R_i = 0\right] \quad \text{(C.3)}\end{aligned}$$

After applying Lemma 1 to the expression above, we obtain the expression of Theorem 2:

$$\beta_i = \mathbb{E}\left[Y_i(1) - Y_i(0)\middle|R_i = 0\right] = \lim_{r\to 0}\left(\mathbb{E}\left[Y_i\middle|A_i = 1, X_i \in W_{r,i}\right] - \mathbb{E}\left[Y_i\middle|A_i = 0, X_i \in W_{r,i}\right]\right)$$

where $W_{r,i} = \{x \in \mathcal{X} : \exists\, x' \in \mathcal{X}, \left\|x - x'\right\| < r, R_i(x') = 0\}$

Identification under local randomization: Under local randomization (Assumption 3),

$$Y_i(a) \perp\!\!\!\perp A_i \mid R_i \approx 0$$

for all $a \in \{0, 1\}$.

Applying this to the second term in Equation (C.2) and using Assumption 2, we obtain:

$$\begin{aligned}\lim_{r\to 0^+}\mathbb{E}[A_i(1)\gamma_i|R_i = r] &= \lim_{r\to 0^+}\mathbb{E}[A_i(1)|R_i = r]\,\mathbb{E}[\gamma_i|R_i = r] = \mathbb{E}[Y_i(1)|R_i = 0] \text{ and} \\ \lim_{r\to 0^+}\mathbb{E}[A_i(0)\gamma_i|R_i = -r] &= \lim_{r\to 0^+}\mathbb{E}[A_i(0)|R_i = -r]\,\mathbb{E}[\gamma_i|R_i = -r] = \mathbb{E}[Y_i(0)|R_i = 0].\end{aligned}$$

Combining these expressions with Equation (C.2) we arrive at:

$$\lim_{r\to 0^+}(\mathbb{E}[Y_i|A_i = 1, R_i = r] - \mathbb{E}[Y_i|A_i = 0, R_i = -r]) = \mathbb{E}[Y_i(1) - Y_i(0)|R_i = 0],$$

and the proof follows as in the continuity case, after applying Lemma 1.

□

D Supplementary material Section 6

D.1 Trace plots of recalibrated simulations

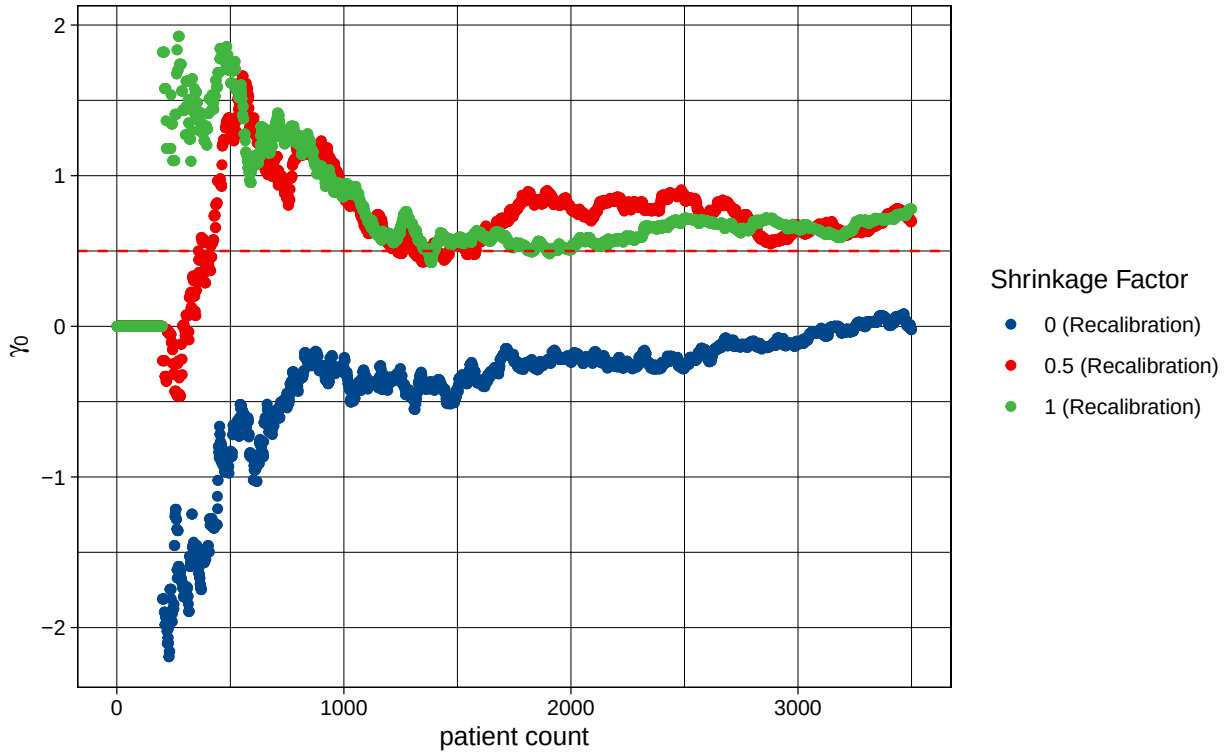

Fig 4. Trace plot illustrating how risk prediction model is adjusting to shifts in γ_0. As the recalibration shrinkage factor increases, the risk prediction is able to capture the shift in the model. Dashed red line marks the shift in the slope $\gamma_0 = 0.5$.

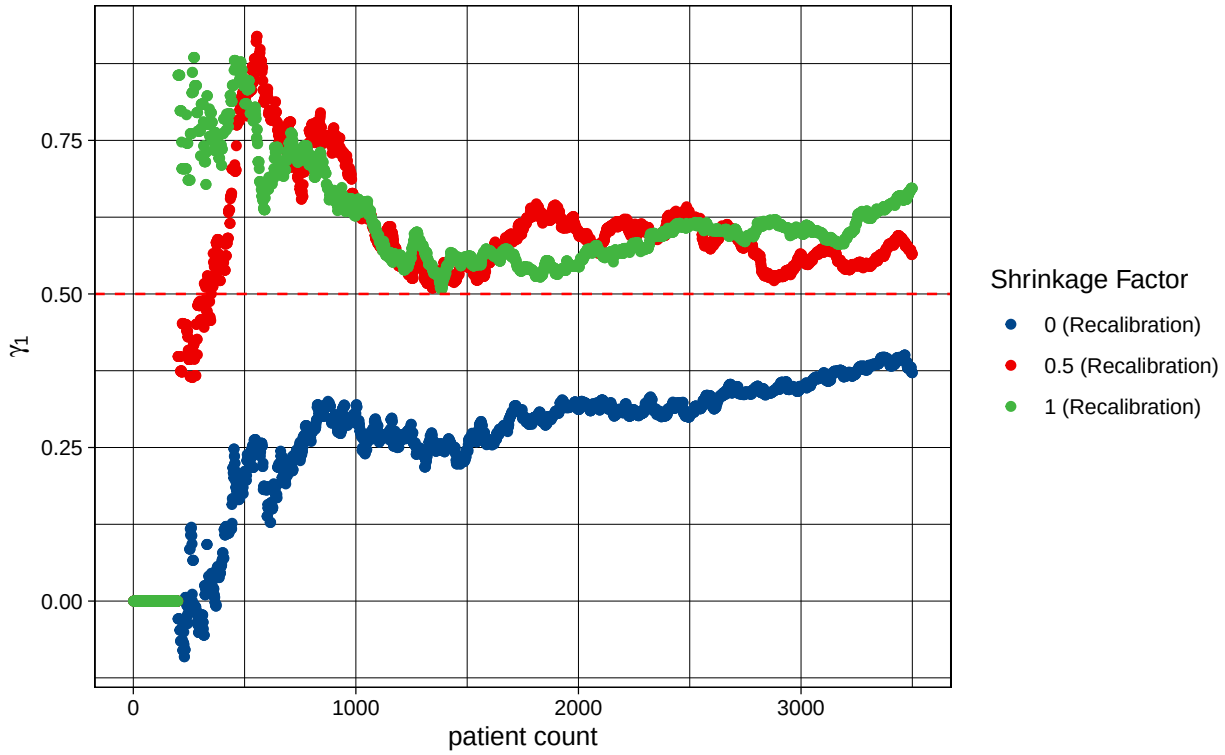

Fig 5. Trace plot illustrating how risk prediction model is adjusting to shifts in γ_1. As the recalibration shrinkage factor increases, the risk prediction is able to capture the shift in the model. Dashed red line marks the shift in the intercept $\gamma_0 = 0.5$.

D.2 Cutoff Plots

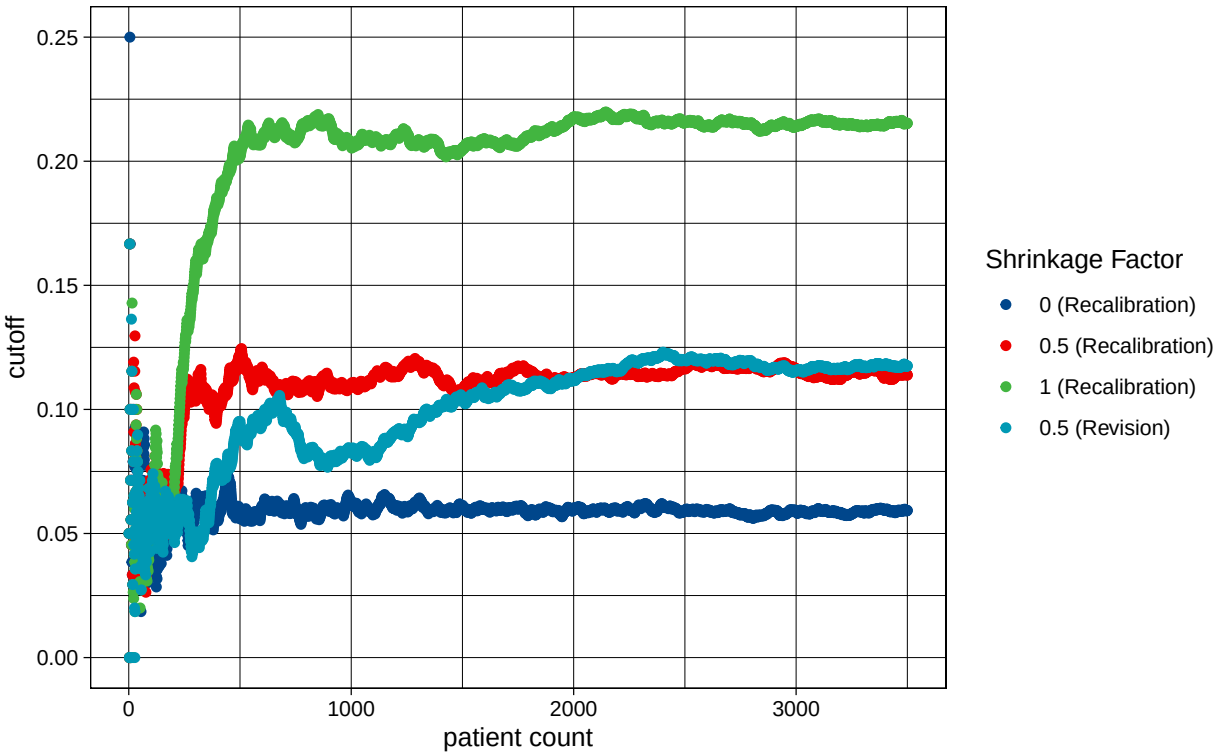

Fig 6. Trace plot illustrating patient-specific cutoff values across main simulation scenarios to achieve an average treatment proportion of 50%. As recalibration/revision is implemented, the median of the risk shifts showcasing the importance of adopting a dynamic cutoff approach in practice.

D.3 Pseudo Code

Algorithm 1 Patient Selection and Estimation Algorithm

Require: $X[1..i]$: observed characteristics of patients
Require: $A[1..i]$: observed intervention assignments
Require: $Y[1..i]$: observed outcomes
Require: R_i: current risk prediction model for patient i
Require: r: bandwidth for patient selection
Require: f_0, f_1: polynomial functions
Ensure: γ_2: estimated parameter (local average treatment effect)

Initialize retroactive risks:
for $j = 1$ to i **do**
    $\hat{r}[j] \leftarrow 0$
end for
Estimate retroactive risks using patient i risk prediction model:
$\hat{r}_i \leftarrow \text{EstimateRisk}(x[i], R_i)$
for $j = 1$ to $i - 1$ **do**
    $\hat{r}[j] \leftarrow \text{EstimateRisk}(x[j], \hat{r}_i)$
end for
Select patients inside window based on the bandwidth criteria:
$J \leftarrow \emptyset$
for $j = 1$ to i **do**
    for each x' in X **do**
        if $\text{Norm}(x[j] - x') < r$ and $\hat{r}_i(x') = 0$ **then**
            Add j to J
        end if
    end for
end for
Perform regression to estimate α_1:
Initialize variables for regression
for each j in J **do**
    Compute polynomial terms based on $a[j]$ and $\hat{r}[j]$
    Update regression variables
end for
Calculate γ_2 based on the regression formula
return γ_2

E Supplementary material Section 7

E.1 Sample characteristics

Table 3: Sample characteristics of patients without diabetes that met inclusion criteria for the diabetes mellitus risk prediction model.

Variable	**N = 1,5921**[1]
Age (years)	55.41 (5.95)
Systolic blood pressure (mm HG)	128.72 (18.81)
Diastolic blood pressure (mm HG)	76.80 (12.45)
Fasting glucose (mg/dL)	107.32 (22.83)
HDL cholesterol (mg/dL)	55.13 (16.71)
Triglyceride (mg/dL)	117.12 (130.23)
Body mass index (kg/m2)	29.59 (6.90)
Initial weight (kg)	82.14 (21.35)
Sex	
Female	840 (52.8%)
Male	752 (47.2%)
Taking blood pressure treatment	
No	313 (19.7%)
Yes	1,279 (80.3%)
History of diabetes	
No	1,204 (75.6%)
Yes	388 (24.4%)
[1]Mean (SD); n (%)	

E.2 Simulation comparison

Table 4: Mean and standard deviation (sd) of the local average treatment effect of intervening on at risk patients based on predictions from diabetes risk prediction model for each simulation. True parameter values for each simulation are also reported.

Simulation	**i**	**r**	**Value**	**Estimate**	**sd**
Static	1500	1.5	-0.2	-0.197	0.236
Static	1500	1	-0.2	-0.199	0.237
Static	1500	0.5	-0.2	-0.201	0.231
Static	2500	1.5	-0.2	-0.197	0.134
Static	2500	1	-0.2	-0.195	0.133
Static	2500	0.5	-0.2	-0.196	0.144
Static	3500	1.5	-0.2	-0.205	0.105
Static	3500	1	-0.2	-0.206	0.107
Static	3500	0.5	-0.2	-0.207	0.112
Recalibration (s=0.5)	1500	1.5	-0.2	-0.190	0.330
Recalibration (s=0.5)	1500	1	-0.2	-0.193	0.333
Recalibration (s=0.5)	1500	0.5	-0.2	-0.189	0.509
Recalibration (s=0.5)	2500	1.5	-0.2	-0.194	0.241
Recalibration (s=0.5)	2500	1	-0.2	-0.190	0.237

Recalibration (s=0.5)	2500	0.5	-0.2	-0.192	0.235
Recalibration (s=0.5)	3500	1.5	-0.2	-0.197	0.208
Recalibration (s=0.5)	3500	1	-0.2	-0.197	0.210
Recalibration (s=0.5)	3500	0.5	-0.2	-0.198	0.191
Recalibration (s=1)	1500	1.5	-0.2	-0.206	0.237
Recalibration (s=1)	1500	1	-0.2	-0.203	0.265
Recalibration (s=1)	1500	0.5	-0.2	-0.204	0.538
Recalibration (s=1)	2500	1.5	-0.2	-0.199	0.209
Recalibration (s=1)	2500	1	-0.2	-0.198	0.217
Recalibration (s=1)	2500	0.5	-0.2	-0.195	0.275
Recalibration (s=1)	3500	1.5	-0.2	-0.202	0.137
Recalibration (s=1)	3500	1	-0.2	-0.202	0.139
Recalibration (s=1)	3500	0.5	-0.2	-0.205	0.155
Revision (s=0.5)	1500	1.5	-0.2	-0.220	0.175
Revision (s=0.5)	1500	1	-0.2	-0.219	0.184
Revision (s=0.5)	1500	0.5	-0.2	-0.222	0.209
Revision (s=0.5)	2500	1.5	-0.2	-0.219	0.148
Revision (s=0.5)	2500	1	-0.2	-0.219	0.160
Revision (s=0.5)	2500	0.5	-0.2	-0.220	0.200
Revision (s=0.5)	3500	1.5	-0.2	-0.213	0.119
Revision (s=0.5)	3500	1	-0.2	-0.212	0.122
Revision (s=0.5)	3500	0.5	-0.2	-0.211	0.139
Static/Weight outcome	1500	1.5	-5	-4.974	1.233
Static/Weight outcome	1500	1	-5	-4.972	1.259
Static/Weight outcome	1500	0.5	-5	-4.982	1.327
Static/Weight outcome	2500	1.5	-5	-5.013	0.867
Static/Weight outcome	2500	1	-5	-5.014	0.886
Static/Weight outcome	2500	0.5	-5	-5.013	0.948
Static/Weight outcome	3500	1.5	-5	-5.039	0.719
Static/Weight outcome	3500	1	-5	-5.044	0.737
Static/Weight outcome	3500	0.5	-5	-5.036	0.778