medRxiv preprint doi: https://doi.org/10.1101/2022.04.30.22274520; this version posted May 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Death by Round Numbers and Sharp Thresholds: How to Avoid Dangerous AI EHR Recommendations

Benjamin J. Lengerich*[1,2], Rich Caruana†[3], Mark E. Nunnally‡[4], and Manolis Kellis§[1,2]

[1]Massachusetts Institute of Technology
[2]Broad Institute of MIT and Harvard
[3]Microsoft Research
[4]NYU Langone Health

April 30, 2022

Abstract

Recommendations for clinical practice consider consistency of application and ease of implementation, resulting in treatment decisions that use round numbers and sharp thresholds even though these choices produce statistically sub-optimal decisions. To characterize the impact of these choices, we examine four datasets and the biases underlying patient mortality risk. We document two types of suboptimalities: (1) discontinuities in which treatment decisions produce step-function changes in risk near clinically-important round-number cutoffs, and (2) counter-causal paradoxes in which anti-correlation between risk factors and aggressive treatment produces risk curves that contradict underlying causal risk. We also show that outcomes have improved over decades of refinement of clinical practice, reducing but not eliminating the strength of these biases. Finally, we provide recommendations for clinical practice and data analysis: interpretable "glass-box" models can transform the challenges of statistical confounding into opportunities to improve medical practice.

Introduction

Evidence-based medicine seeks to develop treatment protocols that reduce the risk of adverse outcomes by collecting and analyzing real-world evidence with probabilistic models; however, real-world evidence observes patients who receive informed interventions. Because treatment choices are based on patient risk factors, data-driven risk models can confound patient risk with clinical practice (Figure 1) and produce incorrect but persistent statistical effects. For example, patients hospitalized with pneumonia have *lower* mortality risk if they *also* have asthma [1]. This finding contradicts biological causality since pneumonia is exacerbated by asthma [2]; however, it reflects medical practice – patients with known risk factors like asthma often request and receive more urgent care.

This example is doubly problematic in the quest for data-driven medicine: not only would a naïve AI model learn an effect that contradicts medical causality, but this contradiction results in the model predicting a lower intrinsic risk specifically for the patients for whom standard treatment would be most beneficial. In general, if we use data-driven risk models to rank patients according to risk, patients who routinely receive effective treatments will be deemed low-risk precisely because they were helped by standard treatments. While risk models can be successfully interpreted by physicians who understand the clinical context [3, 4], most data-driven models do not have this contextual grounding.

*blengeri@mit.edu
†rcaruana@microsoft.com
‡mark.nunnally@nyulangone.org
§manoli@mit.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

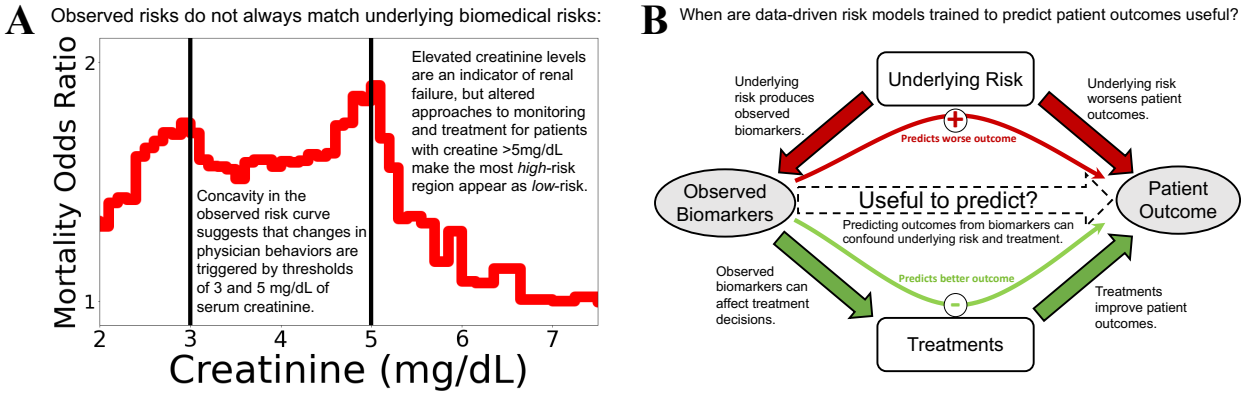


Figure 1: Real-world confounding factors are treacherous to data-driven risk models, especially when data-driven risk models are intended to be used to guide treatment decisions. **(A)** The mortality risk of pneumonia patients suggests that high levels of serum creatinine (which indicate kidney failure) are associated with better survival, even after correcting for other risk factors in a multivariable predictive model. The sharp inflection points in mortality odds ratio at the round numbers of 3mg/dL and 5mg/dL suggest that this association between elevated creatinine and reduced mortality is prompted by discrete treatment thresholds rather than smooth biomedical risk factors. While this confounding between risk factor and clinical decisions is a challenge for data-driven analysis (a naïve triage model would withhold treatment from patients with elevated serum creatinine, precisely those who most need treatment), it is also a blessing: the non-monotonicity alerts us to an opportunity to optimize treatment protocols. **(B)** Causal flow underlying "treatment" effects. Causal arrows are filled, observed variables are shown in gray ovals and unobserved variables in white boxes. Data-driven analyses often estimate $\mathbb{P}(\text{Outcome}|\text{Biomarker})$, but this is only a faithful surrogate for $\mathbb{P}(\text{Outcome}|\text{Underlying Risk})$ if treatments were to have negligible impacts on the outcome. In reality, treatments (broadly defined, including monitoring, therapeutics, diagnostics, and patient behavior) have significant impacts and frequently lead to "paradoxical" effects in which data-driven systems assign low risk to high-risk patients precisely because the high-risk patients are most likely to be effectively treated. To build models which can effectively guide treatment decisions, we require intelligible models and medical domain knowledge to audit the probabilistic models and understand if all relevant treatment effects have been sufficiently corrected.

Instead, data-driven models faithfully recapitulate the effects and biases of real-world clinical practice. This is not a indictment of limited *samples* or a subpar *model*: the confounding effects arise from real-world behavior, so any model that accurately predicts observed outcomes must recapitulate these effects. This is also not a problem of insufficient *features*: even if we were to observe all characteristics of treatment for all patients, standardized treatments without randomization do not provide statistical identifiability to separate the impact of treatments from the impact of features that drove treatment decisions. Without tools that enable medical experts to examine, audit, and contextualize data-driven models, we will not know if all confounding effects have been identified and corrected.

We see this challenge as an opportunity. Paradoxical statistical effects are produced by sub-optimal clinical practice; thus, transparent ("glass-box") risk models allow these paradoxical statistical effects to be detected and uncover opportunities for optimization. By inspecting the risk curves learned by glass-box models, we can transform the challenge of confounding into a solution for improving evidence-based medicine. We find two main types of treatment effects: discontinuous risk profiles caused by a reliance on round-numbers, and paradoxical risk curves caused by overly successful treatments, which reduce the risk of high-risk patients below the risk of untreated low-risk patients. We first examine the pneumonia dataset introduced above to systematically identify other treatment effects. Next, we study several versions of the widely-used MIMIC dataset collected over several decades [5–7] and examine how treatment effects have changed over time in the Beth Israel Deaconess Medical Center. Finally, using glass-box machine learning, we estimate the excess mortality attributable to reliance on round-number thresholds and identify areas for improvement with personalized medicine.

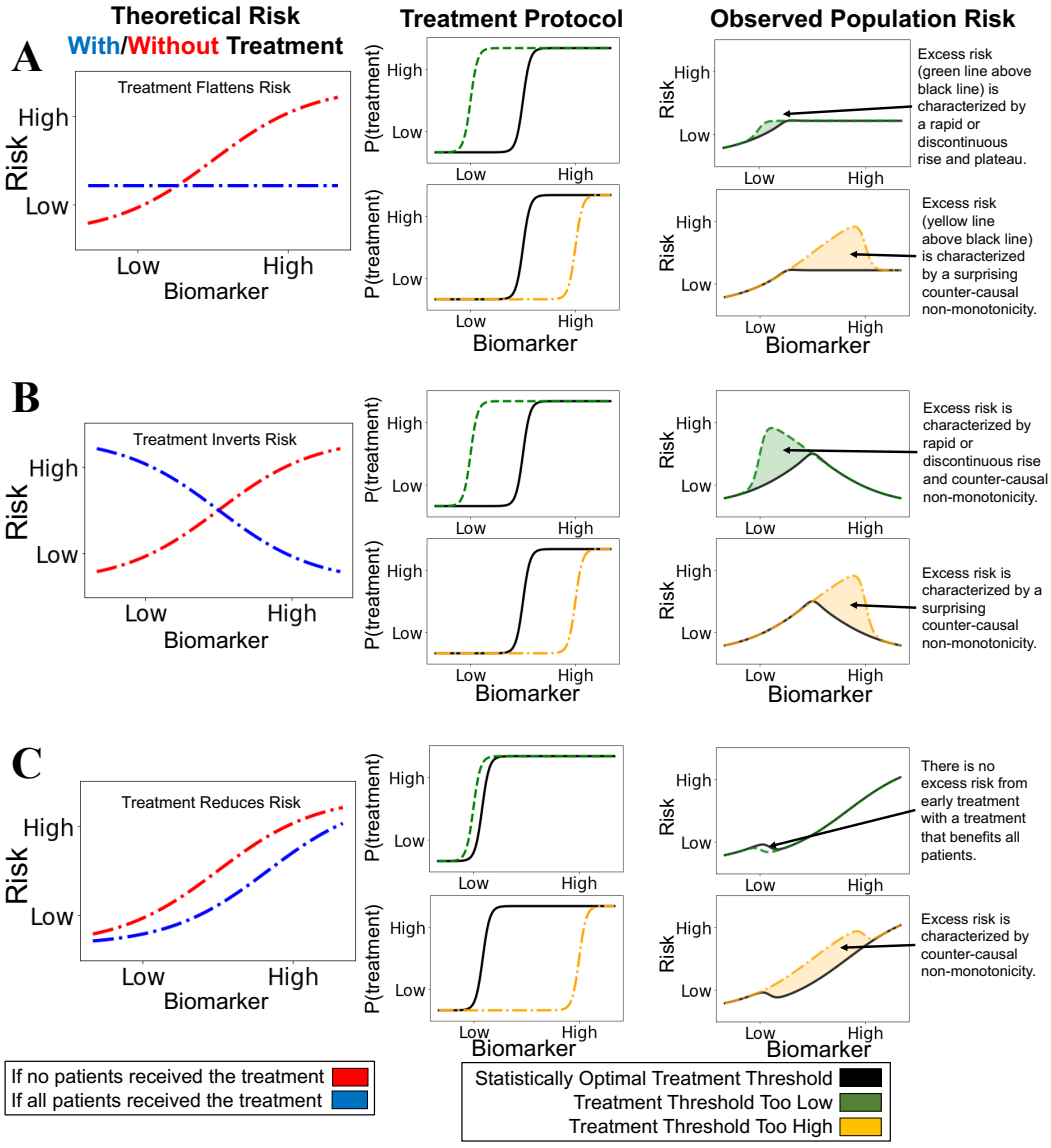


Figure 2: Characteristic types of treatment effects found in data-driven medical informatics. In each panel, we depict the underlying intrinsic risk for treated and untreated patients (left), distribution of treatment decisions (middle), and observed risk (right). When treatment decisions are misaligned from statistical optimality, the observed risk includes excess risk that causes characteristic risk curve shapes. Looking for these shapes can identify opportunities for improving treatment protocols. **(A)** If the treatment induces a constant risk (e.g. surgery), the statistically optimal treatment threshold (black) is the point where the untreated risk crosses above the treated risk. If the treatment threshold is lower than this optimal threshold (green), there is a rapid and/or discontinuous rise in risk, followed by a risk plateau. Conversely, if the treatment threshold is higher than the optimal threshold (yellow), there is a non-monotonic overshoot of risk. **(B)** If the treatment induces an inverse risk (e.g. vasodilators produce an inverse risk with respect to blood pressure), a low threshold (green) produces a rapid and/or discontinuous rise in risk followed by a counter-causal non-monotonicity, while a high threshold (yellow) produces a smooth increase in risk followed by a rapid and/or discontinuous decrease in risk. **(C)** If the treatment induces a constant benefit for all patients (e.g. an idealized treatment), then there is no threshold that is too low, and any threshold produces a counter-causal non-monotonicity.

Results

Real-World Clinical Practice Produces Recognizable Statistical Artifacts

A "treatment effect" is a special case of a confounder in which the "treatment" (broadly interpreted to include monitoring, therapeutics, diagnostics, and patient behavior) influences patient outcomes. When analyzed in a clinical trial, effects of randomly-assigned treatments are statistically desirable; however, when analyzing real-world evidence in which treatments are not uniformly assigned to patients, effects of the treatment can be accidentally included in the observed risk profile of the underlying risk factor (Figure 1). To identify these confounders from data-driven risk curves, we document two classes of recognizable patterns: (1) discontinuities, in which clinical behavior results in step-function changes in risk near clinically-important round-number thresholds and (2) counter-causal paradoxes, in which risk curves contradict underlying biological risk because of the impact of effective treatments. These recognizable patterns exploit the differences between biological risk and medical data biased by informed interventions: underlying biological risk without interventions is smooth while in contrast human medical systems often rely on discrete policies and heuristics to select interventions.

Depending on the treatment, the observed risk profile may be a combination of one or both of these classes of characteristic patterns. For a treatment that completely flattens the risk curve (Figure 2A), the statistically optimal treatment threshold is the point where the untreated risk crosses above the treated risk (when defining the treated risk to include the risk of the treatment itself). If the treatment threshold is lower than this optimal threshold (green), there is a rapid and/or discontinuous rise in risk, followed by a risk plateau. Conversely, if the treatment threshold is higher than the optimal threshold (yellow), there is a non-monotonic overshoot of risk. If the treatment induces an inverse risk (Figure 2B), a low threshold produces a rapid and/or discontinuous rise in risk followed by a counter-causal non-monotonicity, while a high threshold produces a smooth increase in risk followed by a rapid and/or discontinuous decrease in risk. If the treatment induces a constant benefit for all patients (Figure 2C), then any threshold produces a counter-causal non-monotonic drop in risk. In all of these cases, although we only observe the population risk, we can apply our medical understanding of clinical decisions and treatment impacts to reverse-engineer the effects of clinical decision making and suggest improvements to clinical practice. In general, clinical decisions that are statistically near-optimal yield smoother and flatter risk curves. As medicine becomes increasingly precise and personalized, we expect to see risk curves smooth out round-number biases and flatten observed risks — we later discuss an example (Figure 4D) where improvements in clinical decision-making have smoothed and flattened the observed risk curves.

A Model for Systematically Identifying Statistical Artifacts

A practical way to identify these statistical artifacts is to find regions where the difference between the expected and observed risks are large. Given an accurate and high-resolution model of patient risk, we can apply two practical heuristics for identifying these two types of treatment effects. For the first effect, we can find regions in which the underlying condition has smoothly-changing risk but the data-driven risk curve is jumpy, indicating that the change in risk is more likely due to the discontinuous nature of informed interventions rather than continuously-varying underlying risk. For the second effect, we can audit the risk curves for surprising regimes that contradict medical expectations, often assigning low risk to patients that we believe have high intrinsic risk.

To produce an accurate and high-resolution model of patient risk, we trained generalized additive models (GAMs) [8] to estimate mortality risk from patient characteristics, comorbidities, and lab tests. This model class is well-suited for several reasons: (1) the additive models can be precisely decomposed into risk curves of single variables for interpretation, (2) the flexible component functions allow non-monotonic and non-linear risk curves without any implicit preferences, (3) many treatment protocols and clinical decisions (which sum multiple sources of evidence) are inherently additive models, and (4) additive models provide the ability to edit the model [9] and reason about changes to univariable treatment protocol thresholds. Finally, we choose tree-based GAMs [10] from `InterpretML` [11] because: (1) they are scale-invariant models and allow features to be represented in their original units, (2) they allow discontinuities, (3) they capture non-monotonic curves that model decreases in risk just as well as increases in risk, and (4) they achieve accuracy competitive with uninterpretable methods such as deep neural networks on tabular datasets.

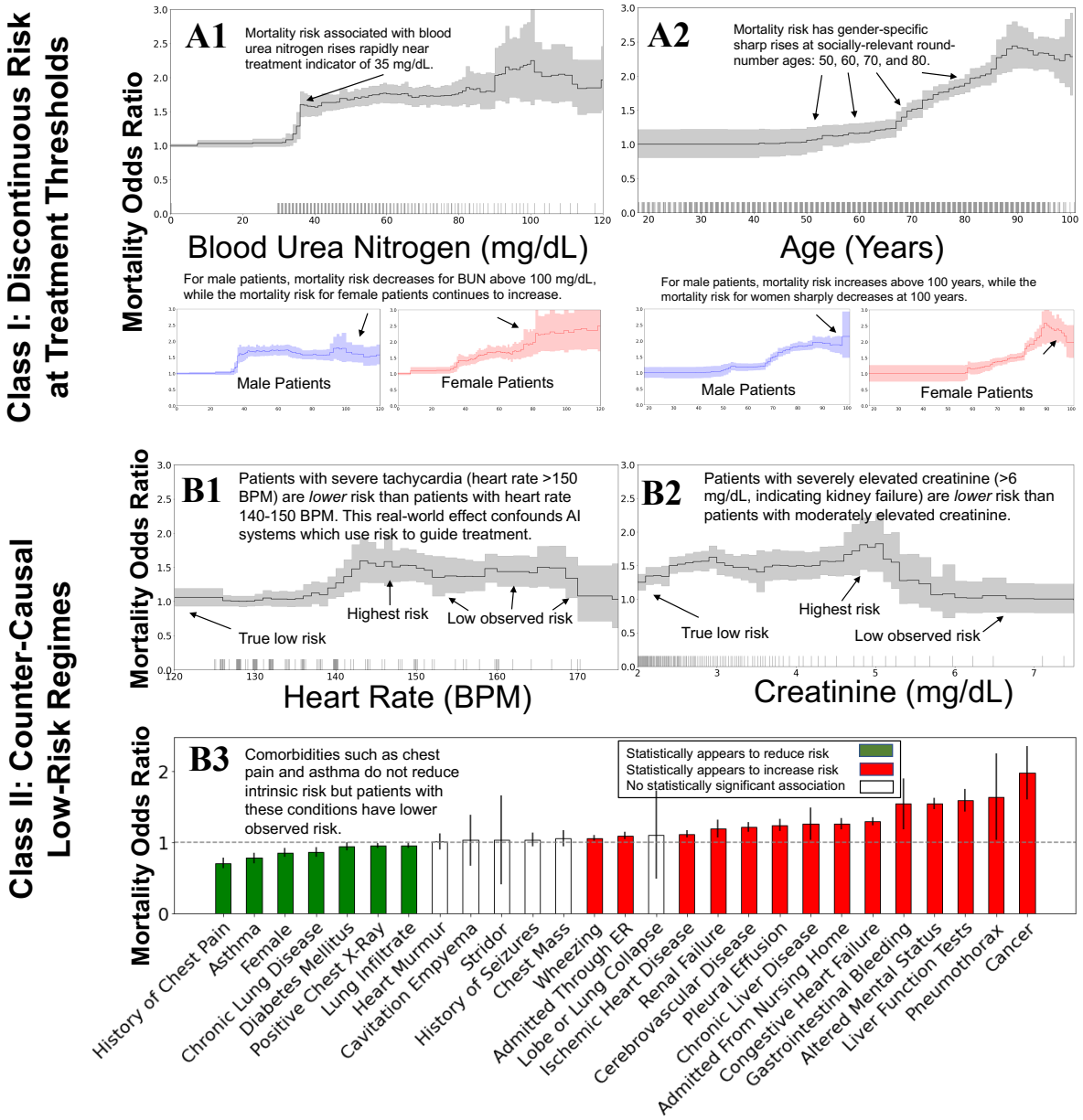


Figure 3: The mortality risks of patients with pneumonia display a number of confounding effects that can be characterized in two broad categories: discontinuities and counter-causal effects. In all plots, we show the effect of each variable on patient mortality risk (with 95% confidence intervals shaded) after correcting for all other observable factors of patient risk using a generalized additive model. Each tick along the horizontal axis denotes 10 observed patients. **(A)** Discontinuities in the observed mortality risk are produced by behavior influenced by discrete round-number thresholds, including blood urea nitrogen above 35 mg/dL and 100 mg/dL (**A1**), and 10-year increments in age, most notably age 100 for female patients (**A2**). **(B)** Counter-causal trends in the observed mortality risk curves lead to lab values and comorbidities that are intrinsically high-risk being predicted as low-risk due to effective treatments. This confounding can lead data-driven AI models to prefer lab values that are intrinsically unsafe, including heart rate above 150 BPM (**B1**) and serum creatinine above 5 mg/dL (**B2**). Finally, chronic comorbidities that increase perceived urgency for treatment, including history of chest pain, asthma, and chronic lung disease, are associated with lower mortality risk (**B3**), even though these comorbidities increase intrinsic risk.

Real-World Clinical Practice Generates Discontinuous Mortality Risk Curves

We next turn to real-world evidence from mortality risk curves of hospitalized pneumonia patients. First, we see that blood urea nitrogen (BUN) has a discontinuous impact on mortality risk (Figure 3A1), with a non-monotone structure that implies different treatments are given at different BUN levels. As expected, lower levels of BUN are associated with survival; however, there is a rapid rise in mortality risk from 30-40mg/dL, with a plateau from 40-80mg/dL, and possibly a reversal of the trend above 100mg/dL. The amelioration of risk at BUN of 40mg/dL is stronger for male patients than female patients, with the risk for female patients rising continuously from 30-50mg/dL. Finally, the risk associated with elevated BUN continues to rise for female patients while the risk for male patients is relatively flat from 40-120 mg/dL, suggesting an opportunity for improved care of female patients with elevated BUN.

Mortality risk also changes in discontinuous steps with respect to patient age, including increases in mortality risk at several round-number ages: 50, 60, 70, and 80 (Figure 3A2). In addition, we see evidence of a "centenarian effect" for female patients approaching age 100, as the mortality risk surprisingly declines (clinicians may be biased to interpret a patient's advanced age as an indicator of resilience and ability to recover from correctable conditions). This centenarian effect is the opposite of the commonly called "old man's best friend"[12] effect where pneumonia may be seen as a blessing for elderly patients with potentially terminal and unpleasant comorbidities. In both of these examples (BUN and age), the biases are so strong that they override the smoothness of inherent biological risk to produce discontinuous risk curves with step changes at round numbers and clinically-meaningful thresholds.

Real-World Clinical Practice Generates Counter-Causal Mortality Risk Curves

We also see that clinical decisions can produce "counter-causal" risk curves in which the observed risk contradicts underlying biomedical risk. In regions where patients would be at high risk without treatment, aggressive treatment can lower risk so much that the observed risk is relatively low. As a result, we observe counter-causal effects with low-risk regions produced by aggressive treatment. For example, elevated heart rate above 145 beats per minute (BPM) (Figure 3B1) and elevated serum creatinine above 5mg/dL (Figure 3B2) are both associated with increased survival (lower mortality risk) of pneumonia patients. These statistical effects (each measured after correcting for all other patient risk factors) contradict medical causality: heart rate above 145 BPM corresponds to severe tachycardia while serum creatinine above 5mg/dL indicates advanced renal dysfunction [13], suggesting that intervening factors including aggressive treatment were given to these patients who would otherwise be at extreme risk absent treatment. Purely data-driven artificial intelligence (AI) protocols would de-prioritize care for patients with tachycardia and elevated creatinine, and may even use these extreme regions as targets toward which patients should be guided.

Chronic comorbidities also produce paradoxical statistical effects, similar to the original motivating example of "protective" asthma for pneumonia patients [1]. For example, history of chest pain, asthma, and chronic lung disease reduce the probability of mortality in pneumonia patients by more than 30%, 20%, and 18%, respectively (Figure 3B3); these effects are robustly estimated and not sensitive to the model class or algorithm used to estimate the effects. We categorize Boolean comorbidities as "counter-causal" effects because *all* Boolean comorbidities (encoded as Yes/No values) have discontinuous risk curves; it is the "counter-causal" effects of Boolean comorbidities that indicate strong confounding with treatment effects. These effects contradict medical intuition of causality, suggesting that risk models trained on these real-world datasets could systematically underestimate mortality risk for patients with prior comorbidities. This systematic underestimation occurs because patients are acutely aware of their own respiratory distress, and their previous experiences and diagnoses will influence how they react as new problems arise. Similarly, clinicians will of course react differently to patients with different medical histories. As a result, the observed risk is a combination of underlying risk, perceived urgency by both the patient and the clinician, and the impact of treatments. When developing data-driven treatment protocols, we must take care to not disadvantage patients with comorbidities that may have warranted aggressive treatment in the training data. For both lab tests and comorbidities, it is important to identify these counter-causal regions of low risk that are produced by aggressive treatments because these are places where AI models that do not understand the causal impact of effective treatments would mistakenly predict patients to be at low risk, possibly denying them the care needed (and routinely provided) to reduce the risk.

Real-World Clinical Practice Has Improved over Years of Protocol Refinement

We next examine three versions of the MIMIC dataset [5–7] that have been used by thousands of researchers to train statistical models of mortality risk. These snapshots were collected over three decades of intensive care units (ICU), providing a large-scale view of medical practice and outcomes over many years. We find that these datasets have strong treatment effects, including both discontinuous and counter-causal risk curves, and while the impact of these treatment effects has decreased over the years the effects are still robustly estimated by data-driven risk models. Finally, we identify treatments recorded in MIMIC-IV [7] that could de-confound the effects of treatment from the underlying biological risk. Overall, we find that the biases and paradoxes in mortality risk are associated with observable treatment patterns, typically based at clinically-meaningful and round number thresholds, and that while these effects have reduced over time, they are still routinely and robustly identified to impact patient mortality risk.

Serum Creatinine Elevated levels of serum creatinine indicate kidney dysfunction[14]; however, this biomarker is not strongly associated with mortality (Figure 4A). Instead, the highest risk (after correcting for all other patient risk factors) is observed for patients with serum creatinine in the range 3-5mg/dL. The plateau in mortality rate at 3mg/dL is associated with a strongly increased likelihood of treatment with intermittent hemodialysis (Figure 4A2). Similarly, the likelihood of continuous renal replacement therapy (CRRT) begins to increase at a creatinine level of 3mg/dL and plateaus at 4mg/dL (Figure 4A3). At a serum creatinine level of 5 mg/dL, the likelihood of intermittent hemodialysis begins to decrease while the likelihood of CRRT correspondingly increases, and the mortality rate decreases. These patterns suggest a possibility that the treatments for kidney dysfunction reduce patient mortality even below the risk level of patients with healthy levels of creatinine. One serious concern is that machine learning models that seek to minimize empirical mortality likelihood might paradoxically try to increase serum creatinine to unhealthy levels.

Serum Sodium As expected from medical intuition regarding hydration, the healthy range of 135-145 mEq/L serum sodium is associated with low mortality risk (Figure 4B1). On either side of this healthy range, the probability of mortality increases, rising rapidly for hyponatremia below 135 mEq/L and hypernatremia above 145 mEq/L. However, this medical intuition is broken at 150 mEq/L: the probability of mortality declines for extremely elevated serum sodium. This threshold of 150 mEq/L is an indicator of moderate hypernatremia [15], which is typically treated with increased attention from care providers, reduced likelihood of oral (PO) food intake (Figure 4B2) and increased electrolyte and water replacements (Figure 4B3). As a result of aggressive treatment of severe hypernatremia and under-treatment for mild hypernatremia [16], the probability of mortality is *lower* for patients with serum sodium above 155 mEq/L than for patients with serum sodium of 145 mEq/L. This effect is robustly supported by statistical evidence and is clinically meaningful — any accurate machine learning model trained on these datasets has learned that severe hypernatremia is lower risk than moderate hypernatremia. Applying these data-driven models in clinical practice could divert treatments away from the group of patients who would benefit *most* from treatment.

Age Patient age has discontinuous impact on mortality rates (Figure 4C1). While the impact of each particular round-number age has changed over the years (possibly due to social changes such as retirement), age 80 confers increased risk in all 3 three datasets. This increase in mortality rate is associated with a change in treatment behavior that can be represented by catheterization: at age 80, there is a reduction in ileoconduit usage and an increase in condom catheter use in male patients (Figure 4C2). The former may reflect shifting surgical thresholds and the latter a risk or comfort bias among treating physicians. Similarly, at age 50, there is a discontinuous decrease in prescription of the analgesic Dilaudid (Figure 4C3), suggesting a greater value placed on pain management for younger patients than for older patients. These value judgements can drive treatment decisions, and biases and heuristics used to make these value judgements are reflected in step changes that lead to discontinuous mortality risks at round number ages. The model used to estimate these risk profiles does not have any algorithmic preference for round numbers, so the repeated discontinuities at round numbers suggest that these changes in risk originate from behavior and treatment.

Blood Urea Nitrogen Patient blood urea nitrogen (BUN) has a non-monotonic impact on mortality (Figure 4D1) with regions of rising and falling risk suggesting there are interactions between underlying biological risk and treatment

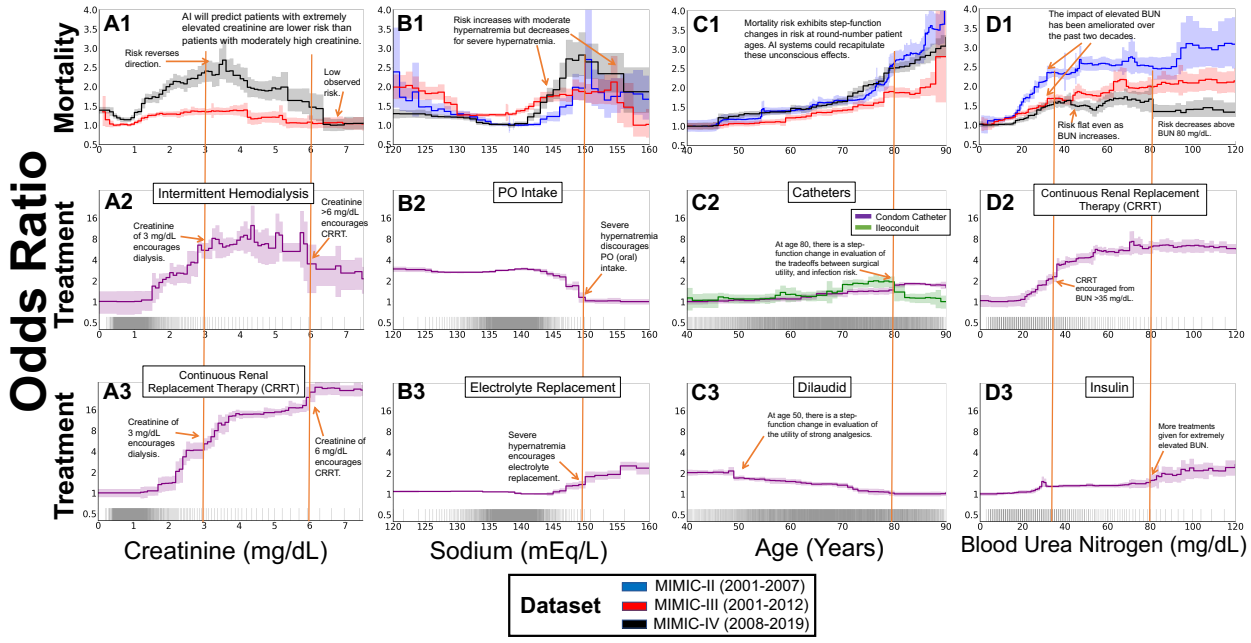


Figure 4: The effects of clinical practice are visible in mortality risk curves for patients in intensive care units, and have changed over time. The three datasets (MIMIC-II, MIMIC-III, and MIMIC-IV) span three decades of intensive care at a single hospital system, allowing comparisons of the effects of clinical practice that have improved over time. The top row shows mortality risk associated with each risk factor after correcting for all other observable factors of patient risk using a generalized additive model (with 95% confidence intervals shaded). The bottom two rows show the treatment propensity associated with each risk factor in MIMIC-IV (with 95% confidence intervals shaded). Ticks along the horizontal axes each denote 10 patients. **(A)** The mortality risk attributable to serum creatinine plateaus at a creatinine level that is only moderately elevated **(A1)**. For extremely elevated creatinine, there is a counter-causal decrease associated with treatment with intermittent hemodialysis **(A2)** or continuous renal replacement therapy **(A3)**. **(B)** The mortality risk attributable to serum sodium decreases at 150 mEq/L **(B1)**, associated with a decreased usage of PO intake (i.e. eating/drinking by mouth, **(B2)**) and an increase in electrolyte replacement (Dextrose 5%, Free Water, Potassium Phosphate, Sterile Water, Sodium Chloride 0.45%)**(B3)**. **(C)** The mortality risk attributable to age increases with discontinuous impacts at 5-year intervals **(C1)**, with a changes in catheterization frequency at age **(C2)** and analgesic prescriptions at 50 **(C3)** demonstrating changing value judgements influencing clinical care as patients progress past round-number ages. The mortality risk attributable to elevated blood urea nitrogen **(D)** exhibits multiple peaks and valleys **(D1)**, with changes in slope linked to propensity for CRRT **(D2)** and insulin prescription **(D3)**. All of these analyses were calculated using a model that is agnostic to round numbers, so the recovery of these effects indicate that the observational data are confounded by real-world treatment decisions that were shaped by round number thresholds. These discontinuities suggest opportunities to fine-tune treatment decisions; in addition, these real-world effects also can confound AI systems to think regions of high intrinsic risk are actually low risk.

effects. As expected, lower levels of BUN are associated with better survival. However, in the elevated regime, we observe several effects: from 35-60mg/dL, risk is flattened and associated with increased CRRT usage (Figure 4D2), and at 80mg/dL, risk is reduced and associated with increased treatments such as insulin (Figure 4D3) in MIMIC-IV. These associations suggest different clinical approaches to patients in these BUN ranges. Over time, as treatment protocols and clinical practice have advanced, the magnitude of these effects has diminished, but has not been eliminated.

Estimating Excess Mortality Attributable to Round Number Biases

With machine learning models that are simultaneously high-resolution and human-interpretable, we can estimate the cost of round number effects. We do this by comparing the observed risk to a locally-flat risk model that smooths out the observed risk to de-confound underlying biological risk from the discontinuous impact of round number biases. For example, in the MIMIC-IV dataset, the mortality rate of patients with serum creatinine 3-6 mg/dL is elevated above the mortality rate of patients with higher serum creatinine >6 mg/dL. If the risk for serum creatinine 3-6mg/dL were no worse than the observed risk for patients with serum creatinine >6 mg/dL, then we would expect a decrease of 91 deaths in the MIMIC-IV dataset of 76,540 hospitalizations with 5,653 recorded deaths (1.6%). Projecting this rate across the 500,000 ICU deaths in the United States per year [17, 18], giving the same risk to patients with *moderately* elevated serum creatinine as is currently given to patients with *severely* elevated serum creatinine could be expected to prevent as many as 8,049 deaths per year in the U.S. alone. This projection is based on the multivariable risk model that corrects for all other risk factors recorded in the MIMIC-IV dataset. While there may be unrecorded factors that would limit the benefit of optimized treatment, the magnitude of this effect suggests that there is an opportunity for improved treatment of patients with moderately elevated levels of creatinine to reduce their risk to that of patients with severely elevated creatinine, or at the very least, to understand the sources of these effects.

Discussion

Increasingly, machine learning is used to estimate and apply evidence-based statistical models of risk. These evidence-based risk models can be more accurate and more reproducible than clinician judgement [19], and offer exciting avenues for refining personalized and precision medicine. There is, however, a fundamental problem in data-driven models that learn to predict risk from observational medical records: patients receive informed interventions with treatment decisions based on patient characteristics. As a result, the observed outcomes are confounded by treatment choices, and hence medical datasets are extremely treacherous: lurking beneath the surface of even large canonical datasets is a complex system of biomedical risk, treatment decisions, patient behavior, and real-world medical systems. Since these confounders are real-world effects, they are robustly supported by large sample sizes and confidently recapitulated by accurate general-purpose models that lack domain-specific grounding to reason about medical context.

We have examined four large datasets and found that the confounding impact of treatment effects are more the rule than the exception. This confounding is not an indictment of the model: treatment effects are real effects, so any accurate risk model trained on these datasets should recapitulate these biases. Furthermore, scale and quality of data collection do not solve this challenge: even datasets that have been collected over three decades and used by thousands of researchers to build data-driven risk models have strong treatment effects that confound data-driven risk models to estimate counter-intuitive effects. Applying such models to guide treatment decisions is risky at best, and could be harmful by incorrectly predicting low risk for patients with high underlying risk. Of all ways for an accurate model to be sub-optimal, incorrectly withholding effective treatments from high-risk patients (due to historical associations with effective treatments) is perhaps the worst offense. Thus, we recommend that clinical understanding be incorporated in all analyses of real-world data, and that black-box machine learning models be used only with great caution.

While these confounding effects are dangerous for blindly applying black-box machine learning models in clinical practice, to the cautious practitioner the confounding effects can in fact be highly beneficial. Treatment effects produce characteristic risk profiles, including non-monotonic and/or discontinuous risk curves, and pinpoint regions in which sub-optimal treatment protocols and clinical decision biases may be revealed. While evaluation of *absolute* risk and the consequences of clinical practice require clinical studies, some evaluation of *relative* risk is possible from observational data alone. In particular, patients with moderately abnormal conditions should have risk no worse than that of patients with more unhealthy conditions; identifying counter-causal instances in which healthier patients actually have higher

risk can identify opportunities for improving clinical practice. All in all, we are encouraged by the potential of glass-box machine learning to contextualize insights from the complex world of real-world evidence and continue to identify opportunities to optimize the practice of personalized and precision medicine.

Materials and Methods

Datasets

Pneumonia The 1989 MedisGroups Comparative Hospital Database (MCHD) pneumonia dataset [20] contains information on inpatients from 78 hospitals in 23 states in the US between July 1987 and December 1988. The MCHD contains over 250 pieces of clinical information that include patient demographic characteristics, history and physical examination findings, and laboratory and radiological results, from which 46 variables were selected [20] with known or postulated association with mortality in patients with community-acquired pneumonia. We used patient data that were collected during the first 48 hr of hospitalization. The tree-based GAMs we trained to predict mortality achieve an AUC of 0.858 on held-out patients, slightly outperforming all other models trained on this dataset [1].

Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) The MIMIC dataset is a high-quality collection of thousands of hospitalizations at the Beth Israel Deaconess Medical Center. This dataset has been provided in several iterations and used by thousands of researchers to estimate mortality risk models. We use three versions of this dataset: MIMIC-II [6] (24,509 hospitalizations), MIMIC-III [5] (27,349 hospitalizations), and MIMIC-IV [7] (76,540 hospitalizations). To standardize the datasets between the versions, we use only patients admitted to an ICU and select only the intersection of patient risk factors available in at least 2 datasets. The tree-based GAMs we trained achieve AUCs of 0.792, 0.756, and 0.740, respectively, on held-out patients, comparable to other models [21].

References

[1] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[2] Sarath C Ranganathan and Samatha Sonnappa. Pneumonia and other respiratory infections. *Pediatric Clinics of North America*, 56(1):135–156, 2009.

[3] Stefan Hegselmann, Thomas Volkert, Hendrik Ohlenburg, Antje Gottschalk, Martin Dugas, and Christian Ertmer. An evaluation of the doctor-interpretability of generalized additive models with interactions. In *Machine Learning for Healthcare*, 2020.

[4] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. *arXiv preprint arXiv:2103.04725*, 2021.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[6] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[7] A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. Mimic-iv (version 1.0. *PhysioNet*, 2021.

[8] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[9] Zijie J Wang, Alex Kale, Harsha Nori, Peter Stella, Mark Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. Gam changer: Editing generalized additive models with interactive visualization. *arXiv preprint arXiv:2112.03245*, 2021.

[10] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, New York, NY, USA, 2013. Association for Computing Machinery.

[11] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

[12] Frederick L Brancati, Joseph W Chow, Marilyn M Wagener, Sara J Vacarello, and Victor L Yu. Is pneumonia really the old man's friend? two-year prognosis after community-acquired pneumonia. *The Lancet*, 342(8862):30–33, 1993.

[13] Marlies Ostermann and Michael Joannidis. Acute kidney injury 2016: diagnosis and diagnostic workup. *Critical care*, 20(1):1–13, 2016.

[14] Sushrut S Waikar and Joseph V Bonventre. Creatinine kinetics and the definition of acute kidney injury. *Journal of the American Society of Nephrology*, 20(3):672–679, 2009.

[15] Qi Qian. Hypernatremia. *Clinical Journal of the American Society of Nephrology*, 14(3):432–434, 2019.

[16] Stanislas Bataille, Camille Baralla, Dominique Torro, Christophe Buffat, Yvon Berland, Marc Alazia, Anderson Loundou, Pierre Michelet, and Henri Vacher-Coponat. Undercorrection of hypernatremia is frequent and associated with mortality. *BMC nephrology*, 15(1):1–9, 2014.

[17] Albert W Wu, Peter Pronovost, and Laura Morlock. Icu incident reporting systems. *Journal of critical care*, 17(2):86–94, 2002.

[18] Derek C Angus, Amber E Barnato, Walter T Linde-Zwirble, Lisa A Weissfeld, R Scott Watson, Tim Rickert, Gordon D Rubenfeld, Robert Wood Johnson Foundation ICU End-Of-Life Peer Group, et al. Use of intensive care at the end of life in the united states: an epidemiologic study. *Critical care medicine*, 32(3):638–643, 2004.

[19] KA McKibbon. Evidence-based practice. *Bulletin of the medical library association*, 86(3):396, 1998.

[20] Gregory F Cooper, Constantin F Aliferis, Richard Ambrosino, John Aronis, Bruce G Buchanan, Richard Caruana, Michael J Fine, Clark Glymour, Geoffrey Gordon, Barbara H Hanusa, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial intelligence in medicine*, 9(2):107–138, 1997.

[21] Fuhai Li, Hui Xin, Jidong Zhang, Mingqiang Fu, Jingmin Zhou, and Zhexun Lian. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the mimic-iii database. *BMJ open*, 11(7):e044779, 2021.