Robust Predictions With Observational Data

Robust Predictions with Observational Data

A dissertation presented

by

William Yuan

to

The Department of Medical Sciences

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Science

Harvard University

Cambridge, Massachusetts

April 2020

Dissertation Advisor: Isaac Kohane William Yuan

Robust Predictions with Observational Data

Abstract

Data science, as currently practiced, is an awkward fit for studying biology or medicine, which currently exist in a state where causal mechanisms to explain many of our observations are often unavailable. While mechanistic deductions are possible in narrow, well defined areas (signaling pathways, binding and protein folding, etc.), a deterministic, internally consistent model of human physiology is still far off. Consequently, the field has developed to serve two purposes simultaneously: both to construct such a framework, but also to help patients in the present with the incomplete information that we have access to. Modern data scientists and researchers utilize massive datasets to attempt to extract insights from a highly complex, largely mysterious system. Given the implications that research recommendations can have on physician behavior, and acknowledged missingness in our understanding, ensuring the reliability and validity of our methods is of paramount importance.

The rise of statistical learning and large datasets has led to significant optimism regarding the ability of such models to influence or even make predictions about patient outcomes. However, constructing inductions that can fit into the otherwise deductive medical and scientific frameworks can be a fraught process. I examine how such work can be framed so as to resultant predictive models "useful" to both clinicians and scientists, and suggest methods for this that can exist within existing research frameworks. In particular, I examine three cases in detail. First, I

describe the basis and implications of temporal bias for the first time, a flaw present in a ubiquitous study design that prevents reliable predictions of the future. Next, I describe knowledge parasitism, a phenomenon where machine learning models piggyback off of the decisions and expertise of clinicians, making their predictions consequently less likely to extend beyond what a clinician may already suspect. Finally, I describe the tendency for propensity matching to "launder" bias in surgical studies, acting to conceal overlooked biases and introduce new biases, reducing the confidence and applicability of the findings.

Acknowledgements

I depended on the support of so many during graduate school- to everyone, friends, family, and mentors, who supported me during this journey, thank you!

Most importantly, I want to thank my parents for their love and support and for inspiring me to set off on my academic pursuits in the first place. They have been the one constant in my life and I could not have made it this far without their encouragement and sacrifice. I am very lucky that they have been there when I needed them.

I want to express my gratitude to my thesis advisor, Dr. Isaac Kohane. I've learned so much about the art of research from him, as well as the way in which data scientists ought to view the world. Zak truly helped me gain my confidence and footing as a researcher and the chance to learn from him has been the opportunity of a lifetime. Not only did he mentor and support me during this journey, he also created a fantastic environment where so many could grow and blossom.

I want to thank all of the members of Zaklab, for all their insight, discussions, and experience. Zak's eye for talent has remained strong, and that's something everyone should be proud of. In particular, I want to thank Dr. Brett Beaulieu-Jones for all his support over a countless number of projects, Dr. Kun-Hsing Yu for his seemingly infinite expertise, Dr. Gabriel Brat for his advice and reminders about the real-world impacts of our projects, and Dr. Scott Lipnick for his patience and optimism. I would also like to thank Dr. Nathan Palmer for all he does to make our work possible, as well as Samantha Lemos, who has done an amazing job of keeping our lab running.

I would also like to thank my committee members, Dr. Paul Avillach, Dr. Tianxi Cai, and Dr. Lee Rubin for their advice and support guiding my progress. I would like to thank Dr. David

Glass for discussions that helped me shape the framework of my thesis. I would like to thank Dr. Anne-Marie Wills, Dr. Fatima Stanford, and Dr. Charles Cook for their domain expertise and guidance of our otherwise wayward analyses.

Finally, I would like to acknowledge the generous financial support for my work from the NVIDIA Graduate Fellowship and the Biological and Biomedical Sciences program.

Citation of Previous Work

Chapter 1 is under review as the following manuscript with changes: Yuan W, Beaulieu -Jones BK, Yu KH, Palmer N, Lipnick S, Cai T, Loscalzo J, Kohane IS, Temporal Bias: Preventing Reliable Predictions of the Future

Chapter 2 has been submitted for review as the following manuscript with changes: Yuan W, Beaulieu-Jones B, Krolewski R, Palmer N, Veyrat-Follet C, Frau F, Cohen C, Bozzi S, Cogswell M, Kumar D, Coulouvrat C, Leroy B, Fischer TZ, Sardi SP, Chandross KJ, Rubin L, Wills AM, Kohane I, Lipnick SL, Characterizing prediagnostic Parkinson's disease and predicting onset in a clinically useful manner

Chapter 3 has been submitted for review as the following manuscript with changes: Yuan W*, Beaulieu-Jones B*, Brat G, Kohane I, Machine Learners as Knowledge Parasites (*equally contributing authors)

Chapter 4 has been submitted for review as the following manuscript with changes: Beaulieu-Jones B*, Yuan W*, Ruffin M, Beam A, Weber G, Brat G, Kohane I, Machine Learning for Patient Risk Stratification: Standing on, or looking over, the shoulders of clinicians? (*equally contributing authors)

Chapter 6 is an expansion upon Yuan W, Cook C, Brat G, Confounding in Retrospective Studies of REBOA, JAMA Surg. 2019;154(12):1167. doi:10.1001/jamasurg.2019.2744

Chapter 7 was previously published as the following manuscript with changes: Yuan W, Yu KH, Palmer N, Stanford FC, Kohane I., Association of Bariatric Surgery with Subsequent Depression Diagnosis, Int J Obes (Lond). 2019 Apr 30. doi: 10.1038/s41366-019-0364-6.

Table of Contents

Introduction

Recent advances in the availability of observational healthcare data have coincided with rapid developments in statistical techniques for processing such data. There has consequently been a significant amount of optimism regarding the ability of "big data" to help guide clinical practices or explore disease etiology (Belle *et al.*, 2015). The British philosopher John Stuart Mill proposed that a scientific framework or theory should primarily be evaluated through its ability to predict future events (Mill, 1843). The contributions of models or studies built on large observational datasets typically take the form of implicit (observed associations) or explicit (risk stratification/disease risk models) predictions about the future. However, there are multiple epistemological and societal flaws that make big data an awkward lens with which to study biology or medicine. These include:

- The epistemological tension between the methods of big data and the goals of medicine and biology.
- The social utility of mechanisms in medicine.
- The societal treatment of big data methodologies.

These will be discussed in turn.

Inductive Methods and Deductive Ends

Tension exists between the inductive methods used to study observational data and the deductive ends that medicine and biology strive for. Most biological sciences aim to generate mechanistic frameworks, which can then be used to derive deductions. The diagnostic process in medicine has also been described as largely deductive (Heneghan *et al.*, 2009). In contrast, the specific brand of induction utilized by modern statistical learning has the potential to highlight

induction's particular weaknesses, while also ignoring principles of experimental design learned from more traditional studies.

Inductive logic is defined as based on "evidential support," and is utilized when the premises provide only partial support to the conclusions. This is in contrast to deductive logic, where the premises provide total support (Hawthorne, 2018). The primary challenge for researchers across times and domains has been to evaluate the situations where the premises fall short and to rationalize the extension of a conclusion to untested waters. While the act of using inductive reasoning and experimental observations to develop deductive frameworks is not new, biomedical researchers gradually refined a balanced approach, by creating restricted domains where deduction can function, and gradually extending these domains with inductive experimentation or observation when warranted (Glass and Hall, 2008). An example of this process is highlighted in the global response to the recent COVID-19 outbreak. Certain public health advice was immediately deduced given previous background knowledge regarding other coronaviruses. For example: heating coronaviruses to a certain temperature renders them inert, COVID-19 is a coronavirus, therefore, heating food to that temperature minimizes infection chance (ANSES, 2020). However, certain properties unique to COVID-19, such as its ability to transmit from asymptomatic individuals, were initially unknown and had to be induced from multiple reliable observations (McIntosh, 2020). Experimental mainstays such as system validation, repeatability, the idea of controls, and the use of domain knowledge in guiding experiments have been critical in bridging the gap between inductive and deductive modes of reasoning (Glass, 2010).

These safeguards on induction, painstakingly built up through hundreds of years of experimentation, have been released with the introduction of machine learning and statistical

tools that serve to accentuate the inherent weaknesses of induction. These include, i) the ability for machine learning/AI techniques to quickly and efficiently (but indiscriminately) induce associations from data, ii) the perceived lack of bias of methods founded on massive data sets or data driven methods, and iii) the hype-driven environment produced by traditional publication biases. In these spheres, the role of the scientist to induce from data has been completely removed. Instead, they are left only to design experiments and interpret the outcomes, both processes fraught with selective pressures.

When dealing with large, complex, high-dimensional datasets, the parable of the blind men and the elephant is particularly illuminating. Often, the totality of the datasets of interest are beyond the comprehension of individual researchers. While researchers were previously limited to extracting linear or combinatorial relationships due to the limits of human conception, "black-box" statistical techniques have enabled comprehensive sets of associations and correlations to be drawn in a manner that is very difficult to both interpret and challenge on methodological grounds.

The size of the datasets used has another pernicious side effect: because constructing a dataset is an expensive and labor-intensive process, researchers analyzing a given dataset are rarely the ones to assemble it. As such, they are often left in the dark regarding hidden biases or assumptions that are baked into the data, biases that consequently influence the model that is created. David Hume's famous critique of induction was predicated on a rejection of the Uniformity Principle: the assumption that identical conditions necessarily produce identical results (Hume, 1739). In the land of observational data, the equivalent principle, the assumption that an identical inference would be uncovered in an alternate dataset collected with an identical methodology, can no longer be assumed to hold. A modern Humean might take the extreme

position that a physical law derived from experiment could fail to function in the future. In contrast, concerns about generalizability are very relevant in observational studies. Data collection processes are subject to myriad intricacies and subtleties that are not perfectly understood, and external validation datasets are increasingly perceived as a luxury rather than a requirement. The lack of repeatability and predictability between models and predictions held up as the nightmare scenario by Humean proponents is in fact a lived reality by data scientists. Despite a tsunami of research, very few big-data or machine learning based models have been deployed due to concerns about generalizability (Rajkomar, Dean and Kohane, 2019; Topol, 2019). Researchers have little conception of how well their datasets represent the wider world or the domain of interest. In this environment, hypothesis validation, let alone identification of causal relationships, becomes incredibly difficult.

(Lack of) Mechanisms

By treating big data methodologies as an inductive process, it follows that mechanisms are a critical area of concern. Hume's description of the problem of induction is again helpful (Hume, 1739): because induction is defined by the incomplete knowledge base used to draw conclusions, how can one be certain that counterexamples do not exist in the unobserved regions of knowledge? The rise of largely uninterpretable statistical modelling has provided an unprecedented opportunity for both inadvertent experimental error and outright fraud, all facilitated by the lack of mechanism.

One of the clearest historical examples of the impact that mechanism can have is the manner in which the original cure for scurvy was lost to medical science. Even when the correct question is asked and the correct experiment is conducted, the findings can lose all utility if the

underlying mechanism is wrong or unchallenged. James Lind is famous for having conducted the first clinical trial on scurvy-afflicted sailors and discovering that citrus fruit was an effective cure. However, more than 150 years after this landmark experiment, Robert Scott's *Discovery* expedition to the Antarctic was famously struck by scurvy, despite awareness of the disease and ample (but misguided) preparations (Baron, 2009).

Although Lind's experiments were powered to detect the curative ability of citrus fruits, the mechanism that was proposed to explain their antiscorbutic properties was insufficient: the vitamin model had not yet been proposed. By the 19th century, improved technology shortened naval voyages such that sailors were no longer at risk of scurvy at all while at sea. Consequently, the impacts of a shift in British Admiralty policy mandating the use of fresh lemon juice (an effective, but expensive antiscorbutic) to the use of processed lime juice (a cheaper, and largely ineffective one) were not noticed (Ceglowski, 2010). At the time of the *Discovery* expedition, skepticism had arisen regarding the power of citrus after high profile incidents where lime juice failed to prevent scurvy in other polar expeditions. The leading proposed mechanism involved a toxicity model: bacterial contamination in meat led to the buildup of acids in the blood, resulting in the characteristic symptoms of scurvy. Rather than carrying lemon juice (regarded as a superstition of the past), Scott's anti-scurvy preparations were largely composed of different methods for preserving meat and preventing spoilage, methods that happened to destroy any vitamin C present in the food. Scott's decision was not unreasonable, as the acid intoxication theory was championed by both one of the most distinguished physicians of the time, Almroth Wright, and one of the most famous polar explorers, Fridtjof Nansen. Scott's shock at the presence of scurvy in his crew despite his best efforts thus does not come as a surprise.

During the 150-year period between Lind and Scott, the mechanistic theories regarding scurvy were subjected to little rigorous experimental testing. Wright described only six case studies in support of his acid intoxication theory, and modern re-analysis has concluded that only one likely had scurvy (Baron, 2009; Wright, 1900). The antiscorbutic efficacy of processed lime juice was never contemporaneously validated through experiment, but Lind himself was willing to advocate for similar products (Lind, 1772). Despite the lack of hard evidence, few physicians acknowledged the weaknesses of their theories. While it is easy to be critical of figures from the past, it is important to remember how much additional understanding we have in comparison. While this particular instance of mechanistic blindness was fueled by a combination of neglect and ignorance, we now have study methodologies that accelerate this process for us.

We are similarly faced today with complex, high-dimensional biological systems that we hope to understand. The failure modes of statistical inference typically stem from deviations between the contexts in which inferences are applied, and the datasets from which the inferences are derived (Kyriacou, 2004). In Lind's case, the gap in understanding occurred between the antiscorbutic power of fresh lemons, and the corresponding lack of power in processed limes. One famous modern example involved the discovery by researchers that a machine learning algorithm for detecting malignant lesions was simply looking for the presence of rulers next to the lesions, which were only placed by dermatologists next to lesions of concern (Schlessinger *et al.*, 2019). Without an understanding for why this model outputs particular predictions, its performance would otherwise seem quite strong.

Mendeleev's famous prediction regarding the existence of particular undiscovered elements, or Le Verrier's prediction of an additional body (Neptune) affecting Uranus's observed orbit provide examples for how biology can develop and mature as a field. The critical

component of a prediction that is often neglected is the mechanistic hypothesis that it proves or disproves. The aspirational, long-term goal of research medicine is to deterministically identify patient outcomes under counterfactual scenarios for the purpose of improving care. While this may be out of reach in the short term, researchers and clinicians continue to develop work that they know will eventually become obsolete because this work forms the foundation for the 'semi-deterministic' future observed in the development other fields. It is therefore important to remember that inductive research methods are a self-defeating tool, where their primary goal is to advance knowledge and understanding to the point where induction from observation is no longer necessary.

Social Perceptions of Data

Despite their mechanistic limitations, machine learning and data-driven techniques occupy a privileged position in society. The perception of these methodologies as infallible or free from bias stems from their impenetrability as well as the high technical requirements needed to truly understand their internal processes. Given this position, burdens of proof for predictions made in these ways ought to be held to a higher standard. Justifications for conclusions such as "the data speaks for itself" or "numbers are numbers" are defeatist in two ways: i) they concede that the system at hand is beyond comprehension and ii) they indicate that the authors do not have enough confidence in their predictions to see them actually deployed. Predictions based on observations can be overturned by a sufficient number of opposing observations, leaving nothing behind, while the falsification of a hypothetical mechanism still contributes to our understanding about a system. Even from a purely aspirational perspective, researchers and scientists should

hold their work to higher standards. Carefully defining where and when such models or results are appropriate is critical in both managing expectations and limiting damage.

More practically, a higher emphasis on identifying appropriate context can help avoid unintended consequences. The roles of domain experts in guiding the training, development, or deployment of such models is often marginalized, often due to the false confidence fostered by the uncritical acceptance of data-driven predictions. The recent entry of the Apple Watch as an automatic monitoring and detection system for atrial fibrillation (AF) is an illuminating example regarding the utility of risk predictions that are separated from the guidance of physicians. Traditional determinations of AF, rendered by a physician, were the result of a specific workflow by three factors: i) the self-selection of patients who are symptomatic to consult a physician, ii) the expertise of the physician to order the right tests and diagnose the condition and iii) the judgement of the physician regarding whether the patient presentation is significant. These criteria define a very narrow, specific set of patients and conditions that could be termed "clinically relevant" AF (Husten, 2019). Consequently, AF criteria are only applicable to those cases where a physician has already made a judgement call. The Apple Watch, in contrast, detected every unfiltered instance of "AF." Without the implicit guidance of physicians to select the most clinically relevant cases of AF, the Apple Watch was consequently affected by significant numbers of false positives (Loftus, 2019; Perez *et al.*, 2019).

Due to the direct impacts that biomedical research has on patient outcomes and popular perceptions of health, it is critically important for us to avoid constructing "soaring efface(s)" (Ceglowski, 2010) upon small foundations of evidence, something that is tragically easy to do with modern datasets and methods. Two high profile controversies emphasize this: the Fleischmann-Pons experiments into cold fusion (Fleischmann and Pons, 1989), which were

ultimately discredited (Miskelly *et al.*, 1989), kicked off a flurry of misguided research activity, but had little impact on the wider public. In contrast, Wakefield's fraudulent study claiming an association between the MMR vaccine and autism (Wakefield *et al.*, 1998) sparked a public health crisis that has yet to be resolved (Flaherty, 2011). The public trust in scientists' findings is critical for further advancements, but is also a precious resource.

The task of generating robust predictions with observational data is made significantly more difficult by the tensions present between the methodologies of big data and the goals of research biology and medicine. Cavalier use of these tools has the potential for significant social harms, due to the inferential power that these tools have, the privileged societal position they occupy, and the direct impacts that the biomedical field has. In this work, I investigate three pitfalls that prevent predictions from reliably informing clinical practice or disease etiology.

In Chapters 1-2, I describe **Temporal Bias,** a flaw present in many observational studies and models that acts to accentuate differences between case and control cohorts, resulting in exaggerated effect sizes and prediction accuracies relative to prospective clinical deployment. I then describe the technical and study development of a clinically deployable model for Parkinson's disease with temporal bias in mind.

In Chapters 3-4, I describe the tendency for machine learning models to act as **Knowledge Parasites,** and the implications in practice of models that may not be able to truly infer beyond what a clinician already knows about a patient.

In Chapters 5-7, I describe the phenomenon where the use of statistical techniques results in **Laundering Bias** in the surgical literature, where the assumptions and limitations of studies are obscured. I describe a series of recommendations for avoiding this, and implement them in observational study examining the risks of bariatric surgery.

Chapter 1: Temporal Bias: Preventing Reliable Predictions of the Future

Chapter Introduction

This chapter introduces the basis for and impacts of **Temporal Bias**, a flaw that stems from the disconnect between the settings in which models are trained and the settings in which they are truly applied in practice. The utility of observational data towards any medical or biological application is contingent on collecting the proper data in the first place, something that temporal bias prevents. This chapter was adapted from a manuscript aimed informing a wide audience about the impacts that this bias could have, ranging from prediction errors to replication failures.

Main Text

The ability to predict future events is one of the defining features of science as a discipline (Platnick and Popper, 1977). Case-control studies have become one of the main tools for predicting events or defining associations using observational data (Song and Chung, 2010). The essence of this study design is to define two populations that differ by some characteristic of interest (helpfully termed "case" and "control") and to measure differing exposures between these groups. In this way, exposures can be interpreted as "predictors" or "risk factors" for "case" status, but are not guaranteed to be causal (Lewallen and Courtright, 1998; Marshall, 2004). With the proliferation of observational datasets and novel machine learning techniques, these studies have expanded into fields where risk prediction is a central focus (Weiss *et al.*, 2012). However, we have identified a structural flaw, seen widely in basic case-control study designs, which we call "temporal bias." At its core, temporal bias represents a failure to collect information along the entire control-to-case trajectory at hand. This temporal bias not only

amplifies reported effect sizes relative to what would be observed in practice, but also obfuscates the prospective use of models or identified risk factors.

A classical example of temporal bias and its impacts can be seen through the initial discovery of lyme disease, a tick-borne bacterial infection. Lyme disease is characterized by i) an initial bite, ii) an expanding ring rash, and iii) arthritic symptoms, in that order (Steere *et al.*, 2016). However, the original 1976 discovery of lyme disease (then termed "lyme arthritis") focused exclusively on patients who manifested with arthritic symptoms (Steere *et al.*, 1977). This enabled researchers to definitively identify the prognostic value of a ring rash towards arthritis, but not tick bites, due to the latter symptom's temporal distance from the researcher's focus. By focusing on predictive features immediately prior to the event in question, researchers capture a biased representation of the full trajectory from healthy-to-diseased. A contemporaneous doctor cognizant of "lyme arthritis" and presented with a patient presenting with a tick bite would miss the possibility of disease until further symptoms developed. Similarly, a predictive model for lyme arthritis focused on ring rashes would report numerous false negatives if it were deployed in practice: patients who had yet to develop ring rashes would contract arthritis at a future time. These errors stem from the incomplete picture of symptoms that was captured.

However, temporal bias is not a problem of the past. The central flaw, an overemphasis on features collected near the case event, still occurs in the literature today. Within the medical domain, there are numerous examples of temporal bias in both clinical medicine and machine learning (Rand *et al.*, 1985; Himes *et al.*, 2009; X. Wang *et al.*, 2014; Chou *et al.*, 2016; Ranganath *et al.*, 2016; Choi *et al.*, 2017; Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium for Early Detection of Lung Cancer *et al.*, 2018; Norgeot *et al.*,

2019). Given the ease and intuitive nature of the case-control study, temporal bias is also not restricted to medical applications. We have identified instances of temporally biased experimental design in areas as diverse as operations research (data center and hard drive failures (Murray, Joseph F, Hughes, Gordon F and Kreutz-Delgado, Kenneth, 2005; Y. Wang *et al.*, 2014; Botezatu *et al.*, 2016; Queiroz *et al.*, 2017; Lin *et al.*, 2018)), meteorology (severe weather prediction, (Marzban and Stumpf, 1996; McGovern *et al.*, 2011)), and political science (identifying factors behind failed states, (King and Zeng, 2001; Goldstone *et al.*, 2010)). As algorithms trained using large datasets and advanced machine learning methods become more popular, understanding biases in the way they were generated is critical. Despite increasing interest in machine learning risk prediction, few tools for use on individual patients have become standard practice (Rajkomar, Dean and Kohane, 2019; Topol, 2019). As evidence-based medicine becomes the norm, a clear picture of the quality of the component studies is required. In this article, we describe the basis for temporal bias and examine three representative instances of temporal bias in the medical, machine learning, and nutritional literature to identify the impact that this phenomenon has on observed effect sizes, predictive power, and experimental reproducibility.

Background

Of interest are the expansive set of studies that focus on predicting future events and obey the following general conditions:

1: Events to be predicted take the form of state transitions (healthy-to-diseased, stable-to-failed, control-to-case, etc.). This implies that there exists a bulk population of controls, from which cases differentiate themselves. Soon-to-be cases progress along a trajectory away from the control population. This trajectory terminates at the occurrence of the case event.

2: Risk-of-event is equivalent to measuring progress along a control-to-case trajectory in time. Because risk prediction utilizes features from the present to assess the chance of a future event occurring, an event that is truly random would not be an appropriate domain for either a case-control study or a risk prediction algorithm. The trajectory can be thought of as the ground truth progression along a pathway towards the event in question, and are defined relative to the specific populations chosen for the study. This assumes that the researchers have taken the exchangeability (Hernan, 2006) of their case and control populations into account: if members of the control population are chosen poorly and cannot experience the case event, then there can be no trajectory.

3: At the population level, the trajectory commences when the to-be-diseased population first begins to diverge from non-diseased population, and reaches a maximum when the disease event actually occurs. This requires that the trajectory is aligned to the event in question. Diseased individuals must consequently referred to using terms such as “days to disease,” while control individuals exist in an undefined point along this timeline. This is

only required due to the retrospective nature of these studies, and is a major departure from prospective deployment.

4: Features actually measured are proxies for an individual's position along the trajectory. Regardless of their positive or negative association with the event, features subject to temporal bias will tend to diverge between cases and controls with a continuous trajectory, and become better at differentiating the controls from cases as case individuals get closer to their event. For example, genome-wide association studies (GWAS) take the form of case-control studies that treat genomic variants as the exposure. Because these variants are constant with respect to time, GWAS studies are not subject to temporal bias.

As a result, we can distill prediction studies into a common structure (Figure 1.1): the members of the diseased population begin as controls at a point in the past, and progress along a trajectory until the disease occurs. Most case-control studies apply a dichotomous framework over this continuous trajectory.

Temporal bias occurs when cases are sampled unevenly in time across this trajectory (Figure 1.1B). (A full theoretical basis for temporal bias is presented in the Methods section.) Note that this is a separate but analogous effect compared to selection bias: the control population may be exchangeable with the diseased population, but must tautologically exist at a prior point along the disease trajectory compared to cases. Rather than operating over the selection of which patients to include in the study, temporal bias acts over the selection of when each subject is observed.

This important temporal feature yields two implications:

1. If the features of diseased subjects are evaluated based on a point or window that is defined relative to the case event (a future event, from the perspective of the feature measurements), features in the end of the trajectory will be oversampled.
2. The resulting model cannot be prospectively applied because the study design implicitly leaked information from the future: a prospective evaluator has no way of knowing if a particular subject is within the observation window defined by the study.

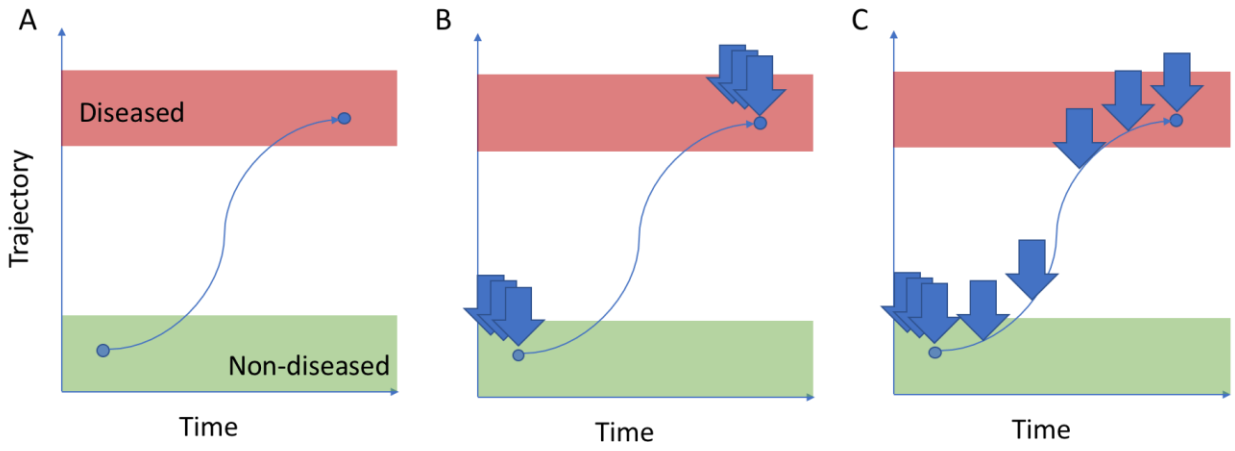


Figure 1.1 Prospective risk can be represented as a trajectory. A) The (single-class) case-control paradigm often imposes a dichotomous (binary) framework onto a continuous trajectory. **B)** Experiments utilizing observations of cases that are concentrated at the time when the case event occurs cannot capture any information regarding the transition trajectory, resulting in temporal bias. **C)** In order to predict a patient's position along the trajectory, experiments capturing the entire transition from non-diseased to diseased are necessary.

Temporal bias is intuitively understood within certain epidemiological circles- in fact: recall bias, caused by the tendency for survey respondents to remember recent events at a higher rate relative to past events, can be interpreted as a specific instance of temporal bias. Similarly, it is understood that case-control studies represent a lower "level of evidence" relative to other study designs (Burns, Rohrich and Chung, 2011). Methodologies have been proposed that, while

not explicitly designed to address temporal bias, are immune to it (density-based sampling, among others (Rothman, 2012)). However, these tend to focus on point exposures or necessitate impractically exact sampling strategies. Despite this important shortcoming, the ease of the case-control framework has allowed temporal bias to proliferate across many fields. We examine three representative examples below.

Temporal Bias Can Inflate Observed Associations and Effect Sizes

The INTERHEART study (Yusuf *et al.*, 2004) examined the association between various risk factors and myocardial infarction (MI) using a matched case-control design among a global cohort. Individuals presenting at hospitals with characteristic MI were defined as cases, and subjected to interviews and blood tests, while matched controls were identified from relatives of MI patients or healthy cardiovascular individuals presenting with unrelated disorders. One risk factor of interest included lipoprotein (a) [Lp(a)], a blood protein (Jacobson, 2013; Hippe *et al.*, 2018). While Lp(a) levels are thought to be influenced by inheritance, significant intra-individual biological variance with time has been reported (Garnotel *et al.*, 1998; Nazir *et al.*, 1999).

One particular analysis utilized data from this study to examine the positive association between blood levels of Lp(a) and MI across different ethnicities and evaluate the possible efficacy of Lp(a) as a risk prediction feature (Paré *et al.*, 2019). However, because cases were only sampled at the time of the MI event, the resulting effect sizes are difficult to interpret prospectively. Indexing case patients by their case status leaks information regarding their status to which a physician prospectively examining a patient would not have access. Intuitively, if Lp(a) was static until a spike immediately prior to an MI event, it could not be used as a prospective risk predictor, even though a significant association would be observed given this

particular experimental design. This limitation cannot be overcome using only the data that was collected, as information regarding the dynamics of Lp(a) over time is completely missing. To evaluate the influence of temporal bias, we estimated the size of the Lp(a)-MI association had the experiment been done prospectively. This analysis was done by simulating control-to-case trajectories using INTERHEART case/control population Lp(a) distributions by imputing the missing data. We conducted extensive sensitivity testing over different possible trajectories to evaluate the range of possible effect sizes. This approach allowed for the recalculation of the association strength as if the study had been conducted in a prospective manner from the beginning.

Table 1.1A summarizes the observed effect size in the simulated prospective trials compared to the reported baseline. In all cases, the simulated raw odds ratio between Lp(a) and MI was significantly lower than the observed raw odds ratio due to apparent temporal bias present in the latter measurement. This is an intuitive result, since case individuals as a group will be more similar to controls (healthier) when sampled at random points in time rather than when they experience an MI event (Figure 1.1B). Although it cannot be definitively proven that prospective effect sizes would be smaller, as this would require longitudinal data that do not exist, this exercise provides a valuable insight. The degree of temporal bias scales with the area under the imputed trajectory: linear and logistic trajectories were more susceptible to temporal bias than logarithmic trajectories. In order to observe the full reported odds ratio, the underlying trajectory would need to resemble a Heaviside step function in which cases spontaneously experience a spike in Lp(a) levels at the point of their divergence from the controls, an assumption that is neither explicitly made in the study nor has a basis in biology. We repeated the imputation process with Heaviside step functions, varying the position of the impulse in the

trajectory (Table 1.1B). As the impulse location approaches the beginning of the trajectory, the effect size relative to the baseline approaches 1. This observation illustrates the assumption intrinsic in the original INTERHEART experimental design: that MI individuals had static risk profiles during the runup to their hospitalizations.

To characterize these findings in a real-world dataset, we examined the Lp(a) test values and MI status of 7,128 patients seen at hospitals and clinics within the Partners Healthcare System - representing Brigham and Women's Hospital and Massachusetts General Hospital among others - who had indications of more than one Lp(a) reading over observed records. This dataset included 28,313 individual Lp(a) tests and 2,587 individuals with indications of myocardial infarction. We identified significant intra-individual variation in Lp(a) values in this population: the mean intra-individual standard deviation between tests was 12.2 mg/dl, compared to a mean test result of 49.4 mg/dl. In this dataset, biased Lp(a) measurement selection among case exposure values led to inflated association strength between Lp(a) and MI by 37.2% (Table 1.1C) relative to random selection. This is a conservative estimate- we would expect the deviation to increase upon correcting for ascertainment bias in the dataset. “Control” individuals would be less cardiovascularly healthy than true controls, while “cases” would typically not be sampled immediately prior to an MI, and consequently appear to be healthier than INTERHEART cases.

Table 1.1. The observed Lp(A)-MI association is magnified by temporal bias. (Table continues on next pages) A) Association effect sizes from simulated prospective trials relative to INTERHEART sizes. Effect sizes less than 1 represent smaller simulated effects compared to those from INTERHEART. B) As the imputed trajectory approaches an idealized step function where the impulse appears earlier in the trajectory, the effect size approaches the observed baseline. C) Association effect size sensitivity to variance in real-world continuous Lp(a) observations.

A)

Initial Case State Imputation Method	Trajectory Type	Effect Size relative to Reported Baseline (95% CI)
Weighted Sampling	Linear	0.172 (0.160-0.196)
Weighted Sampling	Logistic	0.169 (0.150-0.187)
Weighted Sampling	Logarithmic	0.403 (0.390-0.417)
Percentile Matching	Linear	0.518 (0.507-0.528)
Percentile Matching	Logistic	0.517 (0.506-0.527)
Percentile Matching	Logarithmic	0.639 (0.631-0.647)
Percent Shift	Linear	0.389 (0.376-0.401)
Percent Shift	Logistic	0.386 (0.373-0.399)
Percent Shift	Logarithmic	0.539 (0.530-0.549)

B)

Initial Case State Imputation Method	Heaviside Step Function: Impulse Location	Effect Size relative to Reported Baseline (95% CI)
Weighted Sampling	First 10% of Trajectory	0.808 (0.801-0.814)
Weighted Sampling	First 1% of Trajectory	0.980 (0.977-0.984)
Weighted Sampling	First 0.1% of Trajectory	0.998 (0.997-0.999)
Percentile Matching	First 10% of Trajectory	0.898 (0.895-0.901)
Percentile Matching	First 1% of Trajectory	0.989 (0.989-0.989)
Percentile Matching	First 0.1% of Trajectory	0.999 (0.999-0.999)
Percent Shift	First 10% of Trajectory	0.860 (0.857-0.864)
Percent Shift	First 1% of Trajectory	0.987 (0.985-0.989)
Percent Shift	First 0.1% of Trajectory	0.999 (0.999-0.999)

C)

Lp(a) Selection Method	Normalized Lp(a)-MI Coefficient	p-value
Largest Available Lp(a)	1.372	< 2E-16
Smallest Available Lp(a)	0.519	9.08E-4
Random Lp(a)	1	6.34E-12

Prospective Prediction Failure due to Temporal Bias

As the availability of observational data has skyrocketed, event prediction has become a popular task in machine learning, particularly in medical domains. The resulting models are typically intended to act as risk evaluation, serve in triage, or assist the understanding of disease etiology. Because of this focus on prediction, many methods utilize the idea of a prediction window: a gap between when an event is observed and when features are collected. For example, a model that differentiates patients six months prior to MI onset from healthy matched controls may be said to "detect" MI six months in advance. However, because the window is defined relative to a case event, it represents an uneven sampling of the disease trajectory. This finding means that, in the vast majority of cases, this idea shares the same temporal bias as sampling cases at the time of event. While it could be argued that this type of model could be deployed prospectively and would still provide information on which patients were 6 months away from an MI, this prediction requires unfounded assumptions regarding the trajectory of MI onset. For example, if the trajectory is such that patients' risk in the years prior to the MI is approximately

uniform, a model trained in this way would provide many false positive 6-month MI predictions. This outcome can be interpreted as a generalization failure to patients more than 6 months away from an MI. Because window sizes are often chosen without respect to the underlying transition trajectory, significant potential for temporal bias still exists, driven by factors such as differential diagnosis periods or missed exposures.

To illustrate fundamental issues with the use of event-indexed observation windows, we constructed predictors for childbirth that exploit the intrinsic asymmetry of the observable "risk" trajectory of delivery utilizing insurance claims data. Cases and controls are significantly more difficult to distinguish more than nine or ten months prior to delivery compared to later in pregnancy because the case population is not yet pregnant. Features collected while the case population is pregnant are far more informative regarding delivery status. A case-control study that uses a window defined three months prior to delivery will capture these informative, pregnancy related features. In contrast, a cohort study fixed to examine features collected in January will occasionally be presented with largely uninformative features when the case individual's delivery takes place late in the year (Figure 1.2A). Using 2015 data from an de-identified nationwide medical insurance claims dataset, we simulated three studies:

I. Models are both trained and evaluated under the case-control (CC) paradigm: one month of records, three months prior to the delivery date (cases) or matched baseline date (controls) are used (CC-CC model).
II. Models are trained under the case-control paradigm, but evaluated under the cohort paradigm, where records from January are used to predict delivery in 2015 (CC-Cohort Model).

III. Models are both trained and evaluated under the cohort paradigm (Cohort-Cohort Model).

For each simulated study, records within the observation window of diagnoses, procedures, and prescriptions ordered were fed into both deep recurrent neural nets and logistic regression models.

The significant difference in performance (Figure 1.2B) between CC-CC and CC-Cohort models illustrates a central trait of temporally biased sampling. Uneven sampling across the transition trajectory improves validation AUC under artificial validation conditions, but model performance collapses when deployed in a prospective manner. In contrast, models designed with the prospective task from the outset (Cohort-Cohort) had intermediate performance that reflected the inherent ambiguity of the available observations. These findings were robust across both deep neural networks and logistic regression based models. In fact, while the more complex deep neural network performed better than the logistic regression model for the CC-CC task, it was performed worse than the logistic regression on the CC-Cohort task. In this case, methodological improvements on an unrealistic task led to more significant declines in performance on a more realistic task.

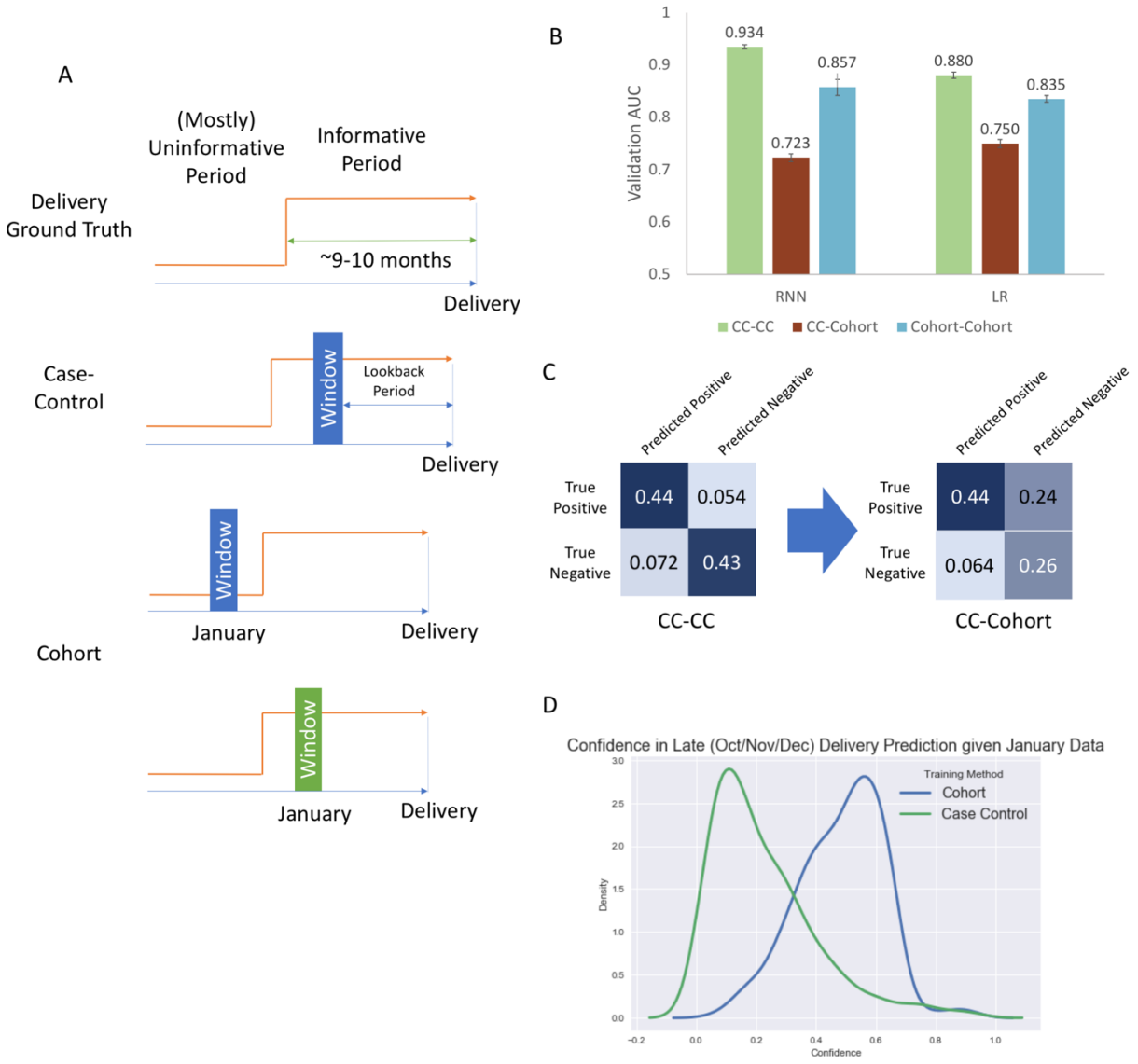


Figure 1.2. Case-control predictors for delivery report false negatives when applied prospectively due to temporal bias. A) The ground truth trajectory for delivery (orange) is composed of parts: an informative period, 9-10 months prior to the delivery, and a largely uninformative period prior. Case-control windows are indexed to delivery/baseline date, and so only sample a single (informative) slice of the trajectory. Cohort windows always occur in January, and so uniformly sample the trajectory. **B)** Model performance (Validation AUROC) for deep recurrent neural networks and logistic regression for each study design. Error bars represent the 95% confidence intervals. **C)** Comparison of confusion matrices for CC-CC (left) and CC-Cohort (right) models. **D)** CC-Cohort validation model confidence distributions for late (Oct/Nov/Dec) deliveries given January features.

For women with October/November/December deliveries, claims data from January are mostly uninformative, and a reliable prediction at that point is not possible at the population level, especially when using features trained during pregnancy. The confusion matrices produced by CC-CC and CC-Cohort models revealed that much of the performance collapse can be traced to false negatives (Figure 1.2C). We examined the confidence that the deep convolutional networks assigned to October/November/December deliveries when evaluated on cohort structured data were predictive (Figure 1.2D). Models trained under a case-control regime incorrectly label these individuals as high confidence controls, while models trained under a cohort regime more appropriately capture the intrinsic ambiguity of the prediction task. Unlike these models, clinicians do not have the luxury of examining only patients three months/six months/one year prior to disease incidence: they must assess risk in real time. Rather than to present a toy example, this is intended to represent the extreme case of the potential consequences of releasing a predictive model trained in this manner.

It is critical to note that this is a problem that cannot be solved methodologically. As evidenced by the comparison of the performance of the deep neural network and logistic regression models, novel or exotic machine learning techniques cannot compensate for the fact that the data fed into the models represent a distorted view of the actual population distribution that would be encountered prospectively. Even with perfect measurement and modelling, temporal bias and the issues that result would still be present: the underlying trajectory would still be unobserved.

A useful analogy exists with the observer effect/Schrodinger's cat in the physical sciences, where the observation of a particle induces the collapse of its wavefunction into a well-

defined state. When a predictor is constructed that "knows" that a certain set of its subjects have experienced a delivery, the set of potential pathways experienced by these subjects similarly collapses to a subset compatible with a delivery. Simply knowing that a subject had a delivery necessitates that the same subject experienced a fertilization event in the past, and implies certain factors regarding the subject's physical health (e.g. fertility, sexual activity). In contrast, a prospective prediction must look into the futures of individuals of every potential pathway, without the guidance of information from the future- a much more difficult task.

Temporal Bias-Induced Replication Failure

The Mediterranean diet (characterized by consumption of olive oil, fruits, vegetables, among other factors) has been implicated as a protective factor against coronary heart disease, but the mechanism for this association is unclear. One paper set out to examine whether olive oil consumption specifically was associated with MI using patients from a Spanish hospital (Fernández-Jarne *et al.*, 2002). MI patients and matched controls were interviewed regarding their olive oil consumption over the past year, and a protective effect against MI was observed among the highest quintile of olive oil consumers. In response, another group analyzed data from an Italian case-control study and were unable to identify the same association between the upper quintile of olive oil consumption and MI (Bertuzzi *et al.*, 2002). Crucially, these analyses differed in the size of the observation window used: one year and two years respectively. As a result, not only were these studies sampling the MI trajectory unevenly, they sampled different parts of the MI trajectory. To examine the degree to which differing amounts of temporal bias present in each study could have contributed to the failure-to-replicate, we utilized longitudinal data from nearly 100,000 individuals from the Nurses' Health Study (NHS) regarding olive oil

consumption patterns and MI to provide a baseline "ground truth." We simulated retrospective case-control studies that considered different "lookback" periods to determine if the presence or magnitude of a protective effect was sensitive to the manner in which an experiment was conducted. Figure 1.3A details the simulation setup: longitudinal records (Figure 1.3A) were used to identify case (red) and control (green) individuals. MI dates and matched baseline dates are chosen for cases and controls, respectively. For each patient, exposures during the lookback time are recorded. The association between MI and the observed exposures were then calculated and the influence of the lookback time on association strength was assessed.

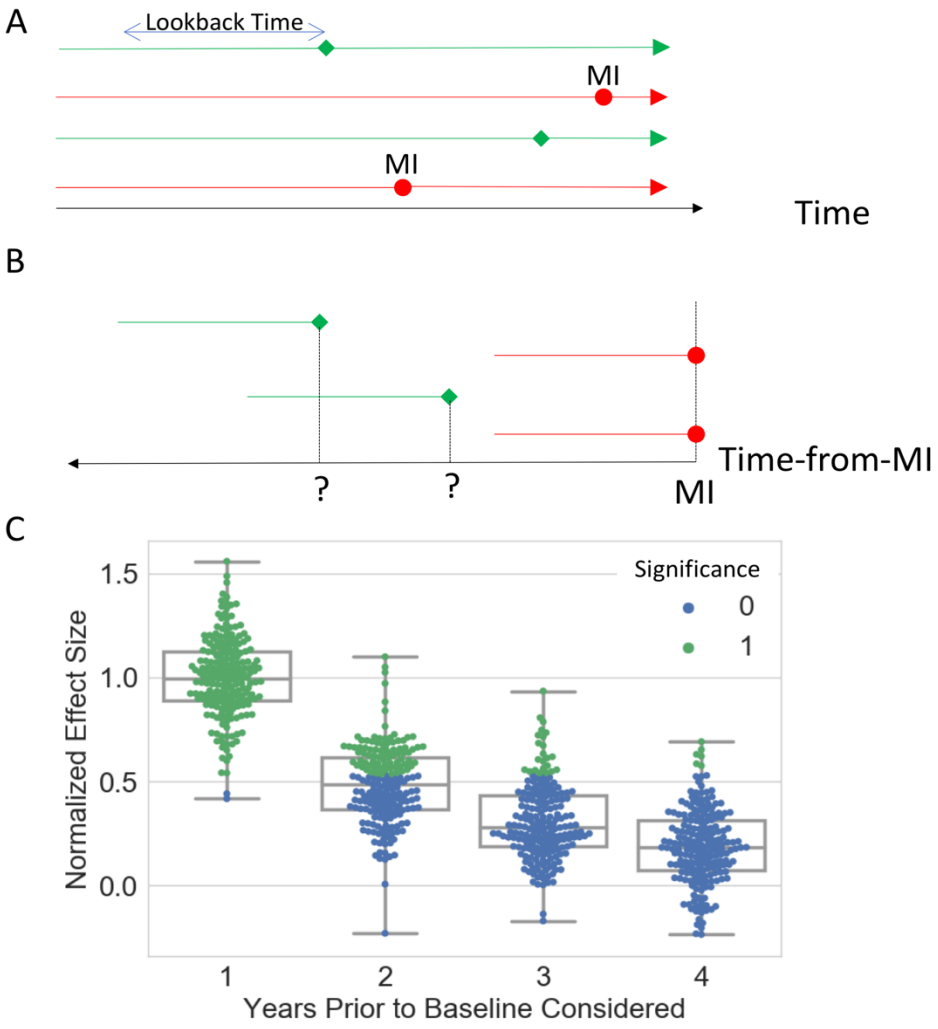


Figure 1.3. Temporal bias results from arbitrary alignment of cases and future-indexed lookback times. A) Over a particular time period, longitudinal data of olive oil consumption is continuous for all cohort members with time. Circles represent MI events, while diamonds represent matched, but otherwise arbitrarily chosen baseline points for controls. **B)** Case-control studies arbitrarily align MI patients at the date of the MI. As a result, the time dimension is inverted and anchored to the MI date, the position of controls is consequently lost. **C)** Strength of olive oil consumption-MI association given years of consumption prior to baseline considered. Effect size is normalized to the average 1-year association strength. Points are colored based on statistical significance after FDR correction.

The simulated studies that examined one year of past olive oil consumption relative to the MI/baseline date detected a protective effect, as originally observed. However, the magnitude and statistical significance of this effect decayed as the size of the lookback period was increased, consistent with the results of the failed replication. When a two-year lookback period was used, only 41% of simulated studies observed a statistically significant result (Figure 1.3C). The observed protective effect in these cases is an artifact of methodology, rather than medicine, physiology, or society. The act of looking back from the MI date/matched baseline has the effect of inverting the time axis to “time-from-MI " and aligning the case individuals (Figure 1.3B). However, no such treatment is possible for control individuals, and their position along the new temporal axis is unknown. As a result, there is no functional basis for comparing healthy individuals to individuals artificially indexed to a future event (MI) because these represent groups that can only be identified retrospectively, after the MI has already occurred. While there may indeed be a prospective association between olive oil and MI, protective or otherwise, these case-control studies were not powered to detect such an effect. Because both olive oil consumption and MI risk are time-varying features, the strength of the instantaneous association between the two will naturally depend on when each feature is measured.

Discussion

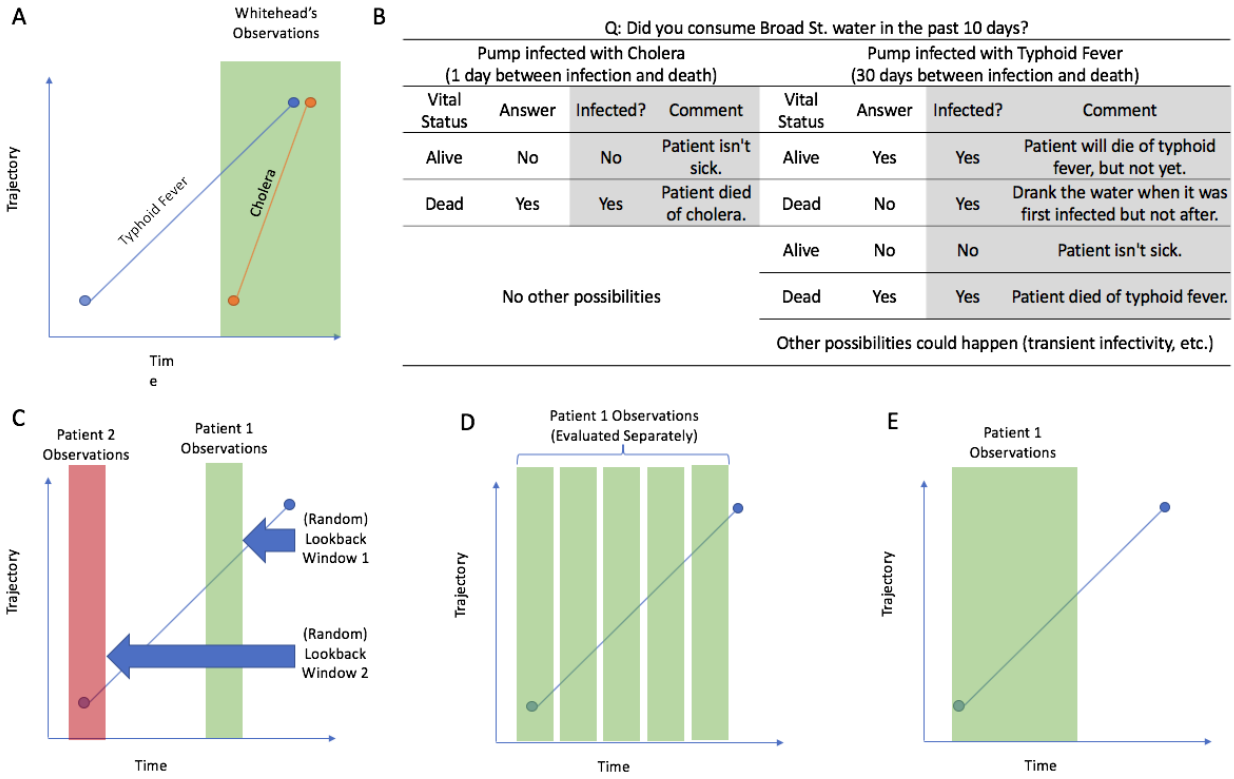
Temporal bias can be thought of as a flaw present in the application of case-control experiments to real-world, prospective applications. Because these experiments do not uniformly sample the control-to-case trajectory, features and observations in certain parts of the trajectory are oversampled and assigned disproportionate weight. Furthermore, because the case observations that are model-applicable can only be identified after the case event actually occurs,

the resulting experimental findings are impossible to use prospectively. Temporal bias serves to amplify differences between the case and control populations, improving apparent predictive accuracy and exaggerating effect sizes of predictors. In prospective cases, it may also result in researchers failing to discover predictive signals that were outside the window considered. Because the magnitude of its effects is a function of an often-unobserved trajectory, temporal bias is poorly controlled for and can lead to replication bias between studies.

The various effects of temporal bias stem from the disconnection between the retrospective case-control experimental protocol and the act of prospective induction or prediction. As a result, any field or domain where prediction, time-varying association, or comparative analysis is important is at risk. Temporal bias is consequently ubiquitous across the literature. Examples in clinical medicine and machine learning (Rand *et al.*, 1985; Himes *et al.*, 2009; X. Wang *et al.*, 2014; Chou *et al.*, 2016; Ranganath *et al.*, 2016; Choi *et al.*, 2017; Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium for Early Detection of Lung Cancer *et al.*, 2018; Norgeot *et al.*, 2019) typically take the form of the examples we have described: studies comparing biomarkers collected at the time of diagnosis to those from healthy controls, or prediction studies that utilize event-based windows. However, we have also identified analogous temporal bias in diverse non-medical domains: data center and hard drive failures (Murray, Joseph F, Hughes, Gordon F and Kreutz-Delgado, Kenneth, 2005; Y. Wang *et al.*, 2014; Botezatu *et al.*, 2016; Queiroz *et al.*, 2017; Lin *et al.*, 2018), severe weather prediction, (Marzban and Stumpf, 1996; McGovern *et al.*, 2011), and identification of factors behind failed states (King and Zeng, 2001; Goldstone *et al.*, 2010)). While the windows or experimental parameters in these studies may have been chosen based on the author's domain knowledge regarding the transition trajectories at hand, analysis on the presence or magnitude of

bias may be illuminating. Several naive examples of sensitivity analyses in hard-drive failure prediction, where the devices themselves collect longitudinal monitoring and performance data, are presented in Supplementary Results.

Temporal bias is not a novel phenomenon. The first documented case-control study in the medical literature was Reverend Henry Whitehead's follow-up (Paneth, Susser and Susser, 2002) to John Snow's famous report (Snow, 1856) on the Broad Street cholera outbreak. Whitehead aimed to evaluate Snow's hypothesis that consuming water from the Broad Street pump led to infection. Whitehead surveyed both families of infected and deceased as well as individuals without cholera regarding their consumption of pump water during the time deaths were observed (Whitehead, 1865; Newsom, 2006).



B

Q: Did you consume Broad St. water in the past 10 days?							
Pump infected with Cholera (1 day between infection and death)				Pump infected with Typhoid Fever (30 days between infection and death)			
Vital Status	Answer	Infected?	Comment	Vital Status	Answer	Infected?	Comment
Alive	No	No	Patient isn't sick.	Alive	Yes	Yes	Patient will die of typhoid fever, but not yet.
Dead	Yes	Yes	Patient died of cholera.	Dead	No	Yes	Drank the water when it was first infected but not after.
No other possibilities				Alive	No	No	Patient isn't sick.
				Dead	Yes	Yes	Patient died of typhoid fever.
				Other possibilities could happen (transient infectivity, etc.)			

Figure 1.4. **Preventing Temporal Bias. A)** Whitehead's cholera study benefited from the short period between infection and death. Had Whitehead been faced with an outbreak of typhoid fever, his sampling strategy would oversample late-stage features. **B)** Hypothetical interview data from Whitehead's case-control study. Lacking underlying knowledge regarding disease etiology, Whitehead's experimental design would have experienced temporal bias given a disease with a longer incubation period. Shaded columns represent information hidden to the investigator. **C)** Randomizing the lookback window among case patients can uniformly sample the trajectory, if the lookback times go far back enough. **D)** Evaluating person-days, person-weeks, or person-months can allow for the entire trajectory to be considered. **E)** Creating a well-defined date from which a "look forward" window is deployed does not uniformly sample the trajectory, but is still prospectively implementable since the starting date can be determined in real time.

The outbreak began on August 31st, 1854 (Snow, 1856), with deaths occurring in the days that immediately followed. Whitehead's subsequent efforts in identifying pump-water exposure among outbreak victims focused on the time period between August 30th and September 8th, corresponding to a lookback time between 1 and 10 days, depending on when the victim died.

This would normally result in temporal bias towards the end of cholera trajectory. Although Whitehead's conclusions were ultimately correct, the comparatively brief incubation period (2 hours to 5 days (Centers for Disease Control and Prevention, no date)) of cholera contributed to the success of the experiment and Whitehead's later ability to identify the index patient. The rapid transition from healthy to diseased ensured that Whitehead's chosen lookback time would have uniformly sampled the disease trajectory. Had Whitehead instead been faced with an outbreak of another waterborne disease such as typhoid fever, which can have an incubation period as long as 30 days (Mintz, Slayton and Walters, 2015), Whitehead's chosen window would oversample exposure status in the runup to death, leading to temporal bias that would overemphasize features in the latter portion of the disease trajectory (Figure 1.4A). Because the disease etiology and trajectory were unknown at the time, the association between Broad Street water and death is much less clear in the case of a hypothetical typhoid fever epidemic (Figure 1.4B).

Since Whitehead's attempt, both practical and psycho-social factors have also contributed to unconscious adoption of bias-susceptible experimental designs. From a data efficiency perspective, case-control studies are often characterized by large class imbalances. A case-control experiment is one of the only ways to take efficient advantage of all minority class observations in a model. The analogous cohort experiment would require identifying a starting alignment date common to all study subjects. Furthermore, longitudinal observational data are often expensive or difficult to acquire, compared to the ease of one-shot, non-temporal case-control datasets. Without the use of retrospective observations, a case-control study is one of the only types that can be conducted immediately after the study is conceived, rather than waiting for observations to be generated, as in prospective studies.

The human and scientific tendency to condense high-dimensional, complex subjects into well-defined states is also a contributing factor. The need to define individuals as “case” or “control,” “diseased” or “healthy” is a necessary part of medicine and research but necessitates a loss of fidelity regarding the patient’s actual condition. This is unavoidable- even in the framework of precision medicine, the conceit is that the smallest groups for which distinct advice is relevant can be identified and that this better than “imprecise” medicine. However, not even the most precise methods have enabled n=1 diagnosis. An awareness and understanding that patients are unique in their complexities and situations can mitigate the potential impacts of temporally biased findings.

More concerningly, publication biases towards larger effect sizes and higher accuracy may have driven researchers towards methods that accentuate the differences between cases and controls. Temporal bias can be interpreted as a relatively invisible symptom of this subconscious aversion towards ambiguity in prognostic models. Strong predictive models (in terms of accuracy) are naturally easier to create when structural differences between the two groups are used to provide additional signal. The increasing popularity of large data sets and difficult-to-interpret deep learning techniques facilitates this strategy.

This is not to say that case-control studies should be abandoned wholesale. These studies for practical reasons (data efficiency, cost, ease of deployment) have contributed countless numbers of discoveries to the scientific canon across fields. However, a systematic understanding of where and why temporal bias exists is critical in the transition of research findings to applications in the clinic and beyond. There are several strategies to minimize temporal bias where it exists and evaluate the size and direction of its effects otherwise. (Figure 1.4B-E)

1) Assuming that a suitable control population can be identified, the following two conditions can enable uniform sampling of the control-to-case trajectory: i) the use of a randomized lookback time, and ii) the length of the maximum lookback time plus the length of the observation window is longer than the transition period (as with Whitehead and cholera).

2) Person-time classification or prediction tasks, where multiple windows are drawn from sufficiently extended case observations for use can also uniformly sample the trajectory in question. This approach takes the form of sampling case trajectories more than once.

3) The use of well-defined baseline dates for exposure evaluation is particularly attractive due to ease of deployment. Assessing exposure after a particular birthday, at the start of a particular month/year, or after a well-defined event makes the prospective deployment population much easier to identify. While, strictly speaking, these forward-facing windows do not uniformly sample the trajectory, they also do not invert the flow of time (Figure 1.3A) and, thus, can be used prospectively.

Finally, sensitivity analyses, such as those demonstrated in our Lp(a)/MI or delivery prediction experiments, combined with researchers' background domain knowledge regarding the state transition trajectory in question can be used to estimate effects of prospective deployment. An increasing focus on considering the deployability of a given model, the nature of the underlying trajectory, or even whether a particular feature can realistically be predicted from features at hand can also serve to prevent temporal bias from infiltrating a study.

While temporal bias is common and has far reaching implications, it is unique among experimental or epistemological flaws in that once understood, it is fairly easy to detect. As experiments grow ever broader in scope and higher in ambition, transparency regarding the

extent to which temporal bias influences findings is key to ensuring the consistency of associations or predictions, allowing for the reproducibility of results, and maintaining greater credibility of the scientific process.

Theoretical Framework for Temporal Bias

We present a theoretical framework for understanding temporal bias for an arbitrary case-control study. These types of studies can have two objectives, 1) to utilize observed features to predict the transition of controls to cases or 2) to evaluate the strength of association between observed features and case-control transition. Variables and descriptions are presented in Table 1.2, while a narrative description of the framework is presented below.

Table 1.2: Variables and Descriptions

Variable	Description
s	Given an observation in a case individual, the time of the observation relative to the future event. Always a negative value.
$f(s)$	Trajectory of observations and predictors towards the future event. General form: $f(s) = \begin{cases} \overline{F}_C \ \ if \ \ s \leq s_c \\ g(s) \ \ if \ \ s > s_c \ \ where \ \lim_{s \to 0} g(s) = \overline{F}_{s=0} \end{cases}$
Case population	Defined by the presence of the event, aligned in time such that events among all individuals happen simultaneously (s=0).
Control population	Defined by absence of event.
$F_{s=0}$	Distribution of observed $f(s)$ at s=0.
F_C	Distribution of observed $f(s)$ of control population.
w_1,w_2	Relative weight of control and case populations respectively.
s_C	Minimum magnitude timegap s when F_C and $w_1 C + w_2 F_{s=0}$ are identically distributed.
F_{s_C}	Distribution of observed $f(s)$ at s=s_C.
i	Lookback time, defined by the experimenter during a study. A positive value.
j	Observation window, defined by the experimenter during a study. A positive value.
k	Defined as $i + j + s_C$

The case population is defined by the first occurrence of a characteristic event that serves to align population members in time. This event is the subject of the study: individuals who experience the event are cases while those who do not are potential controls. We next define a timegap s that represents the amount of time between an observation and the characteristic event. For every member of this population, this event is fixed to occur at timegap $s = 0$. We next define a control (non-case) population whose members are i) representative of the population that produced the cases and ii) are at risk of experiencing the event in question. Let $f(s)$ represent the predictor trajectory (Condition 2, Background). Let $F_{s=0}$ and F_C represent the distribution of observed $f(s)$ for the case population at $s = 0$ and for the control population, respectively. Note that individual observations from F_C do not correspond to uniform values s because controls do not have an event against which to index, by definition. We define s_C as the minimum timegap when F_C and $w_1 F_C + w_2 F_{s_C}$ are identically distributed (Condition 3, Background), where w_1 and w_2 are the relative populations of the control and case population, respectively, and F_{s_C} is the distribution of observed $f(s)$ among cases at s_C. That is, s_C is defined as the earliest point where the cases begin to diverge from the controls. Because controls are selected due to not experiencing the event, their observations correspond to otherwise unknown values of $s \leq s_C$ (Figure 1.3B). By definition, controls transition to cases and not vice versa, so $max(f(s)) = \overline{F_{s=0}}$ and $min(f(s)) = \overline{F_C}$ (Condition 1, Chapter 1, Background). Finally, we note that F_C does not necessarily represent "zero progression"; it is, rather, an abstraction of the baseline progression of the control population toward the event in question. This is a representation of the fact that there is a non-zero chance that controls transition to cases, assuming controls were properly selected with exchangeability in mind. We can now express the general form of $f(s)$: for $s \leq s_C$, $f(s) = F_C$; for $s > s_C$, $f(s)$ increases to $F_{s=0}$ at $s = 0$. This

process appears to introduce a discontinuity in $f(s)$ when $s = s_C$. This is an artifact of the binary nature of the case-control experiment. In essence, although the ground truth trajectory may be continuous, the experiment naturally uses the observed control distribution to define a cutoff beneath which controls are considered exchangeable. The discontinuity occurs when the true trajectory intersects with this cutoff.

During the experiment itself, a look-back time i and observation window j are defined where $i, j \geq 0$. For static time point predictions, $j = 0$. These values are not necessarily constant for all features or across all patients. Observations of the case population correspond to those made at values of s between $[-i - j,\ -i]$, while observations of the control population still correspond to $s \leq s_C$(where all observations are members of the F_C distribution). When $i = j = 0$, the case observations correspond directly to $s = 0$. We assume that i is chosen by the experimenter based on the period of time when a prediction or evaluation would be useful. Therefore, in prospective deployment, individuals with observations made at $s > -i$ will not need to be considered for the purpose of model evaluation. Comparisons between the case observations $[F_{-i-j}, F_{-i}]$ and control populations $F_C$can then be made, leading to odds ratios for observations and the event, or "predictive models" for the event.

Temporal bias occurs when $i + j < |s_C|$ and $\frac{1}{j}\int_{-i-j}^{-i} f(s)\ ds > \frac{1}{-i-j}\int_{s_C}^{-i-j} f(s)\ ds$, while the magnitude of the bias can be estimated by $\frac{\int_{-i-j}^{-i} f(s)\, ds}{f(-i)*(k+j+s_C)}$, where $k = i + j + s_C$. Prospective deployment of an odds ratio or model can be thought of as randomly sampling $f(s)$ over s between $[s_C,\ 0]$. Note that this task is impossible: all values of s are negative and defined relative to a case event occuring in the future. The precise value of s for a given individual cannot be known until this event happens, while if an individual is censored, this value will never

be known. The tension of using s, rather than conventional time, is an artifact of the fact that case-control studies intrinsically leak information regarding case status from the future. The case-control window structure assumes an instantaneous transition between $s = 0$ and $s = s_c + i + j$. As we show, this results in an exaggeration of odds ratios or predictive accuracies due to sampling cases artificially at timepoints close to the event.

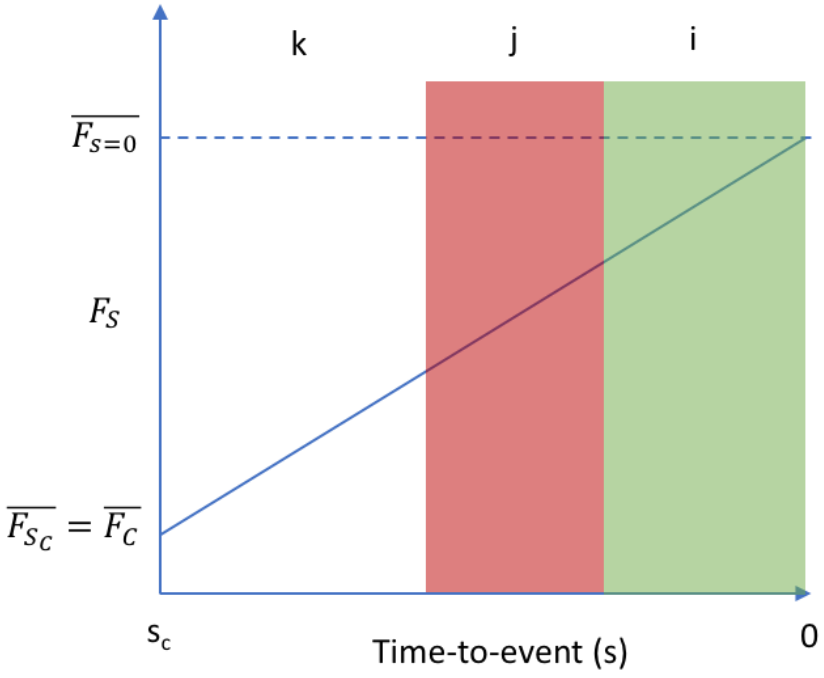


Figure 1.5: Schematic of Generic Temporal Bias Scheme. The investigator attempts to use features in the observation window (j) with a prediction window offset (i) to predict an individual's position across the entire trajectory: an impossible task.

Materials and Methods

Lipoprotein(a) Trajectory Imputation

Centiles of lipoprotein A values [Lp(a)] for myocardial infarction (MI) of 4441 Chinese patients (cases) and healthy matched controls (controls) were utilized as described by Paré, et al.

(Paré *et al.*, 2019) to construct log-normal distributions of Lp(a) values for each cohort. One hundred fifty thousand case and control measurements were drawn and a linear model was fit to establish the baseline coefficient of association between Lp(a) and MI in the presence of temporal bias. For trajectory imputation, for each case patient, a starting Lp(a) value was generated using one of three methods: i) random sampling from the control distribution such that the drawn value is smaller than the case value, ii) percentile matching (if the case value fell in the Nth percentile of the case distribution, the Nth percentile value from the control was drawn), and iii) a uniform shift of 15% (representing the observation that the median control value was 15% lower than the median case). This starting value is understood to represent the Lp(a) measurement of the case patient in the distant past at the point when they were cardiovascularly healthy. For each pair of case-starting and case-ending values, a linear/logarithmic/logistic/step function was fit using the two values as starting and ending points. New case observations were generated by randomly selecting a point along the generated trajectory allowing for the computation of a "prospective" effect size. All individual experiments were repeated 100 times with newly drawn sample cohorts.

To examine the potential impact of inadvertent selection bias on the observed association between Lp(a) and MI, the Lp(a) values and MI for all patients with more than two Lp(a) observations prior to the first recorded MI event were extracted from the Partners Research Patient Data Registry database. This work was approved by the Partners Institutional Review Board (Protocol #2018P000016). Case and control patients were defined based on MI status, and for each patient in each cohort, the i) largest available, ii) smallest available, and iii) mean Lp(a) values were computed and used to identify the observed effect size under each selection scheme. All calculations were conducted in R using the glmnet package.

Delivery Prediction from Sequential Claims Data

Records of health insurance claims in 2015 from a deidentified national database were utilized for this study. Delivery events were identified based on International Classification of Diseases (ICD9/10) diagnostic code, Current Procedural Terminology (CPT) code, or the birth year of newly born members linked by subscriber-parent annotations. Cases were defined as individuals who experienced a delivery between February and December, 2015, while controls were defined as individuals who did not experience a delivery during any of 2015. Thirty thousand cases were randomly selected and matched to 30,000 controls based on age and ZIP code. For each individual, case-control and cohort feature windows were defined. Case-control windows were set as the month of records that was three months prior to the delivery/matched baseline date for cases and controls respectively. Cohort windows were set as the month of records from January, 2015. Three studies were simulated: 1) The CC-CC study consisted of model training using case-control windows and model evaluation using case-control windows. 2) The CC-Cohort study consisted of model training using case-control windows and model evaluation using cohort windows. 3) The Cohort-Cohort study consisted of model training using cohort windows and model evaluation using cohort windows. For each study, deep recurrent neural networks and logistic regression models were trained over the features present in each window. For deep recurrent neural network-based models, the linear sequence of features inside the window was provided in the form of International Classification of Diseases (ICD9) codes for diagnoses, Current Procedural Terminology (CPT) codes for procedures, and National Drug Codes (NDC) for prescriptions. The sequence length was set to 20 events, individual sequences were either padded or clipped to meet this requirement. Logistic regression models utilized binary occurrence matrices for all events as features. Both models contained demographic

information in the form of age. Sex was excluded as a feature because all cohort members were female. All calculations were conducted in Python using the Keras and scikit-learn packages.

Simulation of Olive Oil/Myocardial Infarction Case-Control Study

Data from the Nurses' Health Study (NHS) was used for this analysis. All nutrition and disease incidence surveys between 1994 and 2010 were considered. Internal NHS definitions of first MI were utilized to define the case population. Case individuals were only considered if they had at least two consecutive nutritional surveys with answers to all olive oil related questions prior to the first MI event. Individuals with any history of cardiovascular disease including MI and angina were excluded from the control population. Control individuals were only considered if they had at least two consecutive nutritional surveys with answers to all olive oil related questions. In total, 3,188 total qualifying MI individuals were identified, and 94,893 controls. A baseline date for each control individual was defined based on the availability of consecutive nutrition surveys. For each case, a matched control was identified based using age at baseline and sex. For all individuals, total cumulative yearly olive oil consumption was computed by summing olive oil added to food and olive oil salad dressing consumption, as validated by Guasch-Ferré, et al. (Guasch-Ferré *et al.*, 2015). For each experiment, a lookback time between 1-4 years was selected, and the cumulative total olive oil consumed during the lookback time relative to the MI date/baseline was calculated. For each lookback time, the effect size between the top quintile (based on total consumption) and the remaining population and statistical significance were calculated to match the protocols of Fernández-Jarne et al. and Bertuzzi, et al. (Bertuzzi *et al.*, 2002; Fernández-Jarne *et al.*, 2002). Each experiment, including case-control matching, was repeated 200 times. All calculations were conducted in R using the glmnet package.

Chapter 2: Characterizing prediagnostic Parkinson's disease and predicting onset in a clinically useful manner

Chapter Introduction

This chapter applies ideas about temporal bias to construct a clinically useful classifier for Parkinson's disease diagnosis. Previous work in risk stratification was impacted by temporal bias, making it impossible for clinicians to determine which patients were within the scope of the model in real time. Because the manuscript that this chapter was adapted from was aimed primarily at a neurology audience, temporal bias was only referenced conceptually in the text, rather than by name. A technical addendum is also present at the end of the chapter, describing the development of modeling methods used.

Main Text

Parkinson's Disease (PD) is the second most common neurodegenerative disorder worldwide and is increasing in incidence as the general population ages (Pringsheim *et al.*, 2014). Current treatment strategies focus on alleviating clinical symptoms, which have dramatic effects on quality of life but have little ability to slow disease progression (Lang and Espay, 2018). However, a new class of drugs is being developed that targets recently discovered genomic loci (Hardy *et al.*, 2009; Nalls *et al.*, 2014; Chang *et al.*, 2017; Deng, Wang and Jankovic, 2018; Iwaki *et al.*, 2019) associated with PD to modify disease progression. These include inhibitors of leucine-rich repeat kinase activity, antibodies against α-synuclein, and compounds that modulate glucocerebrosidase activity (Sardi and Simuni, 2019). The targets of these therapeutics are common pathways that may either cause PD or enhance the risk of

developing the disease. The success of the new class of therapeutics will rely in part on the ability to identify PD cases earlier, when an intervention has greater potential to have an impact. The primary goal of this study is to facilitate the early identification of PD onset through prediction and characterization of prediagnostic PD.

Perhaps the most challenging aspect for early intervention using this new therapeutic pipeline is that approximately 30-60% of substantia nigra pars compacta dopaminergic neurons have already died by the time PD is diagnosed (Gaig and Tolosa, 2009; Tabbal *et al.*, 2012) There is evidence that clinical symptoms, including non-motor features, begin to occur several years before a PD diagnosis (Berg *et al.*, 2015; Mahlknecht, Seppi and Poewe, 2015). Early detection of these symptoms may enable earlier identification of people at high risk ultimately leading to faster diagnoses. Certain prediagnostic features, based on clinical observations, have been widely studied and include impaired olfaction, constipation, urinary disorders, disturbed sleep patterns, anxiety and depression, autonomic dysfunction, and many others (Gonera *et al.*, 1997; Abbott *et al.*, 2001, 2005; Ross *et al.*, 2008; Postuma *et al.*, 2012; Lerche *et al.*, 2014; Schrag *et al.*, 2015; Darweesh *et al.*, 2017). Further insight into the first clinical presentations of these prediagnostic features, as well as others not traditionally thought to be components of prediagnostic PD, and their temporal relationships would help to delineate the pathophysiology of early Parkinson's disease progression. This would enable the identification of people at increased risk of developing overt Parkinson's disease, who could be eligible for inclusion in clinical trials of early neuroprotective strategies and ultimately preventative interventions.

Using a curated set of these variables from the period just prior to a PD diagnosis, Schrag et al. (Schrag *et al.*, 2019) demonstrated that it is possible to develop an algorithm to effectively predict whether a person will develop PD within five years. However, one limitation of this

study was their focus on the entire time period up to an individual's PD diagnosis. We hypothesized that this algorithm was primarily influenced by patients where clinicians already suspected PD. In this case, because the data captured included the complete medical history that was used for PD diagnosis at their subsequent visit with a neurologist, we suspected that the signal driving selectivity of this algorithm, and others built on similar methods, derived primarily from features close in time to the diagnosis itself. Given that the delay to diagnosis is well-established in PD and has been shown to take a median of around one year (Breen *et al.*, 2013), this would limit the impact of this diagnostic algorithm.

Furthermore, the technical development of the algorithm makes prospective real-world usage in the clinic difficult. Because the algorithm specifically distinguishes patient cohorts selected based on their future PD status (first presentations within 5 years from a PD diagnosis), a clinician would not know which individuals to apply the algorithm to until after a patient's PD diagnosis. It is not possible to evaluate if someone is within 5 years of a PD diagnosis in real time. Consequently, we sought to develop a predictor with well-defined entry criteria to enable clinical utility based on specific clinical events.

In this study, we utilized two large health record databases to develop a predictor of which individuals progress into PD with a focus on actionability that takes into account the unique features surrounding the trajectory of PD. Ultimately, accurate, prospective identification of high-risk individuals would allow for both earlier diagnosis, intervention, and more effective large-scale evaluation of potential therapeutics.

Methods:

Data

The main components of this study were performed utilizing two data sources:

1) The Partners Healthcare Research Patient Data Registry, composed of electronic medical records (EMR) approximately 6 million individuals in Massachusetts. Data used in this study covers patients records from the early 1990s through the end of 2018.

2) A de-identified administrative claims database from a large private insurance company representing more than 75 million unique members during a period extending from January 1, 2008 through December 31, 2018. Members with zip codes in Massachusetts were excluded from our analyses so as not to overlap with the first dataset.

In both datasets we extracted gender, year of birth, coverage or enrollment duration, zip code, ethnicity, diagnoses (in the form of International Classification of Diseases, 9th and 10th Revision codes (ICD9/10)) and procedures (in the form of Current Procedural Terminology codes (CPT)). Medication prescriptions were not evaluated due to incomplete coverage for this population in the Claims data.

Subject/Control Selection Criteria

We utilized a similar set of case criteria to other studies identifying PD cohorts in large medical records databases (Alonso *et al.*, 2007; Schrag *et al.*, 2015)- these criteria are specifically modeled after inclusion criteria for PD clinical trials. Individuals were first required to have at least two ICD diagnosis codes for PD. The first of these codes was set as their baseline point and a second code was required within the 2 years following their baseline point and at least 90 days after their first. A minimum age of 50 at baseline was set to remove cases not likely

to be sporadic idiopathic PD. Subjects were required to have at least two years of claims data prior to their baseline diagnosis and 2 years following in order to capture the prodromal period of the disease and to track progression. The two years of data prior to their baseline limits the possibility of inclusion of patients with PD that were diagnosed previously (Lewis *et al.*, 2005).

Subjects who presented with diagnoses for encephalitis, Alzheimer's disease or similar cognitive disorders that could phenocopy a true idiopathic PD diagnosis during the window prior to baseline were removed. Subjects with any presentation prior to baseline date of schizophrenia, other Parkinson's like disorders including metabolic neurogenic disorders (e.g. Wilson's Disease), or other degenerative diseases that produce a clinical syndrome of parkinsonism (Multiple system atrophy, Progressive supranuclear palsy) were removed. All codes utilized are listed in Table S2.1.

A control cohort was structured in a similar manner. First, an artificial baseline point in time was established such that the distribution of available records following their baseline point matched the distribution of the same time window for the PD cohort. This was done to ensure a comparable follow-up window in which all controls must have representative data. We finally required that all subjects have at least 2 years prior to and post their baseline date, the latter being a criteria already established from the baseline point matching, and be at least 50 years of age at their baseline point. Finally, we selected a matched subset of controls to PD cases using age (within 5 years) and gender.

We later resampled the databases for anyone having either a gait and/or tremor disorder diagnosis based on ICD codes (Table S2.1). Cases were defined as those patients eventually diagnosed with PD and controls set to those who did not. For both cohorts, we utilized first diagnosis of gait and/or tremor disorder prior to PD diagnosis as their baseline points. All other

inclusion/exclusion criteria were repurposed using this new baseline point. No matching was conducted for these tasks, as the entry criteria were well defined. Subjects with a PD diagnosis at baseline were excluded as the prediction task was already accomplished by the clinician.

Prediagnostic PD Trajectory Modeling/PD Progression Prediction

We conducted two parallel forms of modeling to examine the trajectory of prediagnostic PD: 1) a logistic regression model using an occurrence matrix of individual features; 2) deep learning over a patient's observed temporal sequence of claims. The sequence was sampled at different time points corresponding to different prediction windows sizes: 0, 15, 30, 45, 75, 90, 180, 270, 360, 450, 640, and 720 days prior to PD diagnosis/baseline. For each time point, a two-year long observation from the patient's sequence was used. As an example, for the 75 day time-point, records between 75 and 805 days prior to the baseline were utilized in the model, while records within 75 days of baseline were excluded. The features included were patient demographic data, diagnoses (ICD codes mapped to PheWAS (Denny *et al.*, 2010) codes to reduce dimensionality), procedures (both CPT & ICD), and time between data points. We later repeated these tasks to model progression into PD using the two year window prior to first gait and/or tremor disorder diagnosis.

Static Regression Model

We fit a penalized regression model to predict the diagnosis of PD using a static prediction vector constructed of the values of demographic data and counts of diagnoses, procedures and time between data points. We measured predictive accuracy via area under the receiver operator characteristics using 5-fold cross validation. We then calculated odds ratios and

95% confidence intervals on the entire dataset. We performed univariate association testing using age and gender-controlled logistic regression to identify features that demonstrated an association with PD onset. We first performed this association testing with all features and then again with the features present in at least 0.5% of the population. This process was repeated independently in both the insurance claims dataset as well as the Partners research database. We filtered to identify features with significant p-values after multiple testing using the Bonferroni adjustment (Haynes, 2013) in both datasets.

Deep Learning Temporal Sequence Model

We trained a deep recurrent neural network (RNN) using gated recurrent units (GRU) to predict the onset of PD using each patient's sequence of interactions with the healthcare system (claim or entry into medical record). Sequences for the RNN were constructed using temporal embeddings trained from a separate cohort of one million individuals older than 50.(Beam *et al.*, 2020) Temporal sequences were constructed by interleaving tokens signifying the time between events with tokens representing the events themselves. A co-occurrence matrix was created over all tokens, where events that happened within 7 days of each other were said to co-occur. This window was chosen because events that are temporally close are likely to reflect simultaneous aspects of patient physiology. This matrix was then factored to produce a unique embedding vector for each token. Given an observation window, temporal sequences of events with the window were created for case and control individuals, using the previously created embeddings. Sequences were clipped or padded to a length of 1200 tokens to ensure equal lengths between individuals, with clipping occurring on the earliest events in a window when necessary. Sequences were classified by a deep GRU recurrent neural network. Models were retrained from

scratch using randomly selected train, test, and validation splits to produce confidence intervals. Neural network models were trained only in claims data due to the large amount of data required to construct embedding vectors. A technical addendum on this process is presented at the end of the chapter.

Comparison between Predictive Models trained using different Data Modalities

We compared the predictive model trained using EMR data to the model trained using administrative claims data in two ways: 1) comparing the performance of the model outputs and 2) comparing the features driving the model performance between the two different models. This comparison of relative feature importance was performed by first calculating the Pearson correlation of each data modality separately. This was compared to the correlation of feature importance between the two different data modalities.

Comparative Diagnosis Prevalence

Trends in comparative diagnosis prevalence were identified by first identifying a set population of PD cases and age/gender matched control individuals with coverage prior to and after each particular window. For every given time point, defined as the 365 days relative to the point itself, and a given diagnosis, the prevalence of that diagnosis within that window was computed. Prevalence was computed for PD case and control populations separately. For example, a tremor frequency of 0.08 among cases at day 730 implies that 8.0% of PD cases had a tremor diagnosis between 730 and 365 days prior to their PD diagnosis.

Results

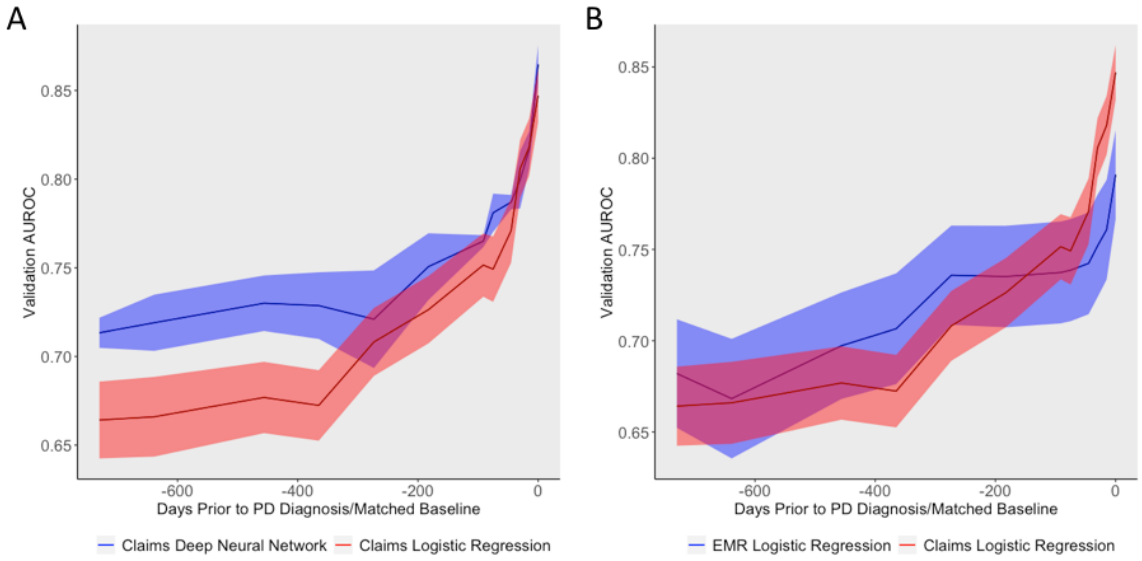
Cohort Demographics

Table 2.1 describes the demographics of the EMR and Claims based cohorts, stratified by the PD case status. Age of first diagnosis was slightly higher in the claims cohort, but was over 70 in both datasets. Our cohorts align with accepted estimates of PD incidence in the population(Poewe *et al.*, 2017) (Table 2.1). Population statistics between cases and matched controls largely align between the EMR and Claims data though the latter population is slightly younger (owing to the transfer of individuals above 65 to Medicare) and has more extended terms of coverage due to the nature of the data sources. EMR records only capture an individual's interactions with that particular hospital system, while claims records capture all of an individual's paid interactions while they were insured.

Table 2.1. Population statistics between cases and matched controls in EMR and Claims Data. Ethnicity data was only available for a subset of patients.

	EMR		Claims	
	PD Cases	Matched PD Controls	PD Cases	Matched PD Controls
Total	3251	18851	5131	23085
Male (%)	1903 (58.5)	11131 (59.0)	3151 (61.4)	14177 (61.4)
Age at Enrollment (STD)	63.15 (10.78)	63.94 (10.909)	69.01 (10.33)	68.71 (10.34)
Age at Baseline (STD)	72.48 (9.33)	72.64 (10.39)	73.70 (10.23)	73.69 (10.23)
Fraction White (among available)	87.6	87.8	81.8	80.6
Percentage African American (among available)	2.4	3.44	1.92	4.25
Percentage Hispanic (among available)	2.4	1.47	3.42	2.85
Percentage Asian (among available)	1.66	1.01	2.84	2.85
Percentage Other Race (among available)	5.93	6.21	1	9.41
Enrollment Months (STD)	209.08 (78.66)	186.77 (75.62)	106.51 (17.84)	104.86 (18.62)
Enrollment Months Prior to Baseline (STD)	111.88 (72.57)	104.29 (68.90)	61.49 (17.59)	64.95 (18.47)
Enrollment Months After Baseline (STD)	97.68 (58.94)	82.99 (51.43)	45.14 (16.06)	40.04 (13.44)

Parkinson's Disease Trajectory is Characterized by a Prodromal Period



C

Description	OR	FDR Adjusted P Value
Abnormality of gait	8.313329	2.80E-188
Abnormal involuntary movements	55.96452	1.49E-168
Malaise and fatigue	3.532172	2.33E-112
Essential tremor	38.04427	1.96E-79
Depression	4.071419	2.96E-76
Back pain	2.533556	2.98E-62
Essential hypertension	1.573989	3.41E-57
Dizziness and giddiness (Light-headedness and vertigo)	3.079242	2.08E-54
Anxiety disorder	3.50637	1.72E-51
Major depressive disorder	4.161642	1.73E-47
Pain in joint	1.983804	6.84E-46
Constipation	3.518131	5.03E-42
Spondylosis without myelopathy	2.683271	5.19E-40
Lack of coordination	10.97439	1.63E-39
Pain in limb	2.258863	6.99E-35
Extrapyramidal disease and abnormal movement disorders	51.00728	8.70E-35
Abdominal pain	2.186139	1.43E-33
Type 2 diabetes	1.776482	4.41E-32
Urinary incontinence	3.458665	9.38E-32

Figure 2.1. Area under the ROC Curve predicting PD onset at various points prior to PD diagnosis. **A.)** Logistic Regression vs. Neural Network in Claims **B.)** EMR vs. Claims Logistic Regression **C.)** Top 20 diagnoses for predicting PD immediately prior to PD diagnosis.

We began by using a deep neural network to construct an unbiased prediction algorithm for future PD diagnosis utilizing two years of observations prior to the PD diagnosis in cases and matched controls. In contrast to prior models, we sequentially compared different time periods before the PD diagnosis date. We found a significant spike in prediction accuracy as the size of this window was reduced, which reached a maximum immediately prior to the PD diagnosis (Figure 2.1A,B). We found that the accuracy of the deep neural network and a logistic regression model trained on identical claims data converged as the diagnosis date approached, implying that the most relevant signal for that time period was additive, with linear relationships between clinical events (diagnoses, prescriptions and procedures), whereas earlier time points appeared to be driven by non-linear, complex relationships between factors that only neural networks could resolve. The increase in performance closer to PD diagnosis date by both prediction models indicated the existence of a pre-diagnostic window during which motor symptoms were present but the diagnosis had not yet been made. Clinicians have described a time period immediately ranging between 3 months to one year (Breen *et al.*, 2013) where PD is suspected and the patient is referred to neurologists or subjected to more rigorous clinical evaluation before a formal PD diagnosis is rendered. Consequently, the strong performance of classifiers that include this period may be illusory: the models draw signal from the actions of clinicians who already suspect PD. We find that the dominant features of this window include diagnoses of abnormality of gait, as well as diagnoses corresponding to tremor disorders (abnormal involuntary movements, essential tremor) (Figure 2.1C), which likely represent proxy diagnoses for PD prior to a neurologist confirming the diagnosis.

Gait and Tremor Disorders highlight PD Differential Diagnostic Window

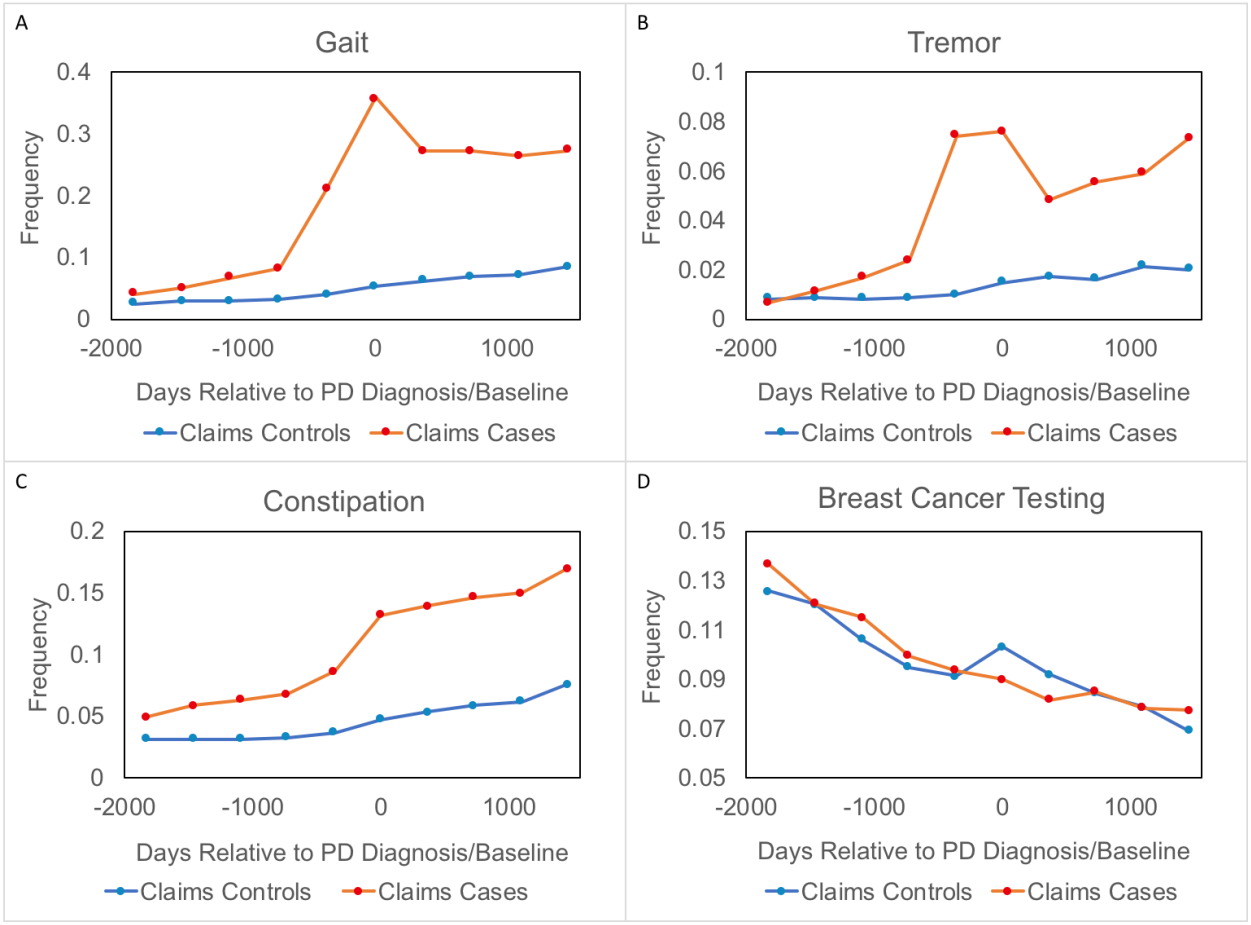


Figure 2.2. Frequency of phenotypes relative to PD diagnosis date (cases)/matched baseline date (controls). Each point represents the frequency of the phenotype among the population in the year defined at the point: a tremor frequency of 0.08 at day 730 implies that 8.0% of PD cases had a tremor diagnosis between 730 and 365 days prior to their PD diagnosis.

In order to better characterize the predictive implications and utility of this pre-diagnostic window, we examined the rates of different diagnoses relative to the PD diagnosis date corresponding to select phenotypes: gait disorders, tremor disorders, constipation (a known prodromal symptom of PD), as well as a phenotype with little if any known physiological connection to PD: breast cancer screening (Table S2.1B). Gait and tremor diagnoses were chosen based on their strength of association and the presence of sufficient patients to create PD

classifiers indexed to their first diagnosis point. In the case of constipation, we found elevated rates of diagnosis prior to the PD diagnosis date, that elevate prior to and post PD diagnosis. A small spike at PD diagnosis is likely due to increased documentation at this critical inflection point in care. In contrast, constipation among PD controls increases more gradually over the whole window, but is agnostic to the baseline date itself. This behavior is consistent with constipation's role as a symptom of PD. Breast cancer testing, a test performed as a part of the standard of care, showed little variance between PD cases and controls throughout the entire window, consistent with the lack of evidence for a physiological association to PD. We find that gait and tremor disorders among PD cases slowly diverge from controls until a large spike approximately one year prior to the PD diagnosis, and fall off in the years post diagnosis likely due to their replacement with a PD code. This suggests that gait and tremor diagnoses are being used as proxy diagnoses in the runup to the PD diagnosis, consistent with the presence of a pre-diagnostic window.

Predicting Parkinson's Disease Progression from First Gait/Tremor diagnosis

Based on the importance of gait and tremor diagnoses in the prediagnostic models and the above finding that they are widely used as proxies for a PD diagnosis, we constructed three new cohorts where baseline classification dates were defined as i) the diagnosis of first gait or tremor disorder, ii) the first diagnosis of gait disorder only, and iii) the first diagnosis of tremor disorder only. In all three cases, all patients were gait/tremor naive prior to their baseline. Two years of features for each patient prior to the baseline were collected. The shift from a predictor based on a case-control study to a cohort study is useful in several ways. Not only are cohort studies considered a higher level of evidence (Burns, Rohrich and Chung, 2011), but the presence of a

well-defined entry date allows for deployment of a predictor in clinical workflow. We used identical model architectures/parameters (both neural network and penalized logistic regression) for gait and tremor indexed models as for prediagnostic models (Figure 2.1). The primary difference was the selection of the baseline point: a point in the future for the prediagnostic models, compared to a point at present for the gait/tremor models. We find that as the models are directed to focus on more specific cohorts, accuracy declines, in both claims and EMR, as well as between both logistic regression and deep neural network based models (Table 2.2). The strongest predictor for future PD diagnosis for all three (gait or tremor, gait only, tremor only) cohorts was bipolar disorder (Table 2.3A, Table S2.2-3), an association that has been highlighted by other epidemiologic studies (Faustino *et al.*, 2019). Progression into PD from gait disorders only was uniquely defined by a histories of features such as urinary tract infection and chronic laryngitis, while progression from tremor disorders only was uniquely defined by parasomnia. While both of these symptoms are known to be early symptoms of PD, the distinction in their contribution towards risk in these gait and tremor defined cohorts may indicate differences between two subsets of disease. The models trained on both data sources showed strong correlation (Pearson correlation of 0.71) between individual feature odds ratios.

We examined the strongest performing model (Table 2.2A), the neural network predicting PD progression from either first gait or tremor in more depth (Table 2.2B, Table 2.3). For this model, we examined the average days-in-advance that the model predicted PD for individuals who truly went on to experience a PD diagnosis on record at various false positive rate (FPR) thresholds. While the mean days saved declined slightly as the FPR threshold is increased, the average was still in excess of 300 days with an FPR rate of 0.01. This indicates

that model performance is not dominated by individuals who immediately go on to develop PD after a gait or tremor diagnosis, and that among this selective cohort, early diagnosis is feasible.

Table 2.2. (Table continues on next page) **A)** Claims, EMR Prediction accuracy at first gait or tremor, first tremor, and first gait. **B)** Analysis of advance prediction time at various FPR thresholds for first gait or tremor Deep Neural Network Model.
A)

	Claims				
	Demographics			Validation AUROC (95% Confidence)	
	PD Cases	Percent Progressing to PD	Average days to PD (SD)	Deep Neural Network	Logistic Regression
First Gait or Tremor	8475	2.43	469 (493)	0.874 (0.869-0.879)	0.803 (0.791-0.816)
First Gait Only	3925	1.37	575 (521)	0.769 (0.759-0.780)	0.791 (0.772-0.809)
First Tremor Only	4550	6.69	377 (447)	0.698 (0.679-0.718)	0.697 (0.674-0.719)

	EMR			
	Demographics			Validation AUROC (95% Confidence)
	PD Cases	Percent Progressing to PD	Average days to PD (SD)	Logistic Regression
First Gait or Tremor	1349	3.08	548 (517)	0.804 (0.792-0.816)
First Gait Only	694	2.23	606 (530)	0.714 (0.679-0.750)
First Tremor Only	681	5.24	479 (490)	0.757 (0.730-0.784)

B)

FPR	FNR	Mean Days Saved (SD)
0.90	0.00	377 (399)
0.80	0.01	375 (397)
0.70	0.03	369 (395)
0.60	0.04	368 (396)
0.50	0.07	360 (390)
0.40	0.12	348 (384)
0.30	0.18	339 (376)
0.20	0.26	334 (372)
0.10	0.33	322 (371)
0.01	0.44	303 (369)

Table 2.3. (Table continues on next page) **A)** Strongest positive features in first gait/tremor cohort **B)** Difference in PD progression odds ratio between deep learning prodromal cohorts (Figure 2.1) and gait/tremor cohorts (Table 2.2).

A)

Description	Gait/Tremor OR	Gait/Tremor Adjusted P Value
Bipolar disorder	3.392	1.04E-88
Major depressive disorder	1.628	1.44E-27
Voice disturbance	2.04	8.74E-21
Memory loss	1.725	5.39E-18
Other non-epithelial cancer of skin	1.334	4.14E-17
Senile cataract	1.233	1.75E-16
Other persistent mental disorders due to conditions classified elsewhere	1.944	1.33E-14
Actinic keratosis	1.219	1.06E-12
Urinary incontinence	1.414	1.16E-12
Depression	1.293	6.30E-12
Symptoms concerning nutrition, metabolism, and development	1.434	2.70E-10
Frequency of urination and polyuria	1.269	1.47E-09
Malaise and fatigue	1.133	5.64E-07
Seborrheic dermatitis	1.475	8.92E-07
Inflammation of eyelids	1.325	1.06E-06

B)

Description	Prediagnostic OR	Prediagnostic Adjusted P value	Gait/Tremor OR	Gait/Tremor P Value
Mental Health Diagnoses				
Major depressive disorder	3.10	2.28E-103	1.63	1.43E-27
Mood disorders	3.62	4.54E-18	1.77	1.50E-05
Bipolar	5.68	1.82E-73	3.39	1.03E-88
Depression	2.57	3.73E-109	1.29	6.29E-12
Anemia-related Diagnoses				
Other anemias	1.13	0.412	0.74	2.67E-18
Iron deficiency anemia secondary to blood loss (chronic)	1.61	3.4812E-05	0.77	0.028
Other Diagnoses				
Constipation	2.24	4.42E-72	1.17	0.0006
Frequency of urination and polyuria	1.66	4.96E-39	1.26	1.47E-09
Urinary incontinence	2.11	2.56E-46	1.41	1.16E-12
Hypersomnia	2.61	3.65E-12	1.43	0.010
Hypotension NOS	1.97	3.53E-19	0.8	0.054
Dizziness and giddiness	2.29	2.35E-133	1.08	0.025

Upon review of the results, we highlighted sets of diagnoses that were significantly different between the first prediagnostic model and the gait and tremor cohort model (Table 2.3B). In particular, the odds ratio directionality of anemia and hypotension reversed when evaluated in the presence of first gait/tremor, meaning that these diagnoses were no longer predictive of future PD. Similarly, while constipation is a known symptom of prediagnostic PD (Poewe *et al.*, 2017), it is less useful at predicting who will progress to PD from gait/tremor than in the original cohorts. These results suggest that distinct trajectories into PD may be present, including trajectories characterized by gait or tremor disorders. These findings also suggest that the controls defined in gait/tremor indexed cohorts represent a distinct population from traditionally defined PD controls, and that the true real-world PD progression prediction task is sensitive to the particular comparisons that a clinician is making.

Discussion:

Disease modification remains a major unmet need for PD. Despite many attempts, not a single study has been successful. One of the potential reasons for these failed studies is the advanced disease state of the studied populations. The present study bridges this gap by providing a novel approach to identify the population at risk of "converting" to PD, before marked symptoms. This main original contributions of this work are: (1) the unbiased characterization of prediagnostic PD, by utilizing a well curated cohort of PD patients and matched controls; (2) the mapping of the temporal relationships of prediagnostic features to evaluate which diagnoses define the later stage of this period (i.e. the "pre-conversion" or suspected PD that we defined as a pre-diagnostic window); and (3) the deployment of novel

machine learning approaches to develop a clinically deployable model for predicting which patients will progress into PD. Implementation of this strategy would facilitate earlier diagnosis and, ultimately, preventative interventions.

The presence of a pre-diagnostic period has complicated and obfuscated attempts to develop predictive models for PD using standard machine learning approaches. Our models, along with those proposed by Schrag, et al (Schrag *et al.*, 2019), that included the pre- diagnostic window all had AUROC values between 0.8 and 0.85, despite the large differences in input data and imputation methods. Hand curated factors and simple linear models performed roughly as well as highly complex neural networks with access to a comprehensive record of interactions with the health care system. This observation implies that within this period, signal is overwhelmingly dominated by proxy diagnoses for PD, and that the signal here is illusory: physicians likely already suspect PD in most of the true positive cases. In order to establish clinical utility for decision support surrounding PD, it is critical for predictive models to identify novel signals at critical times in care, rather than report what a doctor already suspects about a diagnosis.

One way to ensure this bias is through restricting the scope of a predictive model to a more homogenous cohort defined by a specific inflection point in their health. By identifying that the prediagnostic period is, for many, characterized by an initial gait and tremor disorder, we avoid the biases that stem from attempting to determine if a particular model is appropriate for a particular patient. It is feasible for a physician to determine if a patient has their first gait or tremor disorder whereas it is unlikely if a physician can predict if a patient is 1/2/3 years from a diagnosis of PD. This ‘specificity-first’ approach can also yield insights into the heterogeneity of the disease state: as mentioned before, PD can be thought of as a syndrome with numerous

subtypes. An example of a subtype can be seen by the way gait/tremor defined PD trajectory behaves in a different manner than PD as a whole. The algorithm proposed by Schrag et al. (Schrag *et al.*, 2019) nominates an additive relationship between various factors, among them dizziness, hypotension, gait, and tremor. In contrast, at the time of a first gait or first tremor diagnosis, we found hypotension was no longer predictive of PD onset and dizziness had only a very weak effect. This suggests that among gait/tremor defined PD, an algorithm agnostic to latent PD subtypes may overestimate risk of progression among some patients.

Our study has several limitations driven primarily by the use of real world data collected primarily from billing and patient care. First, we used a data-driven approach to define a sufficient quiescence period prior to de novo PD diagnoses. Despite this, there is no guarantee that an individual may not have received a PD diagnosis either prior to appearing in the data (first visit or enrollment into insurance coverage) or outside of the data (in another insurance plan or health system, or through prescriptions, which were not included in this study). Unfortunately, our study was unable to include prescription data to exclude drug-induced Parkinson's cases. The association between Bipolar disorder and PD diagnosis has been previously described(Faustino *et al.*, 2019), although many Bipolar treatments (anti-psychotic medications, valproic acid) are known to cause secondary Parkinsonism. We were not able to clarify this association. Finally, the pre and post-baseline record restrictions that we implemented to ensure the integrity of our cohorts would serve to bias our analysis towards populations with extended lengths of coverage.

Overall, this approach is well suited for not only guiding clinical decision-making regarding referrals and accelerated diagnoses, but also allows for more closely aligning machine learning predictors with the infrastructure around clinical trials. Reliable risk stratification could

identify eligible patients earlier while also providing a proxy endpoint that can be tracked in a continuous manner. By providing the basis for identifying distinct subpopulations and disease progression trajectories, physiological hypotheses regarding the nature of the disease can be elucidated and more precise recommendations made to clinicians.

Technical Addendum

This subsection describes the technical development of the model architecture used earlier in this chapter. Alzheimer's disease was used as the focus due to increased availability of patient data.

Introduction

An important first step in using deep learning for patient risk stratification is the manner of representation of patient sequences (Ching, Himmelstein, Beaulieu-Jones, Kalinin, Do, Way, Ferrero, Agapow, Zietz, Hoffman, Xie, *et al.*, 2018). Clinical data is frequently high dimensional: for example, there are 68,000 ICD 10 diagnosis codes. Traditional approaches of representation (e.g. one-hot encoding) lead to challenges with dimensionality. Consequently, concept embeddings have become a common way to map concepts into a meaningful vector space with a fixed dimension (Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff, 2013). An embedding for a concept can be learned by predicting the concepts likely to occur prior to or after a specific event. These methods can encode longitudinal patient histories into sequences that can be passed directly into machine learning models (Choi, Chiu and Sontag, 2016; Brett K. Beaulieu-Jones, Kohane and Beam, 2019; Beam *et al.*, 2020). Series of embeddings can then be easily passed into deep learning methods developed to predict

using sequences (e.g. Recurrent Neural Networks and Temporal Convolutional Networks) in domains such as natural language processing (Mikolov, Tomas and Karafiat, Martin and Burget, Lukas and Cernocky, Jan and Khudanpur, Sanjeev, 2010; Sundermeyer, Martin and Schluter, Ralf and Ney, Hermann, 2012), signal processing (Giles, Lawrence and Tsoi, 2001), and speech recognition (Graves, Mohamed and Hinton, 2013).

In terms of predictors themselves, Temporal Convolutional Neural Networks (TCN) have been used to predict disease diagnoses from laboratory tests (Razavian, Narges and Sontag, David, 2015). (Choi *et al.*, 2017) employed Recurrent neural networks (RNNs) have been used to predict heart failure (Choi *et al.*, 2017) and mortality from ICU data (Beaulieu-Jones, Orzechowski and Moore, 2018). More recently, RNNs have been utilized to predict inpatient outcomes using raw standardized representations (FHIR) (Bender and Sartipi, 2013) from two different institutions without needing to explicitly map features (Rajkomar *et al.*, 2018). A key challenge to each of these methods is the ability to map gaps between patient events or interactions with the health system in the encoding representation. Many previous studies have focused on stratifying patient risk by predicting outcomes over a relatively short period of time (e.g. 24 hours after hospital admission). Some previous works have imputed measurements that occur in the gaps between interactions, but do not impute discrete events (Futoma, Joseph and Hariharan, Sanjay and Heller, Katherine, 2017). In longer time windows, the only data available occurs when a subject has an interaction with the healthcare system. Over a longer time window, it is not possible to impute events that may occur in between interactions because these events would be missing not at random (Graham, 2009). The data reflects the way the healthcare system operates. Interactions with the healthcare system represent snapshots of deviations in patient physiology, as opposed to continuous monitoring of quantitative biomarkers. Furthermore, the

idea of imputation begins to break down when the features imputed move from continuous lab values to sequence of observed events. While imputation strategies can estimate unmeasured values, this approach does not easily extend to discrete events. Applying machine learning on data created by the healthcare system is an exercise in modeling the dynamics of the healthcare system and not necessarily the health of a patient or progression of disease.

Previous attempts to address this limitation have taken several approaches, including appending the time to the next event at each observation (Graham, 2009; Choi *et al.*, 2017) or constructing a custom RNN structure utilizing masking and time interval vectors as input (Che *et al.*, 2018). These methods still fundamentally provide sequential input to the model in uneven steps. As an analogy, CNNs have demonstrated exceptional performance at image classification by taking into account the spatial organization between adjacent pixels. Their performance would likely suffer if they were presented with a random subset of discontinuous pixels along with annotations of their relative distance to neighbors.

Overview

In this study we used a large insurance administrative claims dataset, containing records of diagnoses and procedures for nearly 70 million individuals between 2010 to 2018, to predict the onset of Alzheimer's disease (AD) using 7 clinically deployable cohorts. By choosing a relevant diagnosis as the index date, models are evaluated in a manner that is relevant to the clinic without the biases that accompany arbitrary index date selection.

To build these predictive models, we introduced a novel encoding scheme for time that can be input to standard sequence-based models. This novel encoding scheme embeds irregularly spaced events in association.

Methods

Cohort Definition

We first identified classes of diagnoses from the literature that were found to have a temporal association with AD or suggested to be potential risk factors of AD. For each of these seven clusters, i) Memory Impairment (Amaducci *et al.*, 1987; Wolk, David A and Dickerson, Bradford C, 2016) , ii) Executive Function Disorders (Wolk, David A and Dickerson, Bradford C, 2016), iii) Depression and Mood Disorders (Speck *et al.*, 1995; Larson, 2016), iv) Motor Function Disorders (Shadlen, Marie-Florence and Larson, Eric B, 2010), v) Seizures, vi) Sleep Disorders, and vii) Cardiology and Vascular Disorders (Luchsinger *et al.*, 2005; Whitmer *et al.*, 2005; Keene, C Dirk and Montine, Thomas J and Kuller, Lewis H, 2016), we utilized Phenome-Wide Association Study (PheWAS) codes (Denny *et al.*, 2010) to identify sets of corresponding ICD9 (Slee, 1978) codes that formed the basis for each examined cohort.

Data from this study was collected using deidentified claims data from a nationwide US health insurance plan that contains data from more than 75,000,000 individuals over a 10 year period. For each cohort, we identified the subset of individuals who had their first annotation of the cohort diagnosis after the age of 60 (this would become their baseline date), had no annotation or Alzheimer's prior to the baseline, and also had at least 24 months of records prior to and 24 months of records after the baseline. The records prior to the baseline served as both a quiescence period in order to select the population who had not been previously diagnosed with Alzheimer's and the observation window leading up to the index event. The records in the 24 months after the index event were used to check for Alzheimer's diagnoses.

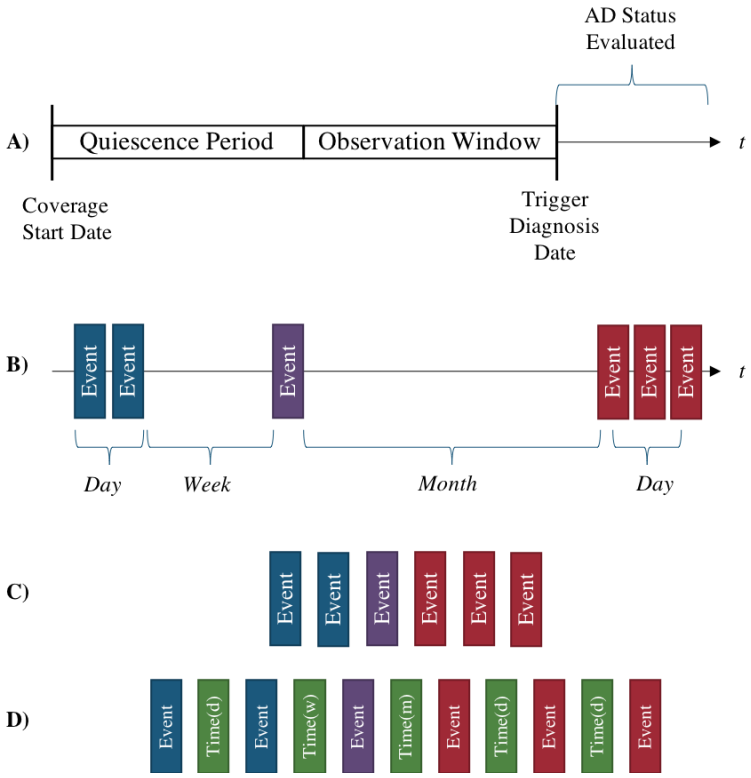


Figure 2.3: Cohort Data Selection/Processing A) Timeline segregation scheme. Only event sequences within the observation window were utilized in the model. **B)** Observed patient events arranged on a timeline. **C)** Traditional event-sequence encoding. **D)** Time-gapped encoding of patient event.

Data Preparation and Event Embedding

Our dataset consists of events characterized by the patient experiencing the event, the time the event occurred, the class of event (diagnosis, procedure, prescription), and the identity of the event (hospital diagnosis of myocardial infarction, inpatient prescription of donepezil, etc.). Furthermore, demographic information for each patient, including age at enrollment, sex, ZIP code, and period of enrollment is also available. External databases for classifying different classes of event were utilized to collapse similar concepts to single classes. Diagnoses expressed as ICD9 codes were collapsed to the phenotype level using PheWAS codes for the purposes of cohort identification and selection. For example, ICD9 code 001, corresponding to Cholera, and

ICD9 code 002, corresponding to Typhoid Fever, would be collapsed into PheWAS code 008: Intestinal infection. Prescriptions were collapsed to the pharmaceutical level using Generic Product Identifier (GPI) codes, resulting in different dosages, name brands/generics being treated according to their main active ingredient. For modelling, all ICD10 codes were cross-walked back to ICD9 for compatibility prior to October, 2015. Embeddings were trained on ICD9 codes. Procedures were expressed using Current Procedural Terminology (CPT) codes. The provider-submitted date the service was started was used for all events.

We first selected a sample of 10 million enrollees who had coverage over the age of 60 and identified their associated events and composed timestamped sequences of events corresponding to an individual's enrollment period. We next composed seven logarithmically grouped bins from the entire distribution of inter-event time gaps (1 day, 3 days, 6 days, 15 days, 36 days, 86 days, 208 days or greater). In order to compensate for the unevenly spaced event sequences, we modeled the time between each event as a discrete event. Between each pair of medical events in an individual's sequence, we inserted a dummy event corresponding to the logarithmic bin that the inter-event time gap was assigned to. As an example, a sequence of three events that occurred on the same day would be transformed into a sequence of five: the original three events separated by "Time gap" dummy events (Figure 2.3B-D). For each pair of events, we computed co-occurence statistics based on 30-day temporal windows. For embedding schemes that utilized time gaps, events were always defined as co-occuring with their adjacent time gaps, even if they were separated from their adjacent events by more than 30 days. We required that events co-occur within 30 days at least 10 times across all data. This included 23,193,710 co-occurring pairs and 9,228,186,699 total co-occurences. These co-occurence matrices were transformed into Euclidean embedding sets as described in Beam et al. (Beam *et*

al., 2020). Unlike other concept embedding schemes, we added additional events for time-spacings and trained additional embedding models for them as well as all observed events (Figure 2.3B, Table 2.4). In order to assess the impact of explicitly modelling the inter-event gaps of time, embeddings were created using the event-only sequences to serve a comparison group (Figure 2.3C, Table 2.5).

Table 2.4: Cohort Demographic Statistics

Cohort	Total Enrollees with Index Event (IE)	AD Onset after IE (count)	AD Incidence after IE	AD Onset Age: Years (STD)	Days from IE to AD (STD)
Any Unspecified Diagnosis	3,598,330	43,791	1.22%	82.54	NA
Memory Impairment	212,994	47,234	22.18%	83.38	363 (211)
Executive Function Disorders	80,200	14,213	17.72%	82.36	369 (213)
Depression and Mood Disorders	661,242	42,260	5.80%	84.19	369 (216)
Seizures	580,788	33,966	5.81%	84.53	392 (219)
Sleep Disorders	361,702	12,389	3.43%	82.97	411 (216)
Cardiology and Vascular Disorders	582,881	29,893	5.13%	83.89	390 (220)

Model Training and Evaluation

For each cohort, individuals were labeled based on if they were diagnosed with Alzheimer's after their baseline. Events from the observation window were replaced with corresponding event Euclidean embedding vectors. Binary prediction models were created using the vector sequences and patient demographics vectors using a stacked gated recurrent unit (GRU) architecture. A branching model with separate processing of vector sequences and patient demographics was utilized. For events with no corresponding embedding, a generic embedding of the corresponding type (diagnosis, procedure, etc) was substituted. These were created by calculating the average embedding vector of all events of that type. For these models, patient sequences were padded/clipped to 600 events (1200 with gaps). For sequences that were clipped, the most temporally distant events were preferentially removed. Event sequences were fed into the model in reverse order, with events closest to the baseline input first. Sequences of event vectors were padded with zero vectors. Individual vectors had the day of the window (0-730) and time-to-next-event (in days) appended. To evaluate the utility of the temporal information in the feature set, models were also trained without gap information, and with their sequences shuffled, rather than arranged by date. Models were initialized and trained using TensorFlow and Keras on four NVIDIA Titan X GPUs.

Results

Study Population

Summary statistics calculated for each of the cohorts, along with a reference cohort of "Any Unspecified Diagnosis" are presented in Table 2.4. All of the phenotypes described by our literature informed cohorts had higher subsequent incidences of Alzheimer's relative to the

reference cohort. All cohorts had a similar age at AD onset and similar time periods between index events and the initial diagnosis. On average, all cohorts represented predictions around one year in advance of the Alzheimer's diagnosis date.

Model Performance

Table 2.5: Model Performance: Influence of temporal information

	Logistic Regression	**Shuffled Euclidean**	**Euclidean**	**Gapped Euclidean**
Memory Impairment	0.51 (0.50-0.52)	0.51 (0.50-0.54)	0.56 (0.54-0.59)	0.65 (0.63-0.67)
Executive Function Disorders	0.65 (0.63-0.67)	0.64 (0.58-0.66)	0.66 (0.62-0.73)	0.71 (0.64-0.79)
Depression and Mood Disorders	0.68 (0.67-0.69)	0.70 (0.61-0.77)	0.72 (0.65-0.77)	0.79 (0.74-0.82)
Motor Function Disorders	0.67 (0.66-0.68)	0.68 (0.57-0.73)	0.70 (0.62-0.74)	0.72 (0.64-79)
Seizures	0.58 (0.57-0.59)	0.62 (0.56-0.72)	0.71 (0.65-0.74)	0.77 (0.72-0.82)
Sleep Disorders	0.66 (0.63-0.69)	0.72 (0.76-0.78)	0.78 (0.72-0.83)	0.82 (0.76-0.85)
Cardiology and Vascular Disorders	0.52 (0.50-0.54)	0.53 (0.51-0.56)	0.58 (0.55-0.64)	0.64 (0.57-0.68)

For all cohorts, increasing the fidelity of the temporal information contained in the embedding improved the validation AUC (Table 2.5). Furthermore, for all cohorts besides Motor Function, the difference in performance between gapped and un-gapped embeddings was larger

than the difference between un-gapped sequences (in the proper order) and shuffled sequences (in random order). We did not observe significant trends in performance between cohorts and factors such as cohort size, Alzheimer's incidence rate, or time to Alzheimer's. We hypothesize that the specific biological relationship between Alzheimer's and the defining phenotype of each cohort is the dominant factor in general performance of a given model, as well as the relative influence that timing and ordering of features has on performance within a cohort. Cohorts defined by Depression & Mood, Seizures, and Sleep disorders had particularly strong predictive performance (AUC > 0.75) when time-gapped embedding schemes were used.

Discussion

The expansion of previous event-sequence paradigms to include inter-event timings effectively doubles the feature-space available to models. It has been shown that the time in which a medical test is conducted can be just as informative as the test result itself (Agniel, Kohane and Weber, 2018)). Similarly, we show here that the knowledge of whether a series of events occurs on the same day, or over a year long period can significantly improve the predictive accuracy irrespective of cohort. In addition, the modeling of event time gaps serves as a way to bridge the gap between health care data and more traditional modalities of continuous monitoring. Unlike biomonitoring data, modalities such as electronic medical records or claims data do not represent patient state. Instead, they represent a decision, on the part of the patient, to engage (or not engage, in the case of time gaps), with the health care system. The knowledge that a patient had not been seen by a primary care physician since their last yearly checkup is intuited by the clinician during a visit, but is often not made available to machine learning models, despite its significant implications regarding patient physiology.

Chapter 3: Machine Learners as Knowledge Parasites

Chapter Introduction

The previous chapter introduced one metric of model utility as "the ability to produce predictions beyond what a clinician would suspect on their own." This chapter, adapted from a perspective aimed at a clinical audience, introduces the idea of a "knowledge parasite:" a model that learns by "looking over a clinician's shoulder", rather than truly assisting with the diagnostic process. This phenomenon is directly caused by the inability for models to justify the decisions or predictions that they make.

Main Text

A patient presents to the emergency department complaining of chest pain. An examination and history prompt you, the physician, to suspect a pulmonary embolism (PE), and you order a d-dimer test. The results of the d-dimer, as well as the subsequent chest x-ray, raise your confidence in a PE, so you order a chest CT scan. Without examining the patient themselves, a rotating medical student arrives at a similar conclusion of PE based on observing which tests have been ordered. While the medical student's belief in a PE diagnosis may be both justified and correct, the medical student has not proven they would arrive at the same conclusion without your presence, knowledge, and prior actions.

Machine learning models ostensibly designed to predict patient outcomes from electronic medical records (EMRs) and administrative data have become a popular focus for research and investment. Physician-initiated data comprises large portions of EMRs and administrative data sources and leads to the generation of models that typically function similarly to the medical

student in the PE example. Instead of basing decisions from a patient's current state, they use the prior decisions and actions taken by a physician. Upon interpretation of the chest CT results, you conclude that the patient is in no danger and order them released. However, a predictive algorithm could classify your patient as "high-risk" based on a CT scan that you ordered and interpreted, even though you determined the patient not to be at risk (Figure 3.1). Rather than surpassing or even capturing your knowledge, the model instead acts as a "parasite," looking over your shoulder and presenting your own expertise without clean attribution to what (or whom) is powering the model.

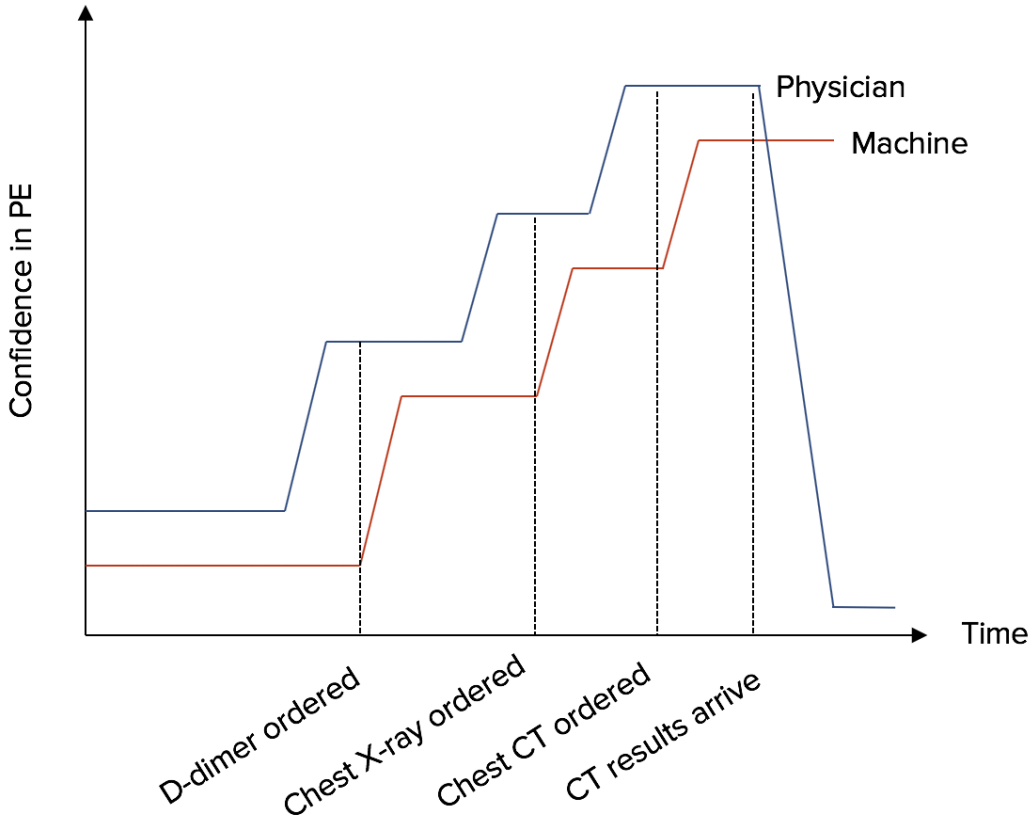


Figure 3.1. Confidence of Physician and Machine in pulmonary embolism (PE) vs. time. Events undertaken by the physician are annotated.

The confluence of machine learning progress in a variety of fields and the accessibility of large medical datasets prompted an explosion of interest in individualized decision support. Given the complexity of patient care and the ability to monitor patients in more dimensions than ever before (e.g. EMR or real time vital sign monitoring), there is significant interest in predicting outcomes of health, for example, mortality or disease onset. Despite strong statistical performance, few, if any, models for predicting risk in conditions as diverse as heart failure to hospital readmissions to Parkinson's disease, have made inroads into clinical workflows.

The majority of healthcare data, including electronic medical records and administrative data, while often thought of as a measure of patient state, are actually a representation of deliberate actions undertaken by the physician. Consider the act of ordering a chest CT. In this case, the physician acts as a filter: only patients with appropriate presentations will prompt a CT. Knowing this, a machine learning algorithm (or enterprising trainee) can borrow the judgement and expertise of experienced physicians by reporting that patients who receive CTs should be treated as higher-risk. This concept introduces an important distinction between physician initiated and non-physician-initiated data elements. Conflating these data elements deemphasizes the role that physicians have in deciding under what circumstances to act, and what course of action to take.

To highlight the difference between these data, contrast the comparative utility of the model that evaluates patient risk based on the presence or absence of a CT scan (physician-initiated data) with a second model that reports risk based on an image-based examination of the CT itself (non-physician initiated data). In the first case, all of the information that the model has regarding the patient's physiology is filtered through your eyes as the attending physician. You decide if and when to order a scan based on your own evaluation of the patient. In contrast, the

second model receives the same scan that the radiologist receives, and could be argued to have access to pixel-level features that human clinicians cannot reliably comprehend. Conclusions that this image-based model reaches are significantly more likely to represent novel information because they are based off of physiological measurements of the patient.

When predicting individual patient outcomes, the points at which physician-initiated data based model guidance would be most useful (the patients with the most unique or ambiguous presentations) are often the points where these models are least equipped to provide it. This is commonly because the signal that they capture is derived from the aggregate behavior and expertise of a large number of physicians: inferred over thousands of observations. This means the resulting predictions are the exact opposite of individualized: they represent the average decision making across many doctors for many patients who happen to present similarly.

This is not to say that these models do not possess value, but that their value is not necessarily to help a clinician make an individual treatment decision. The task of prediction is inextricably linked with the task of inferring causal relationships from real world observations, and consequently, requires very high levels of evidence. Instead, these models have the potential to construct faithful representations of the decisions made by physicians. Doing this could make a significant impact for tasks including real-time monitoring, telemetry, and surveillance. Additionally, studying these models provides an understanding of the standard of care and provides the ability to examine deviation from standards in an attempt to quantify process quality.

For example, while the predictions of a model may arrive too late to influence the decisions of a responsible physician, they collectively can provide a real-time, updating picture of the overall health status of a hospital's current patient load, enabling administrators to better

allocate resources and dictate staffing levels. Alternatively, these models and datasets are perfectly positioned to detect which clinical guidelines might be applicable to a particular patient, or if ordered treatments align with standard practice. High level understanding of where and why deviations exist could enable guideline creation that is simultaneously data driven and more responsive to physician's needs.

There is an important contrast to note. Machine learning algorithms have been shown to provide novel insights by looking for features beyond what could be seen by trained physicians. Imaging analysis of retinal scans for diabetic retinopathy patients or of malignant pathology represent genuine advances in patient-level risk assessment specifically because they capture data that is not physician initiated. Consequently, machine analysis of data modalities such as vital signs, telemetry, imaging or genotype data is far more likely to produce genuine insights that expand upon and complement a physician's existing expertise. Promoting the routine collection of these non-physician-initiated data elements can provide the basis for the development of tools that are better able to offer clinically meaningful insights when deployed alongside, and under the supervision of, physicians.

Capturing the initial promise of machine learning requires a proper understanding of where their inductions are sourced from, and the limits that these inferences are subjected to. Properly navigating the gulf between data that is and is not physician initiated is critical in determining the proper use cases for models. The characterization of machine learning as recapturing rather than surpassing the expertise of physicians will enable both researchers and physicians to determine how and where models can make the greatest impact.

Chapter 4: Machine Learning for Patient Risk Stratification: Standing on, or looking over, the shoulders of clinicians?

Chapter Introduction

This chapter examines the "knowledge parasite" phenomenon from a quantitative point of view, utilizing metrics of healthcare dynamics as a direct simulation of a model that "looks over a clinician's shoulder." I describe the concepts of "clinician-initiated" data elements and the prognosis-diagnosis dichotomy. For these models, I show that the source of a prediction is important in identifying optimal use cases.

Main Text

Machine learning for healthcare promises to have a major impact on the delivery of data-driven personalized medicine (Beam and Kohane, 2016; Topol, 2019). One of the applications with the widest potential is patient risk stratification (i.e. diagnosis, prognosis) (Ching, Himmelstein, Beaulieu-Jones, Kalinin, Do, Way, Ferrero, Agapow, Zietz, Hoffman and Others, 2018). Individualized patient risk stratification requires machine learning models to predict the future disease state of a patient based on his or her current clinical state and available history (Weiss *et al.*, 2012). One major obstacle to this vision is that the true physiological state of a patient is often incompletely characterized and obfuscated through various sources of bias in the electronic medical record (EMR) (Botsis *et al.*, 2010; Weiskopf and Weng, 2013; Crown, 2015; van der Bij *et al.*, 2017; Beaulieu-Jones *et al.*, 2018). Despite this, most current machine learning investigations utilizing these data rely on a major simplifying assumption: that the state of a patient can be inferred through the use of routinely collected data in the EMR (Shickel *et al.*, 2018). However, these data encode information about how clinicians and the healthcare system

as a whole react to the patient, potentially confounding prediction models built to use it. Machine learning models trained using EMR-derived features are consequently linked to the individual decisions and assessments made by clinicians.

When a patient's physiology reaches a state that necessitates examination, the clinician's beliefs regarding potential patient outcomes are updated, which then inform which actions the clinician chooses to make (or not make). These actions, in turn, influence the patient's resulting physiology, and the cycle repeats (Figure 4.1A). Consequently, we define "clinician-initiated data" as data elements that represent the specific insight or expertise of the clinician, rather than direct or routine physiological measurements of the patient. For models that learn from clinician initiated data and are expected to change clinical behavior, there should be an onus to demonstrate that the model is not merely looking over a clinician's shoulder and quantifying a risk the clinician may already suspect. An example of this distinction between clinician and non-clinician initiated data can be seen in the differences between white blood cell counts taken as part of routine testing given to all patients in a ward versus white blood cell counts ordered out of patient concern. It has been observed that patients with abnormal white blood cell counts on average have better 3-year survival than patients who have normal white blood cell counts taken at abnormal times (Agniel, Kohane and Weber, 2018). Clinicians order specific panels of tests based on their clinical suspicion, expectations, or concerns about changes in clinical state. The timing of that order will also often represent a concern on the part of the provider, and many tests require manual orders and are not automatic. The primary difference between a blood test manually ordered in the middle of the night (clinician-initiated) and a routine one (non-clinician-initiated) is the decision-making agency of the physician. In the first case, a clinician chooses to order the test deliberately, based on concern prompted by examination of patient physiology. In

contrast, when a test is part of a routine process, there is no selection on patient physiology or clinician expertise.

Finally, this feedback cycle between patient physiology and physician belief/action highlights the distinction between diagnostic and prognostic tasks, and the differing burdens of evidence and assumptions required for each. The act of diagnosing a patient involves making a direct assessment of patient physiology, and for most patients, there exists a well defined answer. In contrast, prognosis involves making a prediction regarding the outcomes of a patient, and crucially assumes that the patient receives a particular standard of care in the future. Thus, like a physician's beliefs about a patient, the prognosis is dynamic and reacts to specific actions undertaken by the physician.

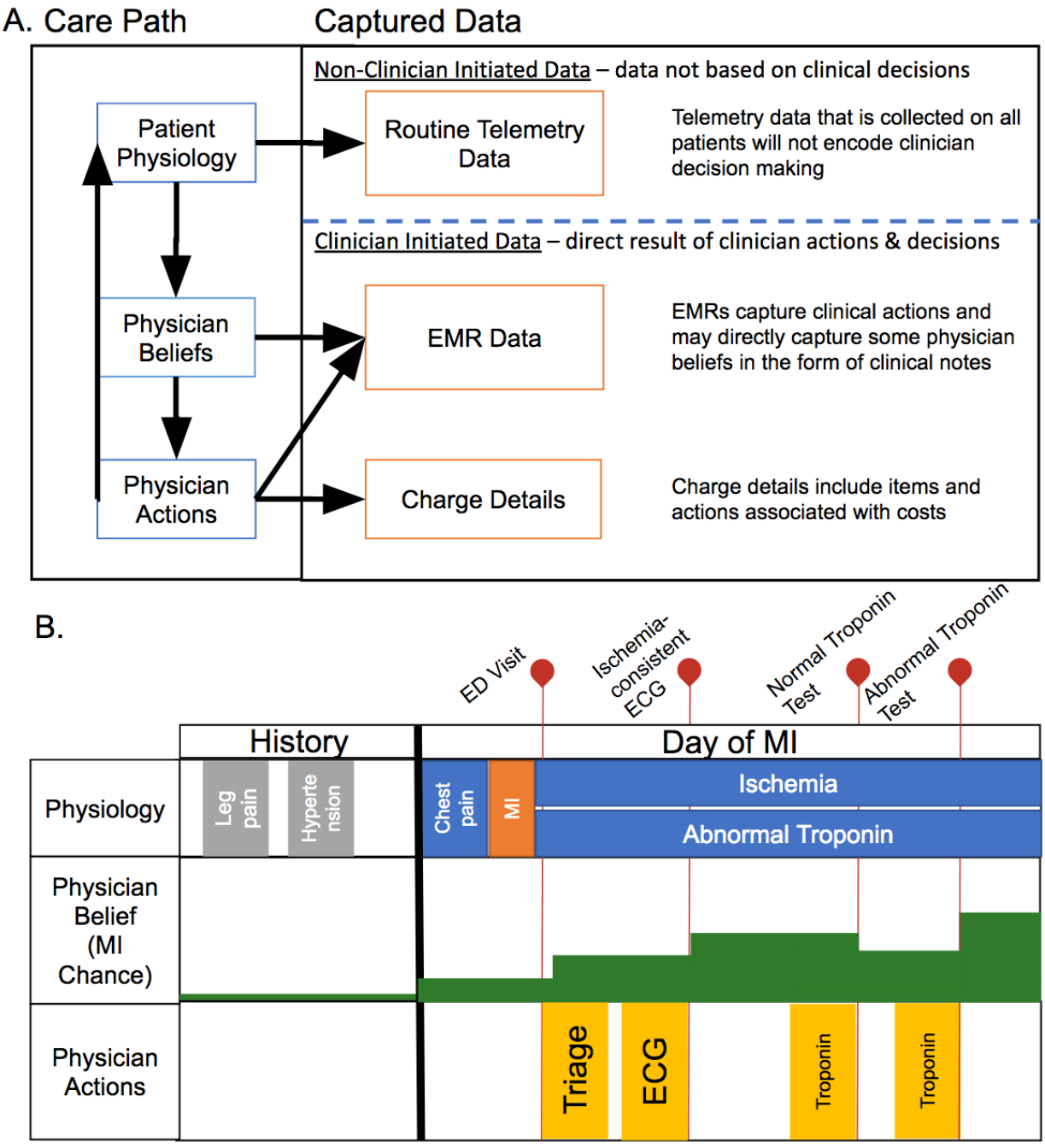


Figure 4.1. Clinician-initiated data alone is a filtered representation of patient physiology **A.)** Clinician-initiated and non-clinician initiated data are distinguished by their proximity as readouts of patient physiology, as well as the presence of the expertise of the clinician. **B.)** Physician actions are a reflection of their beliefs regarding a patient, which are formed through examination of patient physiology.

Diagnostic predictions made from clinician-initiated actions may not accurately predict beyond what the average clinician would decide for the average similar patient. As an illustrative example, we can deconstruct the timing and frequency of actions and orders by a clinician for patient presenting to the emergency department with chest pain (Figure 4.1B) (O'Gara, Kushner and Ascheim, 2013). In this example, a model utilizing this data may learn a test order for troponin means a patient is more likely to have from Myocardial Infarction (MI). Knowing that a patient has MI in combination with demographic risk factors and comorbidities may lead to impressive predictive performance for in-hospital mortality but is unlikely to aid clinical decision-making. It is identifying a behavior or concern by the clinician and not predicting a state that would enable clinical intervention. A trained machine learning model could effectively learn the correlation between abnormal clinical behavior and patient risk, but is not predicting a state that would enable clinical intervention. The model would only label a patient as 'increased risk' after the test had already been ordered, and the window for altering decision-making has passed. This idea, that models are merely interpreting the existing thoughts of clinicians based on their actions rather than identifying true signal, may help explain why increased model performance has not translated to significant clinical impact in most applications of risk stratification (Rajkomar, Dean and Kohane, 2019; Topol, 2019).

To evaluate the hypothesis that machine learning models may be modelling the existing thoughts of clinicians we quantified the ability of a deep neural network to predict patient outcomes using different subsets of data. We trained three models using: 1) patient demographic data only, 2) patient demographic data and data available at the time of presentation to the hospital, and 3) patient demographic data, data available at the time of presentation, and actions

taken during the first day of admission. The performance of these models was compared to published state-of-the-art methods using complete EMR details.

Results

We compared prediction results using charge details to state of the art benchmarks that utilize EMR-based clinical data, including notes, diagnoses, vital signs, histories, and laboratory orders/results. By evaluating the information content of a data source that contained exclusively clinician-initiated data elements, we could evaluate whether it was sufficient to achieve strong predictive performance on its own.

To do this, we utilized chargemaster details, a data modality that represents a record of the specific tasks undertaken by a hospital for a specific patient, and are used to help generate patient bills. These details represent the actions taken by clinicians (clinician-initiated data) and the resources used in order to provide care to a patient during a given encounter (Table 4.1, Tables S5.1-2). However, because they are primarily an administrative product and not used for clinical decision making, they contain only the events that occurred and resources used. Additionally, due to the de-identified nature of the data, timing and order of events within a day cannot be expected to be consistent or reliable. Importantly, because the 24 hour period after admission cannot be identified, all predictions using charge data are done at the end of the first day of admissions, and may include significantly less than 24 hours of data.

Table 4.1. Example first day charge details for a patient with MI.

Description	Department	Quantity
EKG ROUTINE TRACING ONLY	EKG	1
ECHO 2D W/ OR W/O M-MODE COMPLETE W/ COLOR FLOW	CARDIOLOGY	1
ER LEVEL V	EMERGENCY ROOM	1
XR CHEST 2 VIEWS	DIAGNOSTIC IMAGING	1
CULTURE BLOOD	LABORATORY	2
PARTIAL THROMBOPLASTIN TIME (PTT)	LABORATORY	1
PROTHROMBIN TIME (PT)	LABORATORY	1
COMPLETE CBC AUTO W/O DIFF	LABORATORY	1
TROPONIN QN	LABORATORY	2
B-TYPE NATRIURETIC PEPTIDE	LABORATORY	1
LACTATE/LACTIC ACID	LABORATORY	1
CREATINE KINASE (CPK) MB ONLY	LABORATORY	1
CREATINE KINASE (CPK)	LABORATORY	2
COMPREHENSIVE METABOLIC PANEL	LABORATORY	1
THERAPEUTIC/DIAG INJ IV PUSH SINGLE INITI SUB/DRUG	IV THERAPY	1
DOCUSATE NA, COLACE CAP 100MG	PHARMACY	1
ASPIRIN TAB 325MG (EA)	PHARMACY	1
MOXIFLOXACIN, AVELOX IVPB 400MG	PHARMACY	1
MOXIFLOXACIN, AVELOX TAB 400MG	PHARMACY	1
METOPROLOL, LOPRESSOR TAB 25MG	PHARMACY	1
IPRATROPIUM, ATROVENT INH SOL 0.02% 2.5ML	PHARMACY	1
HEPARIN NA VL 5,000U/ML 1ML	PHARMACY	1
FUROSEMIDE, LASIX TAB 20MG	PHARMACY	2
ALBUTEROL, PROVENTIL INH SOL 0.083% 3ML (2.5MG)	PHARMACY	3
R&B TELEMETRY PRIVATE	ROOM AND BOARD	1

Our analysis included 42,896,026 inpatient hospitalizations between 2013 and 2018 from 973 hospitals nationwide (Table 4.2, Figure S4.1). These hospitalizations included over 4.4 billion events occurring prior to and during the first day of admission as well as 21 static features available at the time of admission (demographic and provider details). In contrast, the EHR baseline of only 216,221 patients included more than 46.8 billion data points (Rajkomar *et al.*, 2018). We constructed 3 sets of classifiers, based on 1) demographics only, 2) demographic and provider details only, and 3) demographic, provider, and chargemaster details.

Due to the lack of event timing data, models trained with chargemaster details were only given data up to the end of the first day of admission. In contrast, published benchmarks (Rajkomar *et al.*, 2018) include full clinical details (including clinical notes) for the first twenty four hours after admission. Given that patients are admitted throughout the course of the day, many of the patients used to train our models had significantly less than twenty four hours of data.

To evaluate our hypothesis that clinical machine learning models based on a record of clinician-initiated actions are sufficient to predict inpatient outcomes, we constructed classifiers for three popular endpoints: mortality, readmission within 30 days, and extended length of stay (admissions of seven days or more). We deployed these classifiers over all admissions lasting more than one day, and included only the first day of a given stay in the classifier. Individual patients with more than one stay were classified separately, and no linkage between a given patient's stays was created. Finally, the published EMR baselines performed resource intensive neural network architecture and hyperparameter searches for over 200,000 GPU hours. The models trained on chargemaster data were trained using basic architectures on two GPUs for all outcomes in less than 24 hours.

Table 4.2. Population information for data included for risk stratification using machine learning.

	2013	2014	2015	2016	2017	2018	Total
Hospitals Included	778	783	797	786	770	755	973
Total Encounters	79,209,178	82,145,811	85,037,615	85,391,057	84,448,480	84,641,611	500,873,752
Inpatient Admissions	8,556,411	8,682,382	8,812,595	8,683,133	8,288,089	8,052,278	51,074,888
Multi-day Inpatient Admissions	7,175,154	7,338,193	7,425,860	7,296,849	6,939,021	6,720,949	42,896,026
Total Population: Mortality	120,583 (1.68%)	123,764 (1.69%)	129,640 (1.75%)	126,844 (1.74%)	124,310 (1.79%)	121,549 (1.81%)	746,690 (1.74%)
Total Population: Extended Length of Stay	1,466,580 (20.44%)	1,492,958 (20.35%)	1,518,803 (20.45%)	1,506,125 (20.64%)	1,449,174 (20.88%)	1,437,552 (21.39%)	8,871,192 (20.68%)
Total Population: 30-day Readmission	941,911 (13.13%)	937,562 (12.78%)	950,561 (12.80%)	887,418 (12.16%)	925,833 (13.34%)	901,290 (13.41%)	5,544,575 (12.93%)
All MI Admissions (% of all admissions)	69,448 (0.81%)	71,609 (0.82%)	78,975 (0.90%)	82,952 (0.96%)	84,407 (1.02%)	84,551 (1.05%)	471,942 (0.92%)

	2013	2014	2015	2016	2017	2018	Total
Multi-day MI Admissions (% of total multi-day admissions)	56,594 (0.79%)	57,665 (0.79%)	63,026 (0.85%)	65,925 (0.90%)	66,859 (0.96%)	67,135 (1.00%)	319,539 (0.88%).
MI Cohort: Mortality	3,497 (6.18%)	3,393 (5.88%)	3,625 (5.75%)	3,569 (5.41%)	3,583 (5.36%)	3,337 (4.97%)	21,004 (5.57%)
MI Cohort: Extended Length of Stay	9,172 (16.21%)	8,941 (15.51%)	9,463 (15.01%)	10,024 (15.21%)	10,036 (15.01%)	10,174 (15.15%)	57,810 (15.33%)
MI Cohort: 30-day Readmission	8,764 (15.49%)	8,891 (15.42%)	9,340 (14.82%)	9,811 (14.88%)	10,098 (15.10%)	10,140 (15.10%)	57,044 (15.12%)

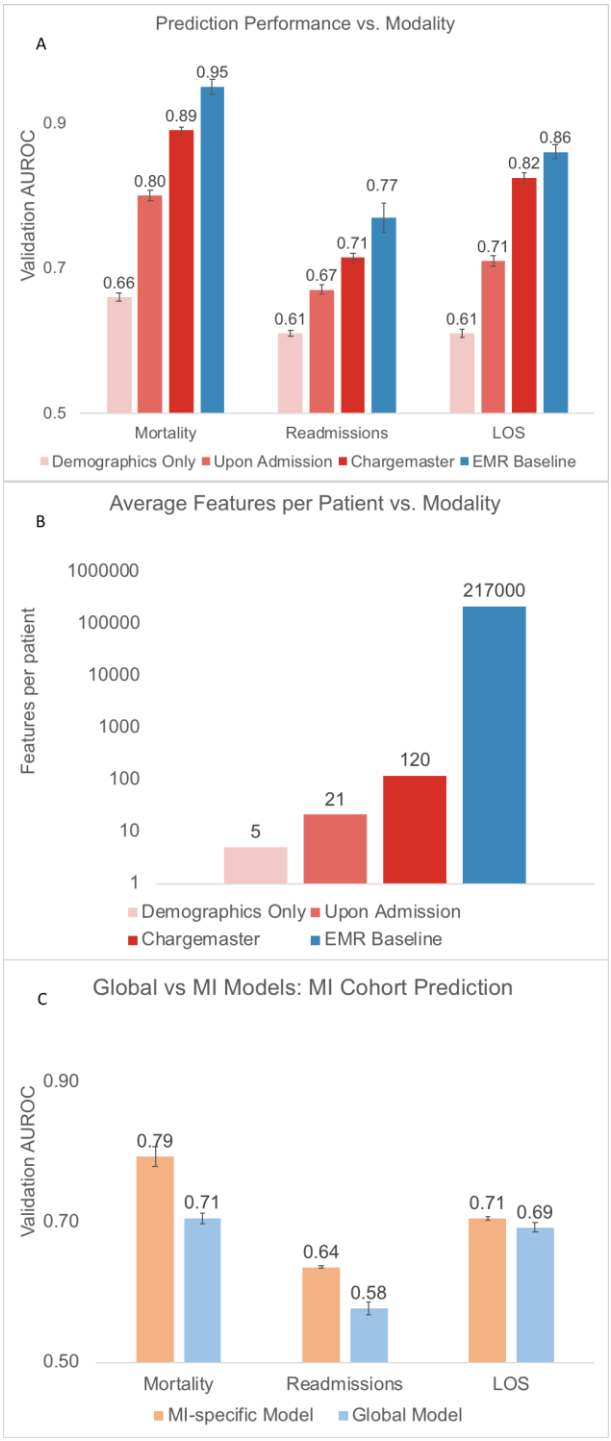


Figure 4.2. Performance comparison between Chargemaster and EMR data across cohorts and outcomes. A.) Comparison of mortality, readmission and length of stay performance (area under receiver-operating curve, AUROC) on randomly selected validation data. **B)** Average relative features per patient for each model version **C.)** Outcome comparison on a myocardial infarction (MI) patient cohort between models trained on MI patients exclusively and all available patients.

We found that, relative to the EMR baseline, abbreviated patient representations were able to capture significant amounts of signal for all three tasks (Figure 4.2A). In particular, chargemaster data only modestly underperformed the baseline (AUCs of 0.89, 0.71, and 0.82 compared to 0.95, 0.77, and 0.86 for mortality, readmissions, and LOS respectively). Performance was attenuated by the significant limitations intrinsic to chargemaster data, including missing data modalities, significantly fewer total and per-patient data elements (Figure 3.2B), lack of reliable event ordering, and the presence of less than 24 hours of data per encounter. Classifiers that utilized crude metrics of patient demographics and provider information captured the majority of signal relative to published EMR baselines over all three tasks. These results suggest that critical elements in EMR-based models are reflections and readouts of a clinician's expertise.

However, clinical practice is highly sensitive to context, and the act of prognosis frequently involves implicit diagnostic prerequisites. Consequently, we hypothesized that models trained on clinician-initiated data would be better able to predict cohort-specific outcomes when patients outside the cohort (representing irrelevant patient presentations) were excluded. We trained a model specifically on patients who arrived at the emergency department suffering from myocardial infarctions (MI). The MI cohort included MI patients hospitalized at hospitals with at least 100 such instances between 2013-2017. Models trained over this restricted subset demonstrated better performance predicting outcomes from MI hospitalizations in 2018 than the general model which was trained over all hospitalizations (Figure 4.2C). The model trained with the more expansive training set (unrefined by diagnosis) underperformed relative to one trained on a targeted subset, where physicians had diagnosed every patient with MI. This emphasizes that the prognostic performance of these models is dependent on the work of physicians to first

establish a well-defined diagnosis. Because these models derive signal primarily from patient interactions with healthcare providers, the observed effect may be caused by the potential for clinical actions to take on divergent interpretations when present in different contexts. The ability for a model trained generally to "guess" at a clinician's thinking may be less effective when required to work across contexts, as the range of mechanisms that must be inferred is much wider.

Discussion

The results of our experiments indicate machine learning models trained only on clinician-initiated administrative data can currently achieve performance close to models trained on more detailed, complete, EMR data. This is an important result because it provides insight into the current utility of machine learning models for patient risk stratification from clinical data and the primary source of signal that these models utilize. The results of our experiments also indicate the value of easier to access, lower resolution datasets (e.g. administrative vs. EMR). Finally, the results provide baseline performance levels that should be exceeded prior to claims that machine learning models can provide tangible guidance to clinicians, rather than simply looking over their shoulders.

Models trained only on clinician-initiated data currently achieve performance close to state-of-the-art models including all available data elements. This indicates that current models extrapolate from the thinking of the clinician and therefore do not demonstrate the ability to diagnose substantially better than clinicians. Operationalizing models requires identifying the specific contexts and situations where they can provide genuine guidance. Future models which aim to guide clinicians should demonstrate the ability to either suggest future actions a clinician

should take or demonstrate improved accuracy through the dominant use of non-clinician-initiated data (e.g. raw imaging results) and data that are difficult or expensive to interpret (e.g. constant real time streaming data).

The idealized use case of machine learning models for patient risk stratification is to have generalizable models that provide specific and personalized projections for individual patients. However, models that derive their predictions from clinician-initiated data may paradoxically produce predictions based on what a physician would do for an average, similarly presenting patient, rather than the individual patient in question. Acknowledging the selective role that clinicians play in terms of what decisions and actions they choose to make on what data is available for models is critical for developing models that can truly assist clinician decision making.

The implications of understanding where and how models derive their signal is important in identifying ideal use cases. While models may superficially display strong prognostic performance, if this performance is derived from the diagnostic efforts of physicians, the model cannot truly be thought of as acting independently. This observation can also explain the necessity for models to be retrained across institutions. The physiological phenomena underpinning disease are largely static, but physicians have diverse behavior profiles corresponding to different disease trajectories that might not be captured in a single training set. Instead, the use of existing clinician-generated data (e.g. EMR and chargemaster) within machine learning is currently most likely to be useful in recognizing divergence in practice in large populations as opposed to guiding prospective clinical decision making. The robustness of population level prognosis stems from the efforts of individual physicians to make diagnoses from the physiology of the patients.

Freed from the expense of collecting clinical details and the epistemological burden of predicting individual patient outcomes in an unbiased manner, machine learning models could have tremendous utility in allowing patients to view quantified prognoses, as well as guiding value-based care decisions, hospital logistics and staffing management. This is especially true using administrative byproducts such as chargemaster details. Acknowledging these models are effective at learning current clinician intuition, rather than attempting to assist individual clinicians, these models could be used to quantify current status across multiple clinicians. An example would be a tool providing administrators with a more holistic view for the current inpatient load and acuity levels of their patients. Such a tool could enable better planning, staffing and resource allocation. The ability to train cohort specific models also suggest values in lower resolution administrative datasets which may have larger patient counts that allow for the training of specialized models. A key challenge in this endeavor will be to identify cohorts prospectively in order to choose which model should be used.

The promise of machine learning in healthcare necessitates an understanding of where the dominant sources of predictive signal are located, as well as what information is truly useful in shifting marginal decisions. Through an understanding of the unique conditions in which healthcare data are created and utilized, researchers can better identify the cases where machine predictions are likely to be beneficial.

Methods

Data

The Premier Healthcare Database (PHD) (Premier Applied Sciences, 2019) is a large-scale, provider-based, all-payer database containing data on more than 215 million total patients

and 115 million inpatient admissions. It includes more than six million inpatient admissions each year between 2013 and 2018 and a total of over 35 million admissions more than one day between 2013 and 2017 (training) and 6.7 million admission lasting longer than one day in 2018 (test) (Table 3.2).

The PHD contains information on providers (hospital, organizational and clinician) and visit characteristics. It includes patient demographics, disposition and discharge information as well as diagnoses for admission and discharge, and billed services such as procedures, medications and devices, laboratory tests, diagnostic and therapeutic services.

Subsets of data:

1. Demographic Data Only -

   Age, Gender, Race, Marital status, Insurance Type (e.g. private, public, government etc.)

2. Demographic Data and information available at time of admission

   All from #1 and Admission Month, Source of Admission (e.g. another healthcare provider, home etc.), Type of Admission (e.g. Emergency, Urgent, Elective), Admitting Physician Speciality, Point of Origin (e.g. emergency department, obstetrics and gynecology etc.)

3. Demographic data, information at admission and charges during the first calendar day of admission.

   All from #1 and #2 as well as charge codes for all actions taken from presentation at the hospital until the end of the first calendar day of admission.

Cohort Selection

We include predictions of inpatient mortality, 30-day readmission, prolonged length of stay (greater than 7 days). All available hospitalizations with length of stay greater than one day

were included, and separate hospitalizations of the same patient were treated separately. Hospitalizations that ended in mortality were excluded from cohorts predicting readmission, and hospitalizations that ended in mortality after less than 7 days were excluded from cohorts predicting prolonged length of stay.

Model Architecture and Training

To make these predictions we first learn 8-dimensional clinical concept embeddings as in Beaulieu-Jones et al. (B. K. Beaulieu-Jones, Kohane and Beam, 2019) for 36,089 distinct charges using 94,708,714 co-occurrence pairs and 146,531,783,286 total relationships. Charges over the first day are converted into a sequence 100 events long and pre-padding with 0's and pre-clipping where necessary.

Two separate model architectures were utilized depending on the type of data utilized: models based on demographics and provider details utilized logistic regression due to the small number of features, while those based on charge data utilized a stacked recurrent neural network (gated recurrent unit (GRU)). Models were trained using the Adam optimizer until convergence based on test accuracy-informed early stopping. Dropout regularization was applied to each model. A table of model hyperparameters is provided in the supplement. All models were trained using the Tensorflow framework.

Evaluation

Models were randomly partitioned into training, validation, and test sets in an 80:10:10 ratio respectively. Validation area under the receiver operating curve (AUROC) was the primary metric for evaluating and comparing model performance.

Chapter 5: Laundering bias: propensity matching and causal reasoning in the surgical literature

Chapter Introduction

Surgery is a domain where observational data has significant potential. The presence of a well-defined intervention from which predictions must be made eliminates many concerns about temporal bias, while a focus on prognostic questions can alleviate some aspects of knowledge parasitism. However, one of the most popular statistical techniques in surgical comparative effectiveness research, propensity matching, has gained the reputation of being a "cure" for bias or confounding. In this chapter, adapted from an article aimed at a surgical audience, we demonstrate the risk of this mindset, and how propensity matching can actually act to launder bias instead.

Main Text

Introduction

It has been well documented that surgical data has unique limitations that, in the absence of advances in recording, monitoring, or data collection, necessitate statistical solutions (Bababekov *et al.*, 2018; Gelman, 2018). Surgical studies often have limited sample sizes at single institutions and enforced standardization is particularly difficult for surgical procedures. To address these limitations, there has been a rise in the use of propensity matching-based studies to conduct comparative analysis of surgical data. In parallel to the growth of large datasets(Haider, Bilimoria and Kibbe, 2018), such as NSQIP or the UK Biobank (Sudlow *et al.*, 2015), the use of propensity matching to reduce bias or confounding has become very popular. In

2018, the share of surgical papers that specifically utilized propensity matching was nearly 100 times higher than levels in 2000.

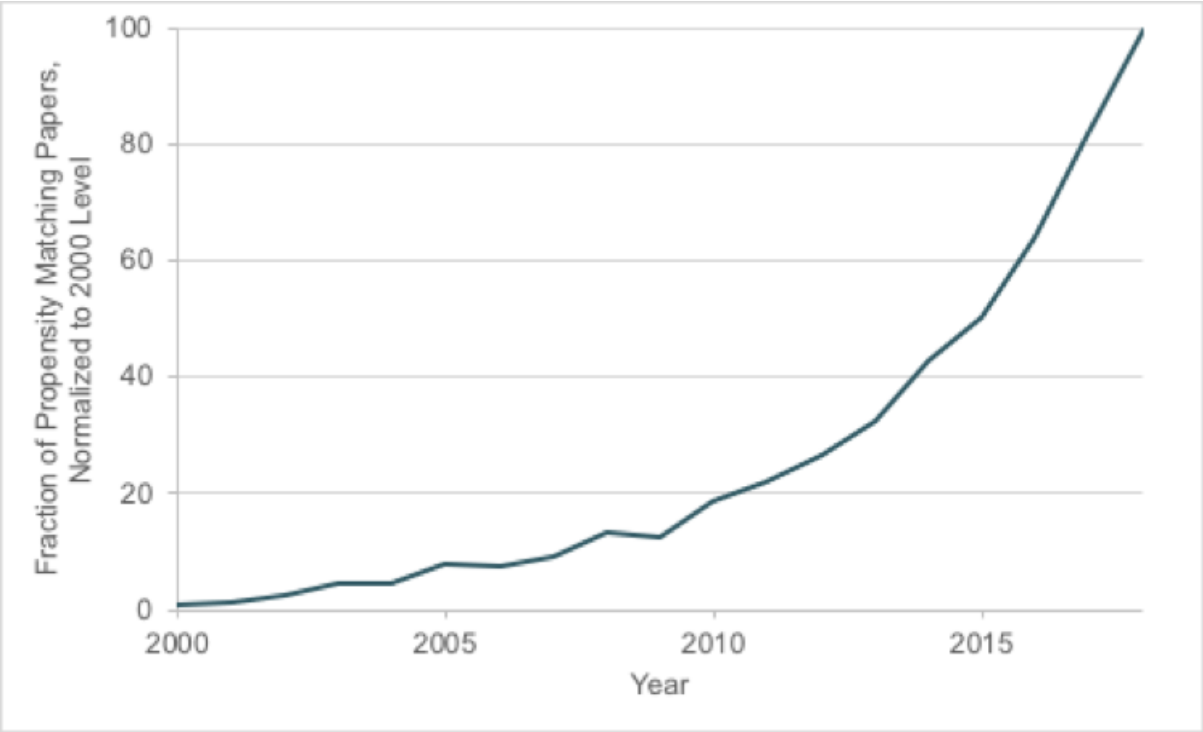


Figure 5.1: Relative Pubmed Publications involving Propensity Matching in Surgery Relative to Growth in Surgical Literature, 2000-2018

Unfortunately, the act of propensity matching does not ensure that the studies utilizing it are free from bias. In fact, given their popular interpretation, propensity score matching may inadvertently be serving to obscure limitations in the datasets and confounders present in the study design, effectively "laundering" bias. Not only can propensity score matching hide existing biases, it can also introduce novel biases of its own, leading to misleading or contrary conclusions later in the analysis. This is a critical issue in surgery, where decisions from limited data can mix with poor research to have serious consequences.

In the current report, we illustrate how propensity matching can provide potentially incorrect conclusions for use of resuscitative endovascular balloon occlusion of the aorta (REBOA), a controversial technique in trauma surgery. REBOA is a method for hemorrhage control that involves the inflation of a balloon in the aorta above the level of injury to reduce bleeding and maintain blood pressure. However, surgery to treat the underlying cause of bleeding is almost always required. While originally conceived for application in battlefield medicine(Stannard, Eliason and Rasmussen, 2011), REBOA use and efficacy in civilian hospitals has recently been the subject of study in several countries using large observational datasets. Four studies in particular attempted to discern the treatment effect of REBOA compared to non-REBOA controls on mortality using propensity score matching and specific inclusion criteria(Norii, Crandall and Terasaka, 2015; Inoue *et al.*, 2016; Otsuka *et al.*, 2018; Joseph *et al.*, 2019). We examined the use of propensity matching in each, and identified oversights in each that could upend conclusions regarding the influence of REBOA on excess mortality.

The issues we identified fell into two central categories:

1. Suppressing data missingness. Consider two patients who present to the ED with identical demographics and vitals- all easily measurable features. In contrast, data on a patient's journey to the ED is less routinely collected. A propensity match using the available data might decide that these patients are equivalent, while neglecting that one had a multi-hour extrication, while the other arrived in the hospital immediately after injury. Propensity matching can act to obscure the quality or scope of the data by collapsing many features into a single matching parameter. Differences that may seem obvious to an attending physician are only visible in the data if the relevant features are deliberately collected.

2. Amplifying bias. The act of using propensity matching with some features may introduce bias where it was previously not present. Consider attempting to measure the effectiveness of ambulatory transfusions on survival, while matching based on surgery type. Because patients who underwent surgery necessarily survived both transport and the ED, patients who expired prior to surgery would not be counted in either group, a bias that would likely reduce the observed benefit of transfusions. When attempting to determine the appropriateness of including a feature in a model, the causal relationships the feature has with the exposure and outcome must be considered.

We used a convenience sampling method to identify 4 major papers (Norii, Crandall and Terasaka, 2015; Inoue *et al.*, 2016; Otsuka *et al.*, 2018; Joseph *et al.*, 2019) in the surgical literature that cited each other as evidence for and against the use of this procedural method for severely injured trauma patients. No effort was made to be comprehensive about the literature review as the purpose of the selection was to serve as an example for our analysis. In the following examples, we identified factors that could raise doubt regarding the presence or strength of REBOA's impact on mortality.

The REBOA-Outcome Effect Can Be Mediated Through Time-to-Surgery

We first examine an instance where the neglecting to include a factor in the propensity match could have allowed confounding to slip through. In this case, the use of propensity matching acted to suppress the data missing from the model.

One research group examining the association between REBOA and excess mortality utilized data from the American College of Surgeons Trauma Quality Improvement Program

data set between 2015 and 2016 (Joseph *et al.*, 2019). The authors found that patients who received REBOA had higher mortality rates than the matched non-REBOA patients and concluded that this excess mortality may be due to the placement of REBOA. They constructed a pre-emergency department profile of each patient, consisting of sex, ethnicity, vital signs upon admission, and the type/number/severity of injuries. These variables form the basis of propensity score matching between REBOA and non-REBOA exposed patients. Crucially, they were only able to utilize factors observed or computed prior to REBOA placement in their analysis, making an implicit assumption that post-REBOA factors were identical between groups. While this is understandable, there are factors after REBOA placement that can act on mortality.

The authors observed that REBOA patients waited significantly longer for their subsequent surgeries compared to the matched non-REBOA patients: a notable exclusion given the well-documented association between time-to-surgery and mortality. To evaluate the relationships between REBOA placement, delay in surgery, and poor patient outcomes, we suggested three plausible narratives: i) the placement of REBOA obstructs blood flow, leading to necrosis in tissue in the lower body, leading to worse patient outcomes (the central hypothesis of the study), ii) placement of REBOA takes a non-trivial amount of time, delaying the surgery, leading to poorer outcomes, and iii) the placement of REBOA temporarily stabilizes a patient, leading them to be triaged differently, delaying their surgery and leading to poorer outcomes. Differential triage could be a result of reduced availability of an on-call surgeon at the time of REBOA placement or a specific decision by the trauma team. The mechanism that the excess mortality operates through is critical in the actionability of the observation. If excess mortality is caused due to physiological action by the REBOA, it could be prudent to discourage its use, but

if the mortality was caused by altered physician behavior, communication regarding this effect could be sufficient.

These narratives are diagrammed in Figure 5.2a. Arrows indicate that the factor at the tail directly causes the factor at the head. We can collapse intermediate terms to view a simplified diagram in Figure 5.2b. This specific effect diagrammed is referred to as effect mediation, as the time-to-surgery mediates the physiological effect of the REBOA on mortality, if indeed such an effect exists.

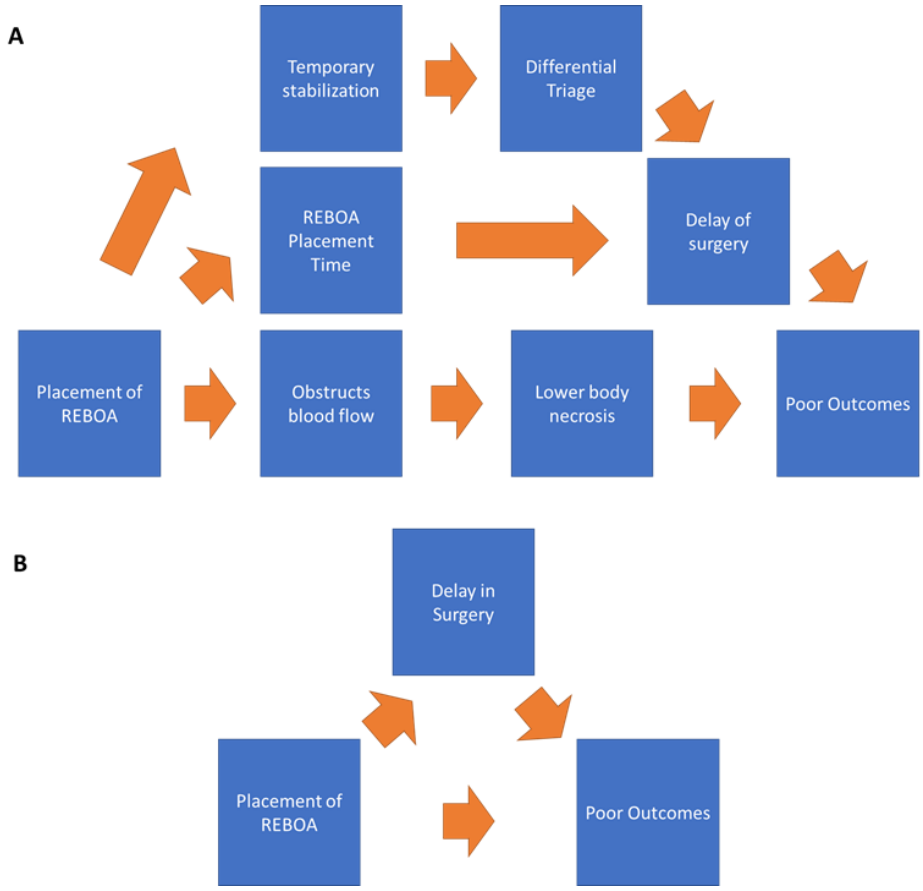


Figure 5.2: Mediation of the REBOA-Outcome Causal Effect by Delay in Surgery

Figure 5.2 summarizes the role of delay in surgery: while the placement of REBOA may cause poorer subsequent outcomes, it also may cause delay in surgery. Therefore, we can hypothesize the existence of an indirect (mediating) effect of REBOA placement on outcomes through delay in surgery. Without controlling for this factor during the matching process, it becomes impossible to determine the relative contributions of the direct physiological and indirect (non-physiological) effects. The assumption that propensity matching alone could eliminate all bias helped obfuscate the assumption that REBOA acted only through physiological mechanisms.

Surgery Type is an Imperfect Proxy for Patient Physiology

We now present an example where the choice of including a feature in a propensity match induced a new bias. One examination(Inoue *et al.*, 2016) of the treatment effect of REBOA utilized records from the Japan Trauma Data Bank (JTDB) (Yokota, 2016). A similar set of factors (demographics, severity/cause/type of injury, vitals, among other factors) were utilized in the propensity score analysis as the other analyses. Uniquely, they included the type of surgery conducted as a measure of patient indication. After taking into account REBOA's role in the ED, this choice ends up embedding a bias into the study that was not originally present.

Figure 5.3 describes the journeys of three hypothetical patients. First, a moderately ill patient is stable enough to proceed to surgery without additional intervention, and goes on to experience better outcomes, consistent with their initial presentation. A second severely ill patient arrives at an institution where REBOA is available, is stabilized sufficiently to proceed to surgery, but ultimately experiences poorer outcomes consistent with their poorer initial presentation. However, imagine a third patient, with identical illness severity as the second

patient, who arrived at an institution where REBOA was unavailable. They might not have survived to receive a surgery, and would therefore be unavailable to serve as a non-REBOA control, despite identical pre-REBOA presentation. If this severely ill cohort represents a non-trivial fraction of the overall non-REBOA population, their systematic exclusion from the controls would cause their group outcomes to seem better. The use of propensity matching, in effect, would be penalizing REBOA for stabilizing critically ill patients enough to receive surgery.

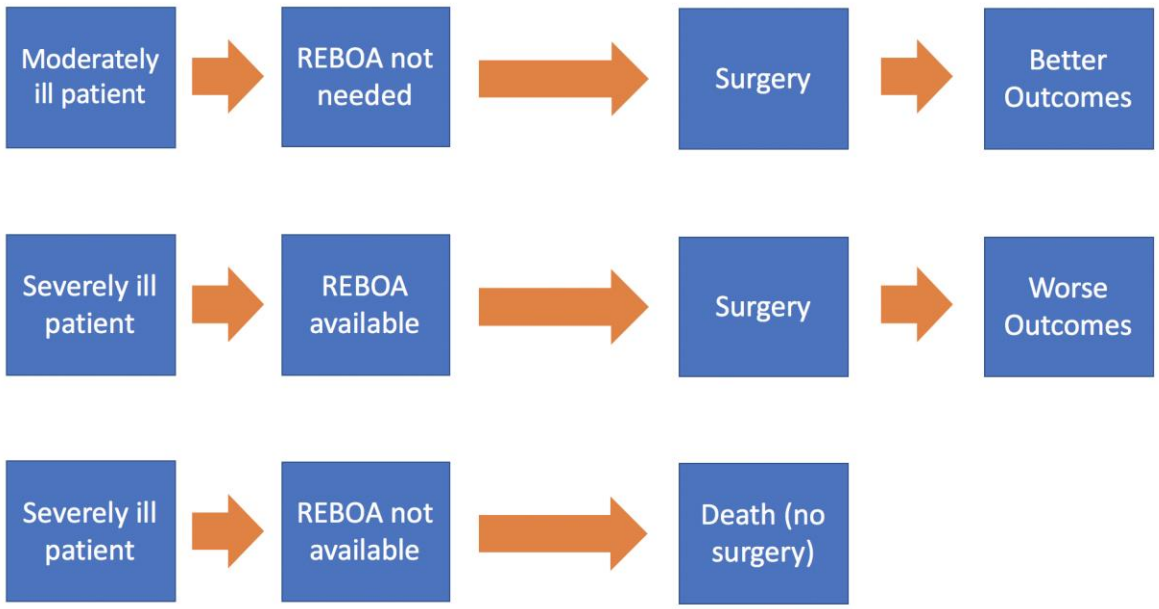


Figure 5.3: Controlling for type of surgery can alter the population of severely ill patients in non REBOA institutions.

Surgery Type is an Imperfect Proxy for Patient Physiology

The act of controlling for surgery type ends up embedding an additional bias of indeterminate direction, due to the multiple care paths that are possible for a given presentation. Using pelvis injury as an example, this can be thought of in two ways.

1) Prior to matching, REBOA patients were significantly more likely to have pelvis hemorrhage surgery compared to pelvis fixation, and vis versa for non-REBOA patients. This encodes the assumption that there are no systematic differences regarding who receives REBOA among patients who receive the same surgery. However, surgery performed is an imperfect proxy for patient indication, particularly when more than one course of treatment for a given injury is possible. Patients who present with crushed pelvis will typically be referred to surgery unless they are stable enough to undergo interventional radiology therapy (IR) instead. However, IR is not available at every hospital. All patients who received both REBOA and pelvis hemorrhage surgery were necessarily severely injured. In contrast, some non-REBOA patients with less severe phenotypes may have arrived at hospitals without IR facilities, resulting in them receiving pelvis hemorrhage surgery. This bias would serve to enrich the non-REBOA cohort with patients with less severe phenotypes. A candidate for IR therapy is known to be an unlikely candidate for REBOA and likely to have positive outcomes. Matching based on surgery rather than indication will embed this effect into a portion of the control population (Figure 5.4a).

2) The non-REBOA population is highly heterogeneous- an opposite causal effect among a different subpopulation could be imagined (Figure 5.4b). There exist patients whose initial indications are so severe that they are immediately sent into surgery without intervening REBOA. For these patients with the most severe phenotypes, knowing that they were sent to surgery could also imply that they were not given REBOA, and, given their initial state, imply that they were more likely to experience poor outcomes. As a result, we are left with two simultaneous, opposite effects induced by the choice to control on surgery type. Understanding the net effect would require knowing the relative size of the IR/severe phenotype subpopulations of the non-REBOA cohort as well as the individual magnitudes of each causal effect.

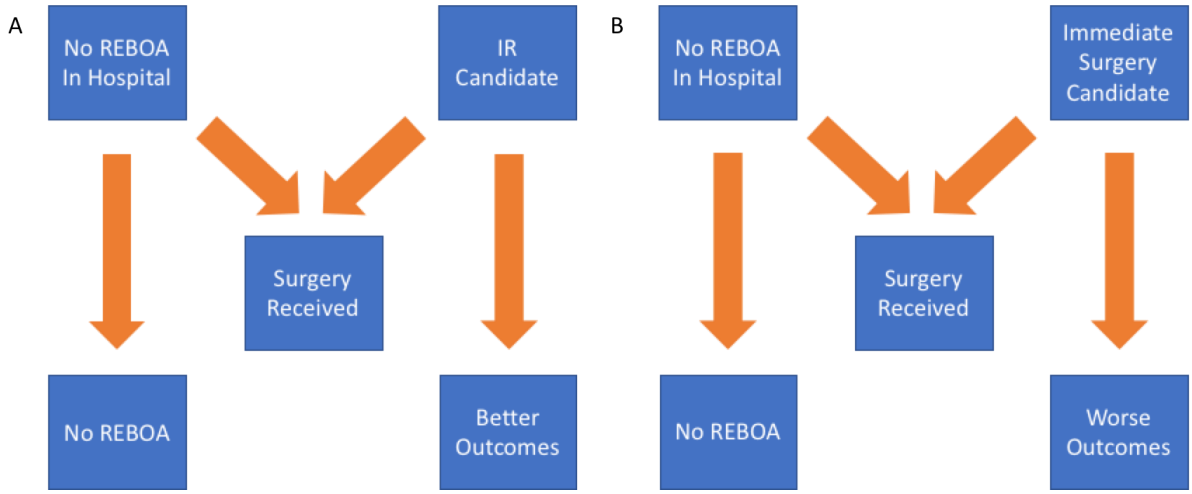


Figure 5.4: Controlling for type of surgery can simultaneously produce a negative (A) and positive (B) treatment effect of REBOA

Bias Through Hospital Facilities is Unavoidable

In other cases, researchers make a deliberate choice to control for a factor to eliminate a particular cause of bias, only to inadvertently induce a new bias.

We examined two studies(Inoue *et al.*, 2016; Otsuka *et al.*, 2018) that only selected patients from hospitals that had the capability to place a REBOA catheter. In the former case, this was an explicit inclusion criterion among JTDB hospitals, while the latter considered only data from a single hospital with the ability to place REBOA. In general, restricting the analysis to only REBOA sites would serve to control for differences between hospitals with and without REBOA. Facilities with capabilities to place REBOA may be larger, more centralized, or better equipped in general, which may influence baseline patient survival (Figure 5.5A). However, within sites who place REBOA, patients are not treated in a random manner. No amount of propensity score matching or adjustment can change the fact that a physician observed the non-REBOA patients in these studies and chose not to place a REBOA. The REBOA cases and

available controls are two fundamentally different populations. The fact that they did not have a REBOA placed implies that they a physician determined that REBOA was not necessary upon examination, implying that they would have better outcomes from the outset. The authors are consequently left with a choice regarding which bias to let through: controlling for one induces the other, and vis versa.

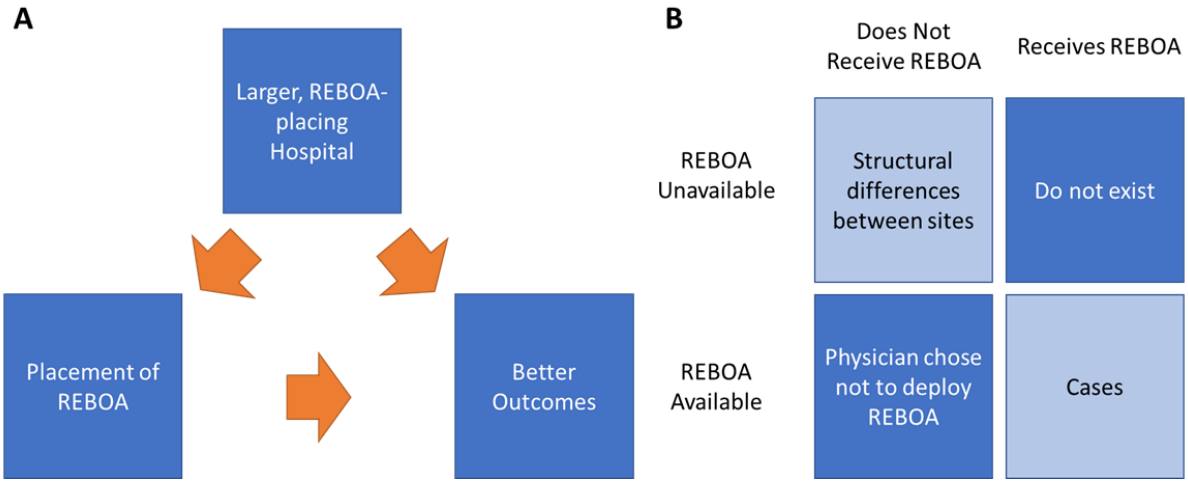


Figure 5.5: Controlling for location bias induces physician-mediated selection bias. **A)** Controlling for the Confounding Factor of Hospital Type **B)** Loss of Exchangeability through Hospital Inclusion

In this case, part of the challenge stems from selection of the question at hand. The central question of evaluating the treatment effect of REBOA relative to non-REBOA controls is an unrealistic one, given both i) the extreme heterogeneity of the non-REBOA cohort and ii) the artificial framing of the study question. Even in the most ambiguous cases, surgeons do not choose between REBOA and non-REBOA, but REBOA and a specific alternative, such as resuscitative thoracotomy. This fact is obscured when the non-REBOA group is described as "propensity-matched." Given the surgical contexts in which REBOA is currently deployed and

the extreme heterogeneity among “non-REBOA groups”, it is unlikely that existing observational datasets can be used to evaluate the treatment effect of REBOA relative to non-REBOA controls.

Discussion

The central risk of an over-reliance on propensity-matched observational studies is unjustified confidence in the strength of the discovered relationships. An understanding of how the features utilized in a particular study align with real life is necessary to determine if a given study can properly be conducted with a particular dataset, or if additional data must be collected. Propensity matching alone cannot compensate for missing features or identify causal relationships. Ultimately, the onus must be with researchers to produce work that is cognizant of its own potential limitations. In those cases where data limitations are not insurmountable, researchers should strive to generate associations that are useful to clinicians or other researchers without extensive reanalysis of the experimental design.

When considering big data research findings that evaluate the efficacy of an intervention, it is critical for clinicians to decide whether a particular result applies to a given patient. This is currently a largely *ad hoc,* clinician-directed process. This results in a conundrum for clinicians, particularly when research findings contradict the way in which a patient presents. The trend of big data driven papers promoting views that are the result of assumptions that aren’t explicit has driven this phenomenon. The lack of transparency surrounding the appropriate interpretation of research findings makes it difficult for an otherwise educated population to weigh the relative value and limitations of each study. What research in this space so often fails to address is how assumptions change directionality or effect size. In this increasingly evidence-driven era of big

data, it is critical for researchers to provide clinicians with the tools to critically evaluate boundaries of what is published and promoted.

Often, the incentives surrounding research publication facilitate experimental designs that inadvertently accentuate structural differences between case and control populations to produce larger and more dramatic effect sizes or studies with unrealistically broad scope. In contrast, the problems that clinicians face where research can provide the most guidance are often very specific with ambiguous prior evidence. Propensity matching ends up functioning as an obfuscation tool, covering up bias where it exists while lending a veneer of credibility.

Ultimately, the promise of observational research hinges on our collective ability to transition learnings and conclusions to the clinic when appropriate and to further experimental testing when needed. Making this determination in an effective manner hinges on a partnership between researchers and physicians and transparent communication regarding what experiments were done and what findings result.

Chapter 6: Recommendations for Transparency and Frameworks in Surgical Research

Chapter Introduction

This chapter, adapted from a companion piece to Chapter 5, aims to justify the need for transparency in communication and provide practical guidance for reporting model design, aiming to deflate the public perception that matching has achieved among surgical audiences.

Main Text

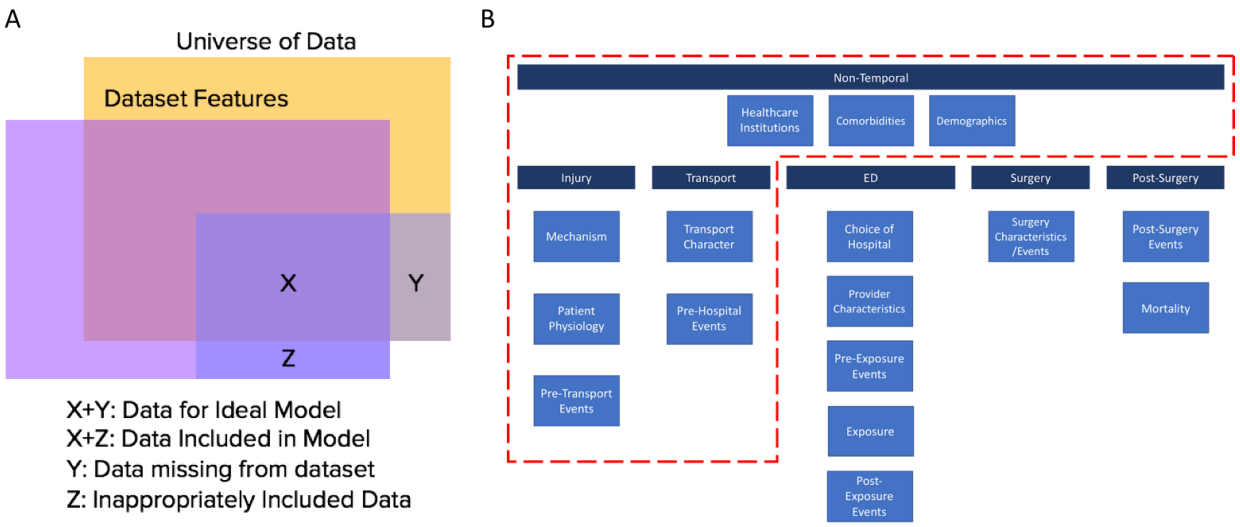
The proliferation of large, observational datasets and the statistical techniques used to study them has coincided with a rising interest among physician-researchers to apply these tools to understand complex systems (Weiss *et al.*, 2012). Unfortunately, this can often lead to an uncritical acceptance of research findings that are generated in this manner. A perception that algorithmically created recommendations are "less biased" or that a large dataset should be allowed to "stand on its own" implicitly concedes that the processes behind the question at hand are beyond comprehension. However, the ability to justify decisions and recommendations is foundational to medicine. The infrastructure surrounding best practice can only exist because decision-making processes can be articulated and these processes represent reproducible phenomena.

Consider the problem of determining whether to intubate an acutely ill trauma patient in the emergency department (ED), and how observational analysis might provide useful guidance. A naive comparison of the patients who were/were not intubated might lead to the observation of better outcomes in one group over the other, but this observation will not be actionable unless i) a clinician can have a sense of why the association exists and what moderating factors may exist and ii) a clinician knows for which patients or situations the association applies. Consequently,

we believe that the limitations and mechanisms that are integral parts of conclusions drawn from observational research can be encouraged by transparency in describing identifiable assumptions utilized as part of a study.

The Strengthening the Reporting of Observational studies in Epidemiology (Vandenbroucke *et al.*, 2014) (STROBE) guidelines are a useful tool in the reporting of medical research, but barriers for effective communication of limitations and mechanisms to clinicians still exist. We therefore propose amendments to enable researchers to more clearly present the assumptions implicit in their studies, and also appropriately consider the contexts in which their findings should and should not be applied.

1. Within Study Design (4): presenting a framework: what are all the factors the authors think affect exposure and outcome (Table 6.1)? What factors do they have or are missing? What is the implication of feature missingness on their conclusions?
2. Within Variables (7): presentation of a causal narrative or directed acyclic graph for each variable chosen to include and exclude from analysis.
3. Within Interpretation (20) and Generalizability (21): proposing a testable mechanism for any observed significant associations and proposing, with as much specificity as possible, which patients are subject to study and where the learning from this study would be applied.



C

Factor	Causal Narrative	Decision
Injury Nature (penetrating vs non penetrating)	Penetrating injuries more likely to: • require intubation • produce worse outcomes.	Include in match to distinguish if poor outcomes are from injury nature or intubation decision.
Total volume of blood transfused	Total blood transfused affected by: • initial presentation • subsequent treatment, including intubation	Exclude from match, unless pre-hospital transfusion data are available.

Figure 6.1: Summary of recommendations for STROBE **A)** A comparison of the dataset features with the Universe of Data, a framework that describes the author's conception of all possible features that could influence exposure or outcome, can reveal deficiencies in the dataset or model. **B)** An example framework for exposures in emergency surgery. Terms highlighted in red are often collapsed into "patient state upon arrival". **C)** Example table of features considered for inclusion/exclusion to the model.

Table 6.1: Examples of features by class

Feature Class	Examples of features
Non-Temporal: Healthcare Institutions	Available health care/medical/pharmaceutical institutions, general practices
Non-Temporal: Comorbidities	Presence of heart disease, diabetes, obesity, etc. in patients
Non-Temporal: Demographics	Age, sex, genetic background
Injury: Mechanism	Piercing vs blunt force, location of injury
Injury: Pre-Transport Events	Patient physiology, time-to-transport, availability/type of first aid
Transport: Transport Character	Ambulance, helicopter, types/expertise of emergency medical services
Transport: Pre-Hospital Events	Patient physiology, time-to-hospital, types of care administered in transport
ED: Choice of Hospital	Distance, presence in dataset (ascertainment bias), size, availability of surgeons/facilities/procedures
ED: Provider Characteristics	Experience level of provider, comfort/preferences regarding procedures
ED: Pre-Exposure Events	Patient physiology, time-to-exposure (if any)

ED: Exposure	Intubation vs. No Intubation
ED: Post-Exposure Events	Time-to-surgery
Surgery: Surgery Characteristics/Events	Types of surgery, comparative risk of complications between procedures
Post-Surgery: Post-Surgery Events	Complications, post-operative care
Post-Surgery: Mortality	Mortality

We first propose for physicians and researchers to think about their datasets as a pool of data that exists within a framework that has direction, assumptions, and limitations. This framework exists as a hypothetical 'universe of data' that surrounds data collected for a specific intervention/exposure and outcome. This universe contains every feature that could conceivably affect the exposure or outcome, even if they cannot necessarily be observed or measured, and represents the author's working model of the factors in play (Figure 6.1A, Table 6.1). In the example of evaluating the effectiveness of an acutely injured patient requiring possible intubation and surgery, this may take the form of a timeline detailing a patient's journey through their encounter (Figure 6.1B, Injury, Transport, ED, Surgery, Post-Surgery), along with a separate set of non-temporal observations, such as comorbidities or demographics. By conceptualizing this universe of relevant factors around an intervention, it is significantly easier to visualize biases, assumptions, and putative mechanisms present in an analysis. These assumptions most frequently manifest in the form of i) the appropriateness of the dataset for the question and ii) the choice of covariate control. As an example, hospital monitoring datasets

typically do not contain annotations of patient deterioration during transport- it is easy to imagine that intubation would be differentially effective on patients depending on how soon after injury that it is utilized. Avoiding this bias would require more precise cohort definitions or an alternative dataset and highlights the instances where certain datasets are not powered to answer certain questions. Identifying the gap between desired and available data is critical to understanding what potential conclusions can be drawn, and when they are applicable. Readers or reviewers could compare their own internal frameworks to those presented by the authors to evaluate the appropriateness of presented findings to their own contexts.

Our second proposal involves a systematic treatment of inclusion criteria and propensity matching factors through the presentation of a causal narrative justifying its presence or absence in a model. This could take the form of a table, presented in the appendix, that would allow a clinician to easily understand the scope and limitations of the presented models (Figure 6.1C). In the case of intubation, the intuition behind controlling for injury location can be clarified by the observation that injury location exists prior to both intubation and mortality and possesses plausible mechanisms of influence over both, confirming its status as a confounding factor to be controlled. A less trivial example involves time-to-surgery, which is influenced by the manner in which intubation is delivered, and has a strong influence over mortality. Controlling for time-to-surgery could be warranted if the authors hoped to disentangle whether the benefits of intubation outweigh the delay to the surgery.

These considerations set the groundwork for the precise definition of clinically relevant cohorts as well as mechanistic, testable hypotheses for why the findings were observed. These elements are critical for allowing readers to determine the appropriateness and applicability of a study to their patients. For instance, lacking other data on patient physiology during transport,

this intubation study might have restricted itself to patients with comparatively less severe presentations who could be stabilized while in an ambulance. A reader would consequently know not to apply the findings of said study to the most severe phenotypes. Finally, by encouraging researchers to be upfront regarding their perception of their experimental systems, other researchers can compare their own sets of assumptions to those made by the study authors. An observation that intubation was associated with increased subsequent mortality would be easier to accept if a plausible, testable mechanism was supported by the data.

The inductive power that big data and statistical learning can provide can obscure the deductive processes of medical research and clinical translation. By encouraging transparency of assumptions and promoting the proposal of mechanistic hypotheses, frameworks empower physicians to identify which research findings are relevant to their patients.

Chapter 7: Association of Bariatric Surgery with Subsequent Depression

Chapter Introduction

This chapter collects the recommendations described previously into a single, constructive, study, examining the association of depression onset after bariatric surgery.

- By focusing on depression onset after surgery, there is a well-defined context and audience for the findings, and any results would be immune to temporal bias due to the fixed index date.
- Rather than making individualized predictions, this study aims to examine population-level prognosis, sidestepping concerns about knowledge parasitism.
- In accordance with the specific recommendations regarding frameworks and matching features, details and rationales regarding the study design are presented. Stratified experiments are also conducted to provide mechanistic hypotheses for observed associations.

Main Text

Introduction

Bariatric surgery is recognized as an effective treatment for severe obesity, resulting in large, sustained weight loss and improved quality of life (Nguyen and Varela, 2017) . Post-operative depression has been implicated as a predictor of poor overall weight loss after surgery. As a result, evaluating the relationship between bariatric surgery and depression is important in ensuring the success of the procedure (Sheets *et al.*, 2015). However, it is difficult to separate the potential of bariatric surgery to reduce pre-existing depressive symptoms from reduction of depression after surgery at the population level due to confounding by the presence/absence of

pre-existing mood disorders. This analysis is further complicated by the bi-directional associations between depression and obesity (Luppino *et al.*, 2010). Previous studies have tended to focus on the reduction in depressive symptoms in the same individuals before and after bariatric surgery, generally reporting a reduction in depression rates between 55-65% over two years (Burgmer *et al.*, 2014; Mitchell *et al.*, 2014; Ivezaj and Grilo, 2015). The Swedish Obese Subjects (SOS) study reported long-term reductions in depression among bariatric surgery patients relative to conventionally treated patients, but the studied individuals seeking surgery had higher baseline incidence of depression relative to the control population (Rydén and Torgerson, 2006). In contrast, an increase in suicide among gastric bypass surgery patients relative to matched nonsurgical controls has been reported (Kennedy, 2008), but this observation was similarly attributed to the presence of preoperative depression (Jones-Corneille, Wadden and Sarwer, 2007). A separate study of patients in Pennsylvania implicated both disappointment with weight regain and lack of follow-up appointments with the association between bariatric surgery and suicide (Tindle *et al.*, 2010).

A clear understanding of the influence of bariatric surgery on post-surgical depression risk will assist the determination of ideal candidates for weight loss surgery from a psychological standpoint, and inform surgical follow-up standards of care. We performed a causal-inference analysis of post-surgical rates of depression in populations undergoing and eligible for bariatric surgery (body mass index (BMI) ≥ 40 or BMI ≥ 35 with a comorbid condition) (Hedley, 2004), and without a history of depression. To our knowledge, our utilization of health insurance claims data for this analysis represents the largest such study to date.

Methods

Using un-identifiable member claims data from Aetna, a prospective cohort study was simulated. The claims dataset contained diagnosis and intervention records for more than 63 million individuals in the United States between 2008 and 2016. Race, ethnicity, and socioeconomic data were not present in the database. International Classification of Diseases, Ninth Revision (ICD-9) codes were used to define diagnoses, while ICD-9 and Current Procedural Terminology (CPT) codes were used to define procedures and interventions. Phenome-wide association study (PheWAS) (Hedley, 2004; Denny *et al.*, 2010) codes were used to map ICD-9 codes according to phenotype. All calculations were conducted using Microsoft SQL Server and R statistical software, version 3.4.3 (Hedley, 2004; Denny *et al.*, 2010; Tierney, 2012). R packages survival and data.table were also used. The Harvard Medical School Institutional Review Board waived the approval requirement, as it determined this analysis of the dataset not to be human subjects research.

Using annotations of diagnoses and procedures, all individuals who were eligible for and who underwent bariatric surgery were identified, with relevant codes and criteria defined using published UnitedHealthcare Commercial Medical Policy (UnitedHealthcare, 2018). Patients undergoing non-bariatric abdominal surgery were identified using CPT codes for anesthesia for abdominal surgeries. For patients with multiple surgeries, the earliest date available was used. Non-surgical groups were assigned a placeholder surgery date to enable comparisons with surgical groups. These dates were calculated by identifying the earliest eligibility date and adding the median eligibility-to-surgery time observed in the bariatric surgery cohort. A similar process was conducted for comparisons against non-bariatric abdominal surgeries, identified using anesthesia codes. Some of the most common surgeries in this category included

cholecystectomies, appendectomies, and hernia repairs. Individuals were required to have at least 6 months of observations prior to their surgery/placeholder date to be included in the study. An analysis of all depression diagnoses in our dataset found that that the mean number of days between depression diagnoses was 58 days, and that 93% of all depression diagnoses were separated by fewer than 6 months, indicating that it was unlikely that unobserved depression diagnoses prior to the observation period were frequent. Subsequent diagnoses of depression were identified using phenotype-level codes corresponding to "Depression" or "Major Depressive Disorder" (Table S7.3). Individuals with codes corresponding to diagnoses of depression prior to their surgery/placeholder date were excluded from the analysis.

To examine the effect of bariatric surgery on subsequent depression diagnosis, Cox proportional hazard models and Kaplan-Meyer cumulative incidence estimates were used to evaluate hazard ratios for depression between three pairs of groups (Table 7.1):

- Bariatric surgery patients vs. bariatric surgery eligible individuals who did not receive bariatric surgery (referred to here as "surgery eligible individuals"). Note that while bariatric surgery patients are also technically eligible for surgery, the phrase "surgery eligible individuals" will refer to those who did not receive bariatric surgery.
- Bariatric surgery patients vs. surgery eligible individuals who received non-bariatric abdominal surgery (referred to as "other abdominal surgery patients"). Patients with both bariatric surgeries and non-bariatric abdominal surgeries were placed in the bariatric surgery cohort.
- Other abdominal surgery patients vs. surgery eligible individuals who received no abdominal surgeries, bariatric or otherwise (referred to as "non-surgery individuals").

Table 7.1: Study Cohort Distribution

All individuals considered are either bariatric surgery patients or eligible for bariatric surgery. While bariatric surgery patents are also eligible for bariatric surgery, the term "surgery eligible" refers to individuals who did not receive bariatric surgery in this analysis. (Table continued on next page)

	Number (%)		
	Bariatric Surgery Patients	Bariatric Surgery Patients	
		Prior Psych Evaluation	No Prior Psych Evaluation
Total	64090	25861	38229
Men	18403 (28.7)	6983 (27)	11420 (29.87)
Age, years, mean (SD)	46.19 (13.59)	43.86 (11.52)	47.76 (14.63)
Post-Surgical Depression Diagnosis (>=1)	7421 (11.57)	2647 (10.24)	4774 (12.49)
Protracted Post-Surgical Depression Occurrences (>= 3 Diagnoses/6 Months)	2550 (3.98)	951 (3.68)	1599 (4.18)
BMI, mean (SD)	44.76 (7.08)	45.37 (7.13)	44.12 (6.98)
6 Month Code Count, mean (SD)	120.63 (123.64)	121.71 (92)	119.9 (141.07)
6 Month Diagnosis Count, mean (SD)	56.21 (55.04)	58.84 (41.96)	54.44 (62.29)
6 Month Procedure Count, mean (SD)	64.42 (71.64)	62.87 (52.99)	65.47 (81.87)
Follow-up Time, days, mean (SD)	748.78 (665.53)	716.38 (647.54)	770.71 (676.56)

	Number (%)		
	Bariatric Eligible Individuals	Bariatric Eligible Individuals	
		Other Abdominal Surgery Patients	Non-Surgery Individuals
Total	713050	220706	492344
Men	327413 (45.92)	97702 (44.26)	229711 (46.65)
Age, years, mean (SD)	53.22 (15.59)	57.69 (12.63)	51.62 (16.66)
Post-Surgical Depression Diagnosis (>=1)	64932 (9.11)	13843 (6.27)	51089 (10.38)
Protracted Post-Surgical Depression Occurrences (>= 3 Diagnoses/6 Months)	20051 (2.81)	3991 (1.8)	16070 (3.26)
BMI, mean (SD)	41.58 (7.06)	41.15 (6.95)	41.8 (7.11)
6 Month Code Count, mean (SD)	53.43 (83.12)	103.06 (132.03)	46.52 (74.16)
6 Month Diagnosis Count, mean (SD)	24.77 (36.93)	48.07 (60.34)	24.86 (42.86)
6 Month Procedure Count, mean (SD)	28.67 (48.28)	54.99 (74.7)	24.86 (33.20)
Follow-up Time, days, mean (SD)	892.33 (654.27)	857.24 (646.68)	985.22 (640.32)

Sex, age, and pre-surgical ICD code count (6 months) were treated as covariates in all hazard-ratio models, and sex-stratified multivariate analyses were conducted. Analyses were further stratified by bariatric surgery type and whether bariatric surgery patients had annotations of psychiatric evaluation 6 months prior to the surgery.

To identify what phenotypes were associated with post-bariatric surgery depression risk, propensity matched case/control groups were created, using both surgery eligible and other abdominal surgery patient cohorts as controls. Sex, age, pre-surgical ICD code count (6 months), and ZIP code were included as parameters in the match. Pre- and post-surgical phenotypes were

examined separately. Phenotype counts and depression status were recorded for all patents in all cohorts, and phenotype-specific risk ratios for depression were computed. Only phenotypes that (i) were found to be significantly associated with depression in the bariatric surgery cohort, (ii) were not found to be significantly associated with depression in either of the control cohorts, and (iii) had at least 100 co-occurring depression diagnoses were examined. Bonferroni correction was used to correct for multiple hypothesis testing. Details on variable selection and experimental design are included in the appendix.

Results

In total, 777 140 individuals were considered, including 64 090 bariatric surgery patients (Table 7.1). Among all considered individuals, there were 72 353 individuals diagnosed with depression (PheWAS code in 296.2 group), for a population incidence of 9.3%. This is lower than previous estimates[15] of depression prevalence among individuals with obesity, but can be rationalized by the exclusion of individuals with histories of depression prior to surgery from the analysis. Among bariatric surgery patients, 7 421 subsequent depression diagnoses were recorded, for a population incidence of 11.57%. The mean follow-up time for bariatric surgery patients was 748 days, compared to 892 days for bariatric eligible individuals and 985 days for non-bariatric abdominal surgery patients.

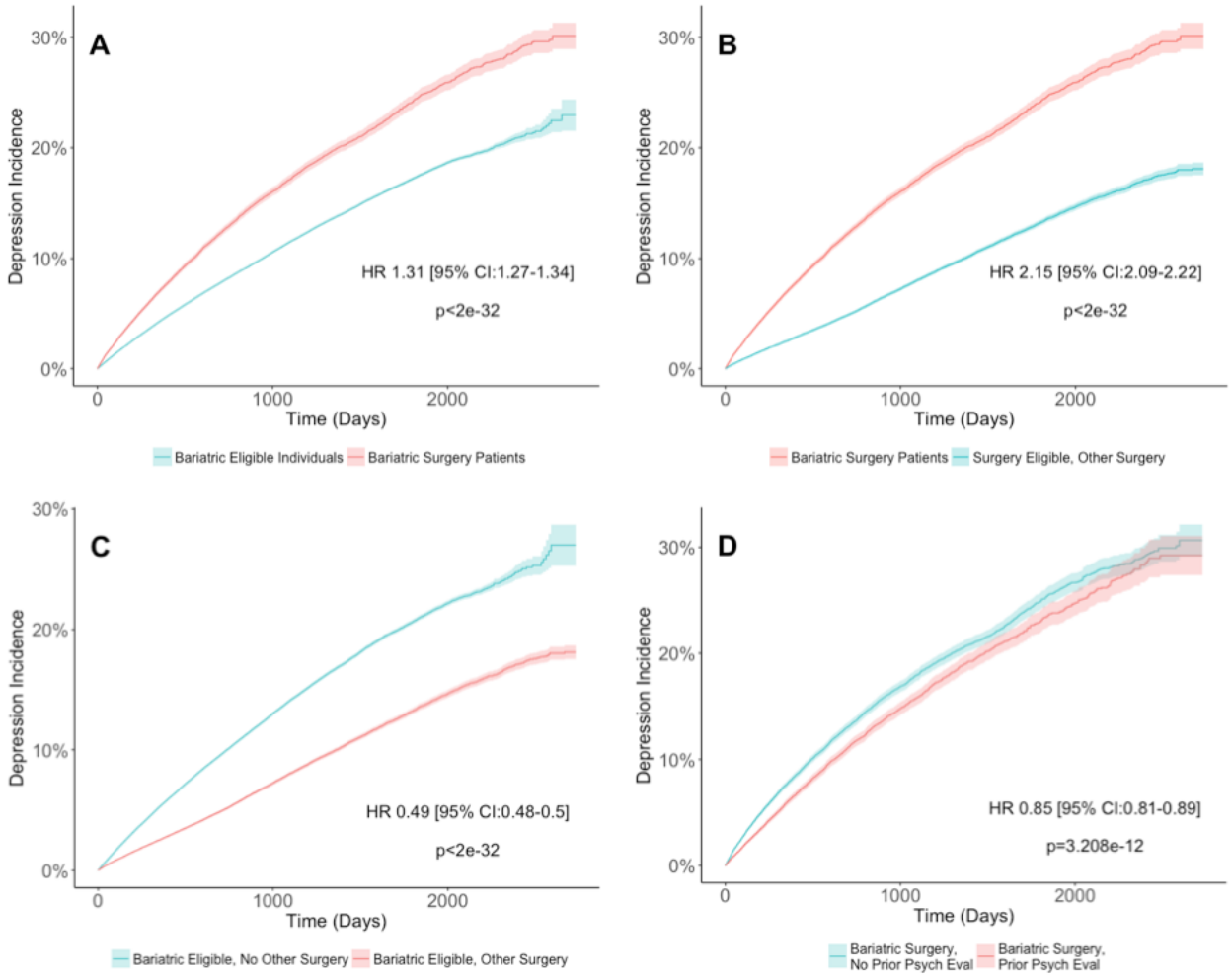


Figure 7.1: Incidence of Post-Surgical Depression
Text represents the hazard ratio and confidence interval between the two plotted cohorts. A) Time-to-depression curve for bariatric surgery patients compared to bariatric eligible individuals. B) Time-to depression curve for bariatric surgery patients compared to bariatric eligible individuals with other abdominal surgeries. C) Time-to-depression curve for bariatric eligible individuals with other abdominal surgeries compared non-surgery individuals. D) Time-to-depression curve for bariatric surgery patients with pre-surgical psychiatric evaluations compared to bariatric surgery patients without pre-surgical psychiatric evaluations. All hazard ratios are adjusted for sex, age, and 6-month claim count.

Figure 7.1 summarizes the results of Cox regression models over the patient cohorts. Bariatric surgery was found to have a hazard ratio of 1.31 (95% CI, 1.27-1.34, P < 2e-32) towards subsequent depression when compared to surgery eligible individuals. Furthermore, bariatric surgery was found to have a hazard ratio of 2.15 (95% CI, 2.09-2.22, P < 2e-32)

compared to other abdominal surgery patients. Other abdominal surgery patients had a hazard ratio of 0.49 (95% CI, 0.48-0.50, P < 2e-32) relative to non-surgery individuals. Age and code count 6 months prior to surgery were found to have minor effects (hazard ratios between 0.98-1.02, all P< 2e-32) on the hazard ratio. Pre-surgical BMI measurements were available for a subset of patients examined (n = 132 000). Within this cohort, BMI was observed to have a small but statistically significant hazard ratio (1.01, P = 6.1e-12) with respect to depression (Figure S7.2). As an additional point of comparison, the depression hazard ratio for patients undergoing laparoscopic bariatric surgeries (n = 58 536) compared to bariatric eligible individuals undergoing non-bariatric laparoscopic surgeries of the stomach and esophagus (n = 2 679) was found to be 1.39 (95% CI, 1.30-1.50, P < 2e-32), consistent with the finding that the non-bariatric laparoscopic surgery cohort had no elevated risk of depression relative to the non-surgical group (HR = 1.03, 95% CI, 0.98-1.09, P < 2e-32) (Table S7.2).

Psychiatric evaluation prior to bariatric surgery is commonly recommended (Brolin, 1996; Roberts *et al.*, 2000). The effect of psychiatric evaluations or testing 6 months prior to bariatric surgery on the risk of subsequent depression was examined. Relative to bariatric surgery patients without pre-surgical psychiatric evaluations/tests, patients who received them (n = 25 861) prior to bariatric surgery had a hazard ratio of 0.85 (95% CI, 0.81-0.89, P = 3.208e-12). For this study, depression was defined as a single observation of a depression code patient's record. These annotations do not make reference to the severity or degree of diagnosis. The prevalence of protracted depression, defined as 3 or more diagnoses of depression within a 6-month period, was examined with respect to bariatric surgery. Although the absolute incidence of protracted depression was lower compared to standard depression, similar trends in hazard ratio were observed (Figure S7.1).

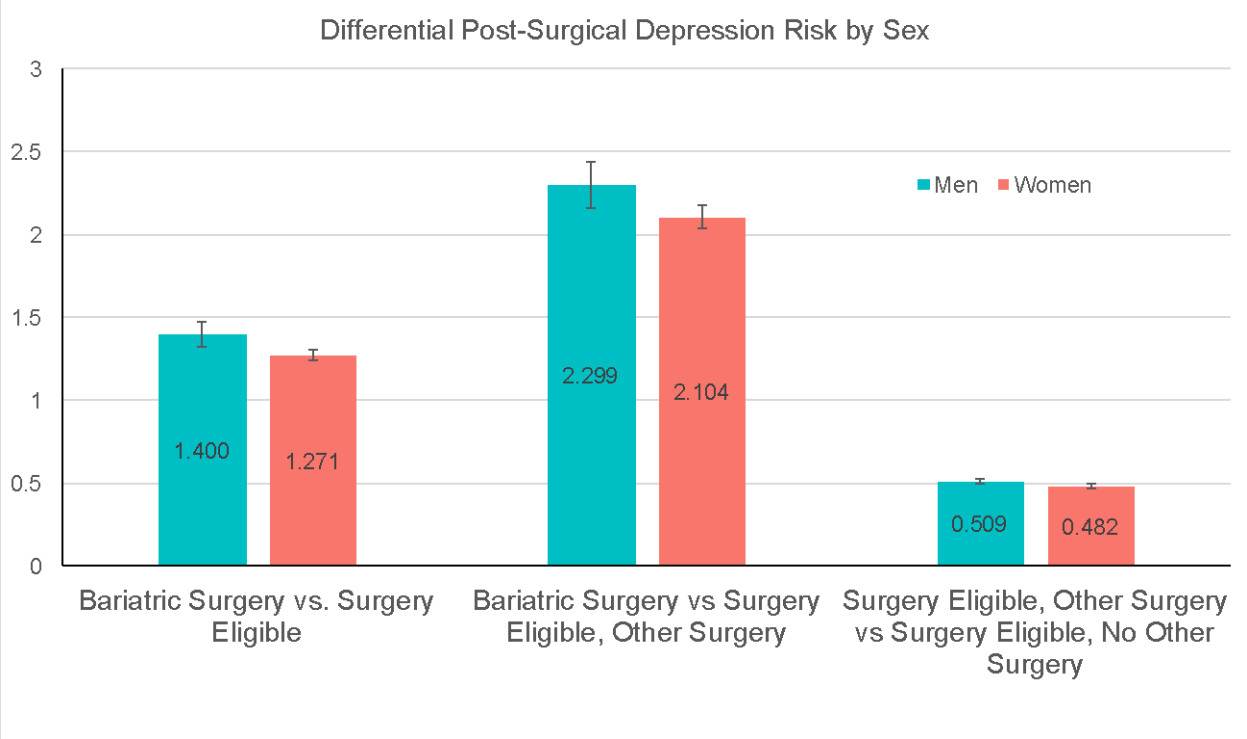


Figure 7.2: Sex-Stratified Time-to-Depression Diagnosis Curves
Red curves represent the case groups in each plot (bariatric surgery or other abdominal surgery patients) while blue represent the control groups (surgery eligible individuals, other abdominal surgery patients, non-surgery individuals, respectively). Error bars represent 95% confidence intervals.

Figure 7.2 shows the hazard ratios for the bariatric surgery comparisons stratified by sex. Consistent with previous studies (Mechanick *et al.*, 2013), the rate of depression was found to be higher among female individuals, but men were typically more susceptible to a post-bariatric surgery depression effect. With respect to surgery eligible patients, male patients undergoing bariatric surgery had a hazard ratio of 1.40 (95% CI, 1.329-1.475, P < 2e-32) while female patients had a hazard ratio of 1.271 (95% CI, 1.236-1.307, P < 2e-32) (Supplementary Figures 3-4). Compared to the other abdominal surgery cohort, male patients undergoing bariatric surgery had a hazard ratio of 2.299 (95% CI, 2.165-2.442, P < 2e-32) while female patients had a hazard ratio of 2.104 (95% CI, 2.031-2.178, P < 2e-32). The effect of sex on postsurgical depression

diagnosis risk was significantly smaller in magnitude when considering non-bariatric surgeries. When evaluating the hazard ratios for other abdominal surgery patients compared to non-surgery individuals, the hazard ratio for men was 0.509 (95% CI, 0.492-0.526 P < 2e-32) compared to 0.482 (95% CI, 0.471-0.494, P < 2e-32) for women. The difference in hazard ratios between men and women for all comparisons was found to be significant with $p < 0.0001$ based on bootstrap analysis.

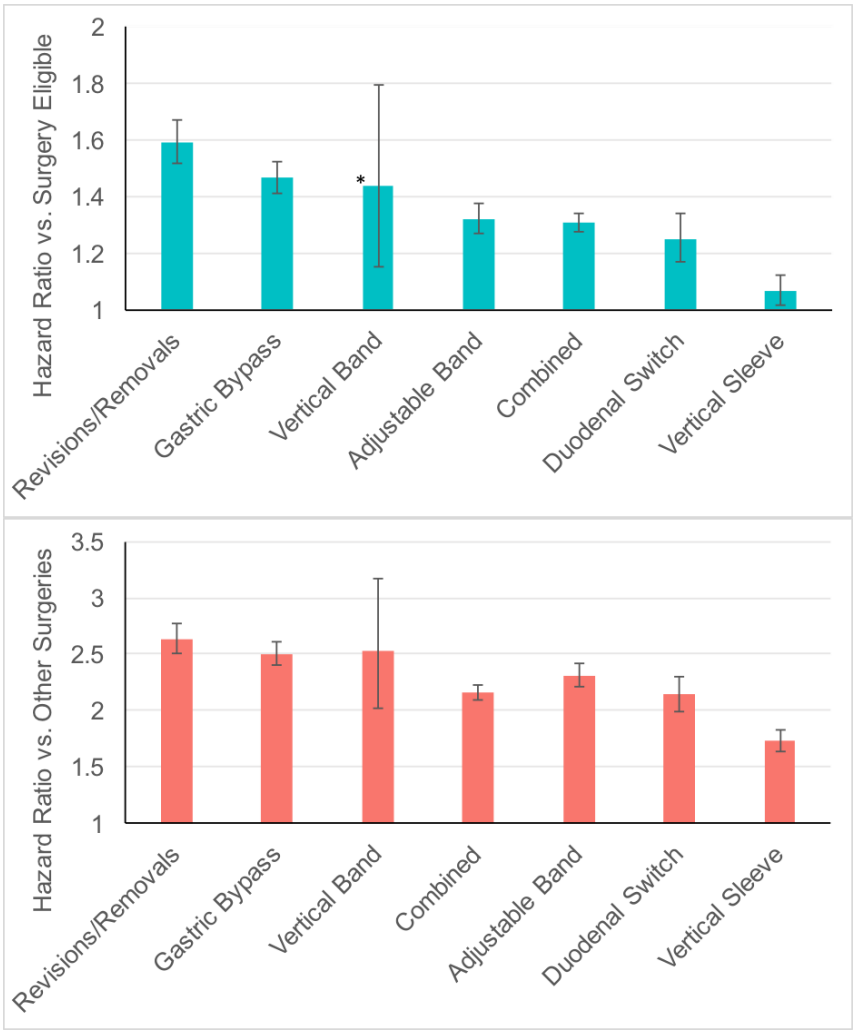


Figure 7.3: Procedure-Stratified Time-to-Depression Hazard Ratios

The error bars represent 95% confidence intervals. Blue bars represent hazard ratios relative to surgery eligible individuals, while red bars hazard ratios relative to other abdominal surgery patients. The asterisk represents a non-statistically significant hazard ratio.

Figure 7.3 describes the hazard ratios of post-surgical depression stratified by surgery type. Individuals with more than one surgery were classified based on the last annotated surgery present in their record. The individual surgical cohorts are summarized in Table S7.1. Vertical sleeve gastrectomy (VSG) surgical patients (n = 19 852) had significantly lower rates of post-

surgical depression compared to the overall population, while Roux-en Y Gastric Bypass (RYGB) surgical patients (n = 18 877) and patients who underwent revision/reversal surgeries (n = 12 319) had significantly higher rates. Other surgical groups included adjustable bands (n = 15 799), vertical bands (n = 506), and duodenal switches (n = 6 686). Consistent with the group-level sex-stratified analysis, in every examined surgery type, men had higher hazard ratios than women. In particular, a significant hazard ratio for depression was not observed for female VSG patients relative to female bariatric eligible individuals (Supplementary Figures 5-8).

Table 7.2: Phenotypes Associated with Depression in Bariatric Surgery Patients Only
Phenotypes are grouped by PheWAS code. Depression risk ratio is observed in the bariatric patient cohort. A p-value threshold of 3E-05 was used to account for multiple hypothesis testing.

Presurgical Phenotype	Depression Risk Ratio	P-Value
Memory Loss	1.997	8.11E-08
Chronic Airway Obstruction	1.408	6.58E-07
Postsurgical Phenotype	**Depression Risk Ratio**	**P-Value**
Sleep Disorders	2.043	1.95E-12
Peritonitis and retroperitoneal infections	1.783	8.42E-10
Postoperative infection	1.638	4.83E-09
Complications of medical procedures NOS	1.71	4.58E-08
Pleurisy; pleural effusion	1.484	4.58E-08
Cellulitis and abscess of trunk	1.496	1.58E-05

Table 7.2 summarizes the results of PheWAS analysis of post-surgical bariatric surgery, split between pre- and post-surgical phenotypes. Phenotypes were only included in these lists if they were not found to have significant associations with depression in surgery eligible and other abdominal surgery cohorts. Post-surgical phenotypes related to infections and surgical complications were significantly associated with depression in the bariatric surgery group but not in the cohort who underwent other abdominal surgeries, implying a specific association with the specific bariatric surgeries undergone by the patients with subsequent depression.

Discussion

We report a robust association between bariatric surgery and subsequent depression. Because the follow-up times in the control cohorts were longer than those in the cases, it is unlikely that the observed association is due to early censoring. This observation is likely an artifact of utilizing insurance claims records: the case group was significantly younger than the controls, and were consequently more likely to experience changes in employment. We, however, note that age was not found to be a significant covariate in our analysis. This also implies that the presented risk ratios are conservative, due to the possibility of missed instances of depression among the cases. The observed trends were also found to be robust to the definition of depression used (protracted or not). Two findings suggest that that the specific processes and impacts unique to bariatric surgery may be responsible for our observations: i) non-bariatric abdominal surgery patients have reduced depression risk ratios, likely due to the therapeutic value of the surgery itself, and ii) laparoscopic bariatric surgeries also have an elevated hazard ratio against depression relative to esophageal/stomach laparoscopic surgery patients. Differential susceptibility towards this effect based on sex was also observed. Despite a higher population incidence of depression among female individuals, the post-surgical effect size was found to be higher among men. Among the subset of patients with pre-surgical BMI measurements, we observed only a very small effect size of BMI with respect to depression risk, implying that our observations are unlikely to be driven by higher BMI measurements in the bariatric surgery group. Previous studies reported more significant effect sizes (Bjerkeset *et al.*, 2008) or a U-shaped relationship (de Wit *et al.*, 2009) between BMI and depression. We hypothesize that the lack of observed effect in our cohort is due to our study design: all patients were eligible for bariatric surgery, and had significantly elevated BMI and baseline depression

risk as a result. Therefore, trends observed over a wider BMI range might not be applicable to our cohorts.

A survey of 81 bariatric surgery programs found that only half required formal psychiatric assessment prior to surgery (Martin-Fernandez, Heinberg and Ben-Porath, 2019). Our finding that post-surgical depression was less common in patient populations with these screenings reinforce the value of these evaluations even in populations without histories of depression. These screenings may be interpreted as indicators of programs with greater priority on mental health care or stricter patient selection mechanisms.

Based on our control of confounders and the temporal control we applied to cohort selection, our results lead us to hypothesize a potential causal relationship between the surgery and subsequent depression in a subset of patients, relating to the success and frequency of surgery. Postsurgical phenotypes related to infection and surgical complications were strongly associated with depression among patients undergoing bariatric surgery, and revision/removal surgeries had some of the highest hazard ratios for subsequent depression. Crucially, postoperative surgical complications and infections were only associated with depression in bariatric surgery patients, and not in other abdominal surgery patients. These associations could provide hints at the mechanism and time-scale of post-bariatric surgical depression, such as perturbations of the gut microbiome or disappointment at the outcome of the surgery, though further study would be needed to investigate any of these hypotheses. Previous reports of post-surgical depression often implicated long-term weight loss as a primary factor (McGuire *et al.*, 1999), while our phenotype association data provides evidence for a shorter-term mechanism as well. Similarly, we observe a higher risk of post-surgical depression among RYGB patients

relative to other surgeries, which could be attributed to the fact that patients undergoing it typically have more severe baseline conditions (English *et al.*, 2018; Griggs *et al.*, 2018)

We were able to achieve a significantly higher sample size than comparable clinical studies through our use of clinical insurance records, allowing us to focus on a specific subset of the population (those without histories of depression) and compute comparable placeholder surgery dates for comparisons with non-surgery populations, which emulated a prospective clinical trial and were able to controlled for potential confounders not addressed in previous studies. Furthermore, we were able to utilize matched diagnosis data for our patient population to uncover new phenotypic associations with post-surgical depression.

Our study contains several limitations. First, although we enforced similar inclusion criteria on our cases and controls, based on the inclusion criteria for consideration for bariatric surgery, our cases still had higher pre-surgery claims counts relative to the controls, raising the possibility that they had poorer baseline health. Second, although we removed individuals with pre-existing diagnoses of depression, other comorbid conditions present among the population who received bariatric surgery could contribute to the observations noted. The limited demographic and sociological information present in the insurance claims dataset restricted the pool of covariates that could be controlled or adjusted for in the analysis (Table 7.3-4). Features that we were unable to measure include ethnicity, marital status, and socioeconomic status, as well as information about the providers. Furthermore, the population of individuals with health insurance may not be representative of the population at large. Annotations of procedures and interventions required the presence of billed procedure codes. Discrepancies between procedures billed for and procedures actually carried out as well as differences between when a procedure occurs and is billed for exist. Finally, we were unable to completely eliminate the possibility that

the observed association results from a hypothetical association between the decision to undergo bariatric surgery and subsequent depression, as opposed to the surgery itself.

Conclusions

This study examined the frequency of depression diagnoses after bariatric surgery in individuals without a prior history of depression, relative to both non-surgical and non-bariatric surgery controls. We report an increased risk of depression following bariatric surgery that is amplified in men and reduced in patient cohorts with pre-surgical psychiatric evaluations. Our findings also show that this effect is comparatively larger in patients undergoing gastric bypass, and smaller among patients undergoing vertical sleeve surgeries. Furthermore, we find that this effect is most pronounced in the presence of post-surgical infections or complications, as well as in patients with pre-surgical histories of memory loss or chronic airway obstruction. For patients considering bariatric surgery and their physicians, this research provides a clearer estimate of post-surgical depression risk and associated exacerbating and mitigating factors.

Appendix: Experimental Design and Variable Selection

In accordance with the recommendations made in Chapter 6, a framework describing the total features influencing bariatric surgery and depression is presented in Figure 7.4, as well as a comparison of what features are available and utilized in the model in Table 7.3. Finally, a table justifying the treatment of matching variables is presented in Table 7.4.

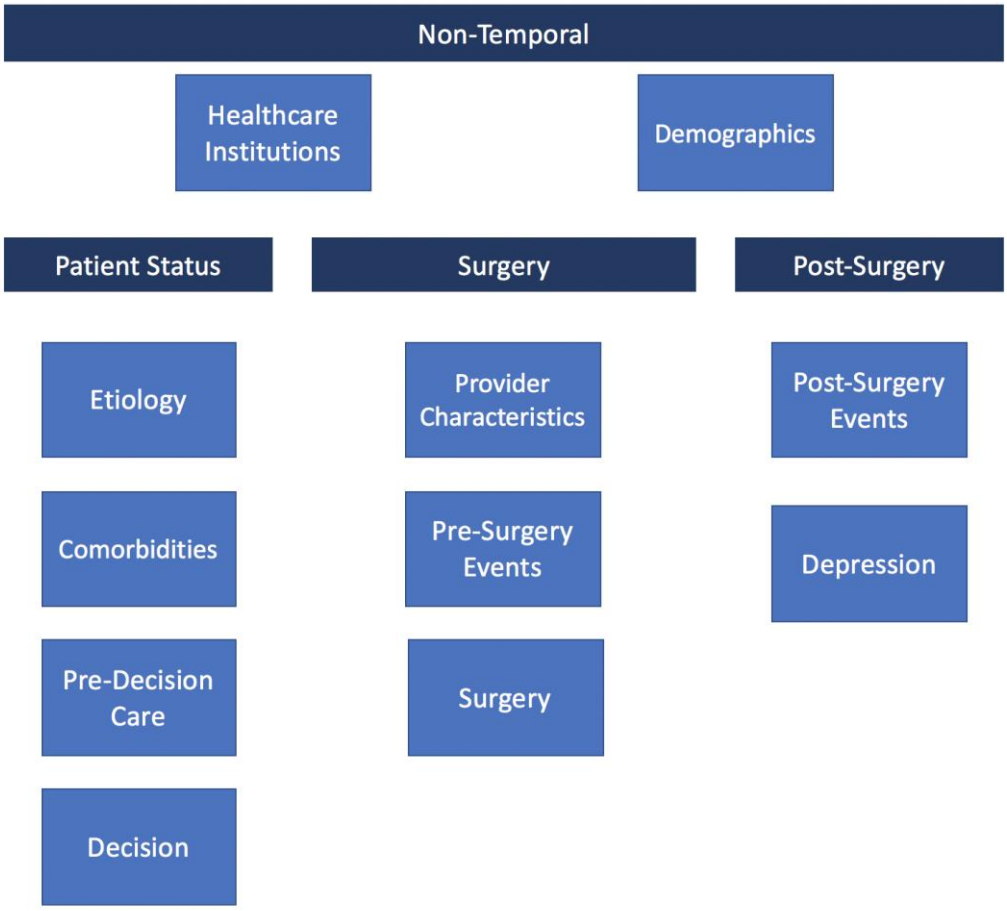


Figure 7.4: Framework of features influencing bariatric surgery status and postsurgical depression

Table 7.3: Comparison of features present in dataset to features present in framework (Table continues on next page)

	Features Present in Dataset	Features Missing from Dataset	Implications of Missingness
Health Care Institutions	Proxy for rural/urban (ZIP code)	N/A, all patients were in the US and had health insurance.	None
Demographics	Age, sex, proxy for socioeconomic status (ZIP code)	Race/ethnicity Marital Status	Propensity for both surgery (Cheung *et al.*, 2013; Johnson-Mann *et al.*, 2019) and depression (Simpson *et al.*, 2007) are influenced by race. Marital status is known to affect adherence to postoperative guidelines (Wheeler, 2008) as well as depression.
Etiology	Weak proxy for obesity etiology (diagnosis/procedural records)	Actual obesity etiology	True etiology is often not recorded or even known to patient/provider, potential source of unmeasured confounding.
Comorbidities	Strong proxy for comorbidities (diagnosis/procedural records) BMI	Unmeasured/acted upon comorbidities.	Can affect propensity for depression, but unlikely to influence propensity for surgery if provider is unaware.
Pre-Decision Care	Strong proxy for pre-decision care (diagnosis/procedural records)	Care at out-of-network providers	Provider is likely to be aware of out-of-network treatments, influencing propensity for surgery, potential source of unmeasured confounding.

Decision	None	All factors influencing a decision to undertake surgery	Likely the strongest confounding factor, as individuals do not make the decision to undergo surgery randomly. We attempt to control for this by comparing to non-bariatric surgeries, but this is imperfect.
Provider Characteristics	None	All information about providers	Propensity to recommend surgery may be different between small vs. large hospitals, for-profit vs. not-for-profit, etc.
Pre-Surgery Events	Strong proxy for pre-surgery care (diagnosis and procedural records)	Care at out-of-network providers, but unlikely, as surgery was covered through insurance. What pre-surgical care was available.	Lack of availability of pre-surgical care may indicate non-comparable surgeries for psych eval vs. no psych eval comparisons.
Surgery	Directly measured Weak proxies for success of surgery (complications that require procedures)	Success of surgery	Expectations regarding surgery results influence both propensity for surgery and propensity for depression.
Post-Surgery Events	Strong proxy for post-surgery care (diagnosis and procedural records)	Care at out-of-network providers, but unlikely, as surgery was covered through insurance. Patient weight loss/expectations.	Presence or absence of strong post-surgical care is an indicator of program quality, influencing propensity for surgery and depression.
Depression	Directly measured	Severity	Associations among subgroups of depression will be undetectable.

Table 7.4: Rationale for including/excluding features in matching

Factor	Rationale/Causal Narrative	Decision
Age/Sex	Age and sex are a strong influence on both the comorbidity profile of the patients and the risk of depression.	Included in match.
ZIP code	ZIP is the only proxy available for socioeconomic status and provider, both of which are strong influences on propensity for surgery and depression.	Included in match but confounding is still present.
BMI	BMI is one of the central criteria for receiving bariatric surgery.	Include in match so cases/controls have similar levels of pre-surgery obesity.
Pre-surgery comorbidities	Comorbidities that influence surgery through policy include diabetes, cardiovascular disease, coronary artery disease, cardiomyopathy, and sleep apnea (UnitedHealthcare, 2018), while those that influence surgery through patient choices may include any that interfere with day to day life. To select just one, diabetes is known to significantly increase the odds of comorbid depression (Anderson *et al.*, 2001).	Include in match, but controlling for every individual factor would massively increase the number of hypotheses evaluated, making statistical significance difficult. We use a combination of surgery eligibility + code count as a proxy, but this is explicitly a compromise.
Total Claims Count	Healthcare utilization can be captured by counting total claims over an individual's coverage period. However, the presence and success of a surgery contributes to this count.	Exclude from match to prevent anomalously sick controls from being utilized.
Post Baseline/Surgery Enrollment	Health insurance coverage in our dataset is highly correlated with employment. Success of surgery or presence of depression could influence changes in employment that affect coverage time.	Exclude to prevent controls with unnaturally extended coverage from being utilized.

Conclusions

We can distill the critical components that underlie the ability to make robust predictions from observational data into several points:

- Have a question worth answering: determining whether or not this is present for a given study requires extensive consultation with domain experts and reflection regarding the available data. Domain experts and clinicians provide valuable context about the intricacies of a domain and are the end consumers of the work that we produce. It is therefore critical to design studies that address questions they care about, and so that they can identify when and where our predictions are most valid. This can take the form of narrowing the focus of a study, both to eliminate flaws such as temporal bias as well as to make the identification of hypothetical mechanisms easier. Furthermore, certain questions are incompatible with certain datasets. Making this concession is not an admission of defeat, but an indicator that we have a strong understanding of our system.
- Benchmark against the real world: performance and utility are intertwined, but are often not treated as so. Optimizing for accuracy, AUC, or effect size is not helpful if the strong performance can only be realized retrospectively, or if the predictions do not represent an advance beyond what a human practitioner already knows. Prediction in the real world is inherently a hard task filled with uncertainty- particularly among questions that are worth answering, it is not always reasonable to expect AUC = 0.9. Real-world benchmarks, particularly "clinical judgment," may be difficult to approximate or evaluate in a study, but are the ultimate standards that models will be judged against.
- Work to make prediction obsolete: prediction from observation is fundamentally a response to uncertainty about the causal relationships of interest. A machine learning

model to predict chemical reaction products given starting materials would be of dubious utility since the systems are sufficiently well understood. Similarly, when observed reaction products are not what was expected, the fault typically lies with the execution of the reaction rather than with the prediction. The idea of a prediction becomes obsolete once comprehensive mechanistic understanding of the system is achieved. In order to maximize impact, the predictions that we make should be towards the specific end of eliminating the uncertainty that exists, through improving mechanistic understanding or advancing hypotheses.

Resolving the tension between big data and biomedical research will involve significant shifts in the ways that research is conducted and evaluated. Conducting research with a focus on utility, modesty, and mechanism will be critical in delivering on the promise that new datasets and methodologies bring.

Supplementary Materials

Supplementary Materials for Chapter 1

Temporal Bias in Hard Drive Failure Prediction

We examined instances of temporal bias in a non-medical application: hard drive failure prediction. Hard drives utilize Self-Monitoring, Analysis and Reporting Technology (SMART), an internal self-monitoring and reporting process that records various physical internal metrics. There have been numerous attempts to develop algorithms or techniques to utilize these internal metrics for early warning systems (Hughes *et al.*, 2002). Hughes, et al. (Murray, Joseph F, Hughes, Gordon F and Kreutz-Delgado, Kenneth, 2005) utilized a dataset containing the records of the 600 immediately prior to failure for a set of disk drives, as well as records from a number of healthy drives.

To establish the presence of temporal bias, we examined the heterogeneity of the disc decay trajectories from an information content point of view. We identified "early" and "late" sections from each failed drive, defined as the earliest and latest 20-hour sections from each failed drive respectively and constructed logistic regression classifiers for future disk status. Assuming a homogeneous trajectory, features learned from late sections should be similarly predictive when deployed over early sections; however, we observed a significant drop in test AUC when temporally selected sections were tested against non-matching sections (Figure S1.1A). This implies the presence of temporal bias within this dataset: the early sections were distinguishable from controls, and late features fail to generalize to earlier sections. These imply that the observed period does not uniformly sample the hard drive failure trajectory.

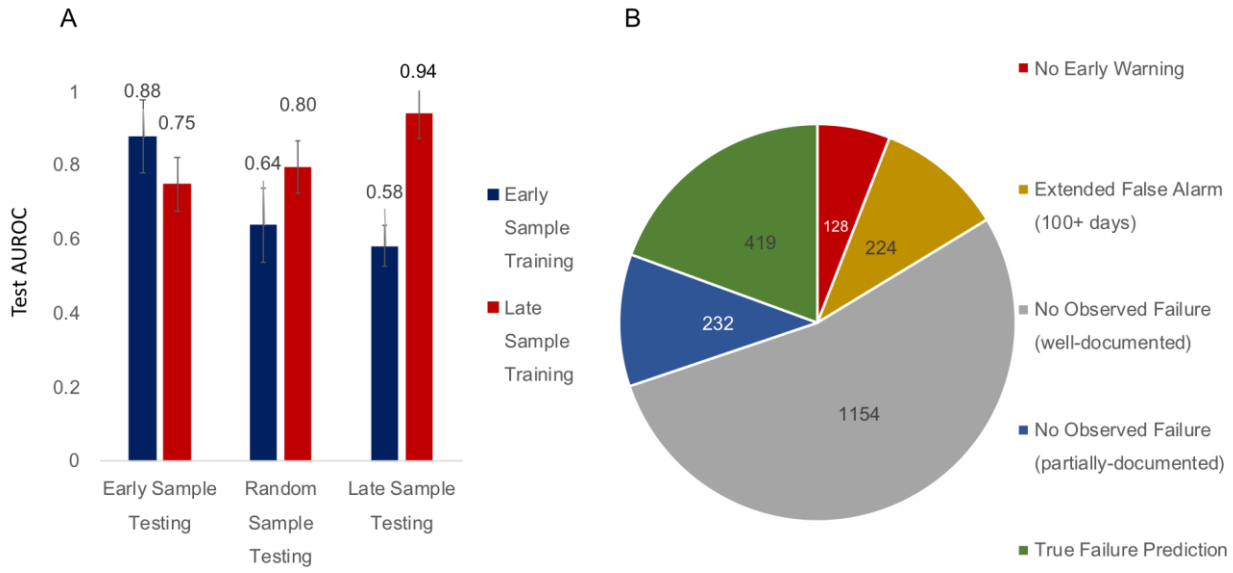


Figure S1.1: Temporal Bias in Disk Failure Prediction. A) The disk drive failure trajectory is heterogeneous over the final 600 hours. Features trained on early/late sections of the trajectory fail to generalize when tested over late/early sections respectively. **B)** Simulated prospective deployment of rules trained in a temporally biased manner show significant numbers of false alarms.

Botezatu, et al. (Botezatu *et al.*, 2016) constructed an algorithm built on the Backblaze monitoring dataset, containing continuous readings of SMART statistics and failure times for more than 50,000 hard disks. This algorithm examined the distribution of changepoints for particular parameters based on a "days-before-failure" metric when considering what observation window size to utilize. We examined the prospective performance of the Seagate-specific decision tree rules generated using their algorithm on the same Backblaze dataset. We compared predicted replacement dates to actual failure dates where available for disks from 2015-2018. We defined a "true failure" prediction as a predicted replacement date between 1 and 99 days prior to the true replacement date. Extended false alarms had replacement dates more than 100 days in advance of predictions. We grouped no observed failures into two groups based on when observations were censored: "well documented" (censor date > 1 year after predicted replacement date) and "partially-documented" (censor date < 1 year after predicted replacement date. For these two groups of disks, all disks were operating normally in all observations. If the algorithm predicted a replacement date on the same day the disk failed, the disk was classified as "no early warning." Figure S1.1B summarizes the distribution of result

Supplementary Materials for Chapter 2

Table S2.1A: Included and excluded codes for initial analysis

Initial Inclusion Criteria			
Disease	ICD9	ICD10	CPT
Parkinson's Disease	332, 332.0	G20	NA
Pre-PD exclusions			
AD/Cognitive Issues	331*	G30*	NA
Dementia	290*	F03.90	NA
Multiple Systems Atrophy/Progressive Supranuclear Palsy	333.0	G90.3, G23.1	NA
Schizophrenia	295*	F20*	NA
Lewy Body Dementia	331.82	G31.83	NA
Encephalitis	323*	G04*	NA
Wilson's Disease	275.1	E83.01	NA

Table S2.1B: Included and excluded codes for feature tracking and gait/tremor indexed analysis

Feature	ICD9	ICD10	CPT
Screening mammography, bilateral (2-view study of each breast), including computer-aided detection (cad) when performed	NA	NA	G0202
Tremor/Abnormal Movements	781.0, 781.7, 333.1, 333.90, 333.99	R25.0, R25.1, R25.2, R25.3, R25.8, R25.9, R29.0, G25.0, G25.1, G25.2	NA
Gait Disorders	781.2	R26.0, R26.1, R26.81, R26.89, R26.9	NA
Constipation	564.00, 564.01, 564.02, 564.09	K58.1, K59.00, K59.01, K59.02, K59.03, K59.04, K59.09	NA

Table S2.2: Tremor-only Cohort, PD associated Diagnoses

Description	OR	Adjusted P Value
Bipolar	2.03	8.69E-16
Difficulty in walking	1.43	0.000320
Senile cataract	1.16	0.00124
Lack of coordination	1.54	0.00568
Voice disturbance	1.47	0.00811
Memory loss	1.38	0.00860
Other non-epithelial cancer of skin	1.19	0.00861
Osteoporosis NOS	1.21	0.0193
Parasomnia	1.92	0.0348
Symptoms concerning nutrition, metabolism, and development	1.26	0.0380

Table S2.3: Gait-only Cohort, PD associated Diagnoses

Description	OR	Adjusted P Value
Bipolar	4.32	2.40E-49
Major depressive disorder	2.20	1.09E-35
Other persistent mental disorders due to conditions classified elsewhere	2.61	6.23E-21
Urinary incontinence	1.68	1.73E-16
Depression	1.55	2.34E-16
Other non-epithelial cancer of skin	1.40	2.4712
Memory loss	1.87	3.99E-12
Voice disturbance	2.21	4.88E-12
Malaise and fatigue	1.26	5.44E-12
Degeneration of intervertebral disc	1.28	3.79E-11
Frequency of urination and polyuria	1.41	3.35E-10
Actinic keratosis	1.26	6.65E-09
Senile cataract	1.24	8.76E-09
Dizziness and giddiness (Light-headedness and vertigo)	1.28	1.75E-08
Orthostatic hypotension	1.85	2.35E-08

Psychosis	1.99	3.80E-08
Symptoms concerning nutrition, metabolism, and development	1.55	4.83E-08
Generalized anxiety disorder	1.75	1.43E-07
Syncope and collapse	1.33	5.95E-07
Mood disorders	2.53	9.28E-07
Seborrheic dermatitis	1.68	3.93E-06
Functional disorders of bladder	1.57	3.93E-06
Retention of urine	1.36	5.25E-06
Urinary tract infection	1.22	6.61E-06
Chronic laryngitis	2.32	5.46e-04

Supplementary Materials for Chapter 4

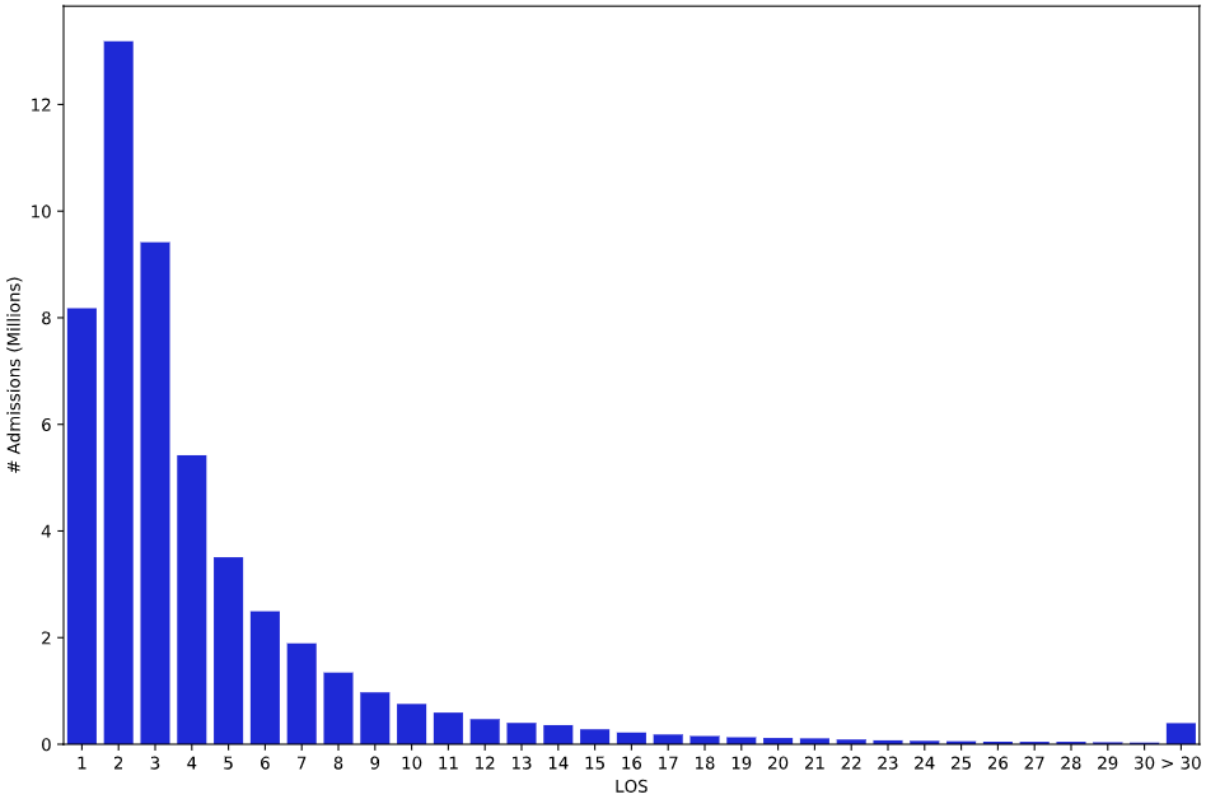


Figure S4.1. Histogram of length of stays for included admissions (in millions).

Table S4.1. Example first day charges for an MI patient with little data available.

Description	Department	Quantity
PF ER LEVEL V	PROFESSIONAL FEES	1
EKG ROUTINE TRACING ONLY	EKG	2
ER LEVEL V	EMERGENCY ROOM	1
THERAPEUTIC/DIAG INJ IV PUSH SINGLE INITI SUB/DRUG	IV THERAPY	1
*LORAZEPAM, ATIVAN INJ 2MG	PHARMACY	1
R&B ONCOLOGY PRIVATE	ROOM AND BOARD	1

Table S4.2. Example first day charges with unclear treatment actions. (Table continues on next page)

Description	Department	Quantity
PF ER LEVEL V	PROFESSIONAL FEES	1
EKG ROUTINE TRACING ONLY	EKG	2
PRESSURIZED OR NONPRESSURIZED INHALATION TX	RESPIRATORY THERAPY	1
ER LEVEL V	EMERGENCY ROOM	1
CPAP PER DAY	RESPIRATORY THERAPY	1
CPAP PER DAY	RESPIRATORY THERAPY	1
RT TIME FLAT RATE	RESPIRATORY THERAPY	1
*XR CHEST 1 VIEW PORTABLE	DIAGNOSTIC IMAGING	1
CULTURE BLOOD	LABORATORY	2
PARTIAL THROMBOPLASTIN TIME (PTT)	LABORATORY	1
PROTHROMBIN TIME (PT)	LABORATORY	1
COMPLETE CBC AUTO W/AUTO DIFF	LABORATORY	1
TROPONIN QN	LABORATORY	1
LACTATE/LACTIC ACID	LABORATORY	1
GLUCOSE BY DEVICE	LABORATORY	1
CREATINE KINASE (CPK) MB ONLY	LABORATORY	1
CREATINE KINASE (CPK)	LABORATORY	1
COMPREHENSIVE METABOLIC PANEL 80053	LABORATORY	1
NEW THERA PROPHY/DIAG INJ EA ADD SEQ PUSH SUB/DRUG	IV THERAPY	2
IV INFUSION SUBSTANCES/DRUG CONCURRENT 96368	IV THERAPY	1
IV INFUSION SUBST/DRUG EA ADD SEQ UP TO 1 HR 96367	IV THERAPY	1
IV INFUSION SUBSTANCES/DRUGS EACH ADDL 1	IV THERAPY	3

HR 96366		
IV INFUSION SUBSTANCES/DRUGS UP TO 1 HR 96365	IV THERAPY	1
0.9% NACL VL 10ML	PHARMACY	1
0.9% NACL 250ML	PHARMACY	1
ASPIRIN TAB CHW 81MG (EA)	PHARMACY	4
ALBUTEROL, NONCOMP INH SOL 1MG	PHARMACY	5
INSULIN ASPART PROT/ASPART, NOVOLOG 70/30 PER DOSE	PHARMACY	0.05
IPRATROPIUM/ALBUTEROL, DUONEB INH SOL 3ML	PHARMACY	1
LIDOCAINE, XYLOCAINE VL 2% 10ML	PHARMACY	1
HEPARIN NA VL 1,000U/ML 1ML	PHARMACY	1
HEPARIN NA VL 1,000U/ML 1ML	PHARMACY	1
*FUROSEMIDE, LASIX VL 40MG 4ML	PHARMACY	3
FENTANYL, SUBLIMAZE AMP 0.05MG/ML 2ML	PHARMACY	1
CEFTRIAXONE, ROCEPHIN VL 250MG	PHARMACY	1
AZITHROMYCIN, ZITHROMAX VL 500MG	PHARMACY	1
NON REVENUE ITEM	ADMINISTRATIVE FEES	5
R&B ICU	ROOM AND BOARD	1

Table S4.3. Admitting Physician Specialty (where available). (Table continues on next pages)

Admitting Physician Specialty	# Admissions
INTERNAL MEDICINE (IM)	8,683,840
HOSPITALIST (HOS)	7,668,355
OBSTETRICS/GYNECOLOGY (OBG)	4,295,512
UNKNOWN	3,254,355
PEDIATRICS (PD)	3,209,914
FAMILY PRACTICE (FP)	2,286,362
ORTHOPEDIC SURGERY (ORS)	1,824,100
PSYCHIATRY (P)	1,765,802
GENERAL SURGERY (GS)	1,479,856
NEONATAL - PERINATAL MEDICINE (NPM)	892,949
CARDIOVASCULAR DISEASES (CD)	876,613
UNSPECIFIED (US)	672,255
NEUROLOGICAL SURGERY (NS)	433,030
PULMONARY DISEASES (PUD)	426,366
PHYSICAL MEDICINE AND REHAB (PM)	348,622
EMERGENCY MEDICINE (EM)	331,833
OTHER SPECIALTY (OS)	297,251
CRITICAL CARE MEDICINE (CCM)	255,352
NEPHROLOGY (NEP)	235,647
UROLOGY (U)	226,721
THORACIC SURGERY (TS)	225,508
CARDIOVASCULAR SURGERY (CDS)	219,539
HEMATOLOGY/ONCOLOGY (HO)	199,818
VASCULAR SURGERY (VS)	198,145
TRAUMA SURGERY (TRS)	171,277
NEUROLOGY (N)	170,404
OBSTETRICS (OBS)	141,561
COLON/RECTAL SURGERY (CRS)	141,120
PULMONARY CRITICAL CARE MEDICINE (PCC)	100,151

PEDIATRIC CRITICAL CARE MEDICINE (CCP)	92,561
GYNECOLOGICAL ONCOLOGY (GO)	84,902
CERTIFIED NURSE MIDWIFE (CNM)	80,432
MEDICAL ONCOLOGY (ON)	79,390
INTERVENTIONAL CARDIOLOGY	73,798
GASTROENTEROLOGY (GE)	73,657
MATERNAL AND FETAL MEDICINE (MFM)	73,394
CHILD AND ADOLESCENT PSYCHIATRY (CHP)	72,002
GERIATRICS - INTERNAL MEDICINE (IMG)	63,311
GENERAL PRACTICE (GP)	63,155
SURGICAL CRITICAL CARE (CCS)	62,820
INTENSIVIST (INT)	62,410
INFECTIOUS DISEASES (ID)	61,761
GYNECOLOGY (GYN)	60,553
PEDIATRIC HEMATOLOGY/ONCOLOGY (PHO)	57,319
PLASTIC SURGERY (PS)	52,931
PEDIATRIC SURGERY (PDS)	47,671
OTOLARYNGOLOGY (OTO)	46,239
NURSE PRACTITIONER (ARNP)	45,990
ORTHOPEDIC SURGERY OF THE SPINE (OSS)	40,801
SURGICAL ONCOLOGY (SO)	35,813
CARDIAC ELECTROPHYSIOLOGY (ICE)	34,263
ANESTHESIOLOGY (AN)	33,296
PODIATRY (POD)	27,802
ENDOCRINOLOGY AND METABOLISM (END)	27,738
GERIATRIC MEDICINE - FAMILY PRAC. (FPG)	26,737
HEMATOLOGY (HEM)	24,599
TRANSPLANT SURGERY (TTS)	22,886
SPORTS MEDICINE - ORTHOPEDICS (OSM)	17,005
PHYSICIAN ASSISTANT (DRA)	15,025
PEDIATRIC GASTROENTEROLOGY (PG)	14,909

HEMATOLOGY (HMP)	14,270
RHEUMATOLOGY (RHU)	12,860
DENTAL/ORAL SURGERY (DOR)	12,790
PEDIATRIC PULMONOLOGY (PDP)	12,645
ADDICTION MEDICINE (ADM)	11,677
ADOLESCENT MEDICINE (ADL)	11,538
PEDIATRIC NEPHROLOGY (PN)	11,536
PEDIATRIC INFECTIOUS DISEASES (PDI)	11,382
PEDIATRIC EMERGENCY MEDICINE (PEM)	11,054
PEDIATRIC CARDIOLOGY (PDC)	10,281
HAND SURGERY (HS)	10,273
PSYCHOANALYSIS (PYA)	9,921
PAIN MANAGEMENT (APM)	9,236
PEDIATRIC NEUROLOGY (CHN)	8,820
VASC. & INTERVENTIONAL RADIOLOGY (VIR)	8,050
HOSPICE & PALLIATIVE CARE	7,694
RADIOLOGY - DIAGNOSTIC (DR)	7,337
CERTIFIED REG. NURSE ANESTHETIST (CRNA)	7,067
RADIOLOGY (R)	6,868
ABDOMINAL SURGERY (AS)	6,725
RADIATION ONCOLOGY (RO)	6,093
PEDIATRIC ORTHOPEDICS (OP)	5,976
OPHTHALMOLOGY (OPH)	5,900
PEDIATRIC SURGERY - NEUROLOGICAL (NSP)	5,247
ALLERGY AND IMMUNOLOGY (AI)	4,745
MEDICAL GENETICS (MG)	4,503
PEDIATRIC ALLERGY (PDA)	4,327
NUCLEAR MEDICINE (NM)	4116
REPRODUCTIVE ENDOCRINOLOGY (REN)	4,003
DERMATOLOGY (D)	3,986
SPORTS MEDICINE (FSM)	3,540

PEDIATRIC ENDOCRINOLOGY (PDE)	3,336
PEDIATRIC UROLOGY (UP)	2,557
SLEEP MEDICINE	2,533
OCCUPATIONAL MEDICINE (OM)	2,460
MAXILLOFACIAL SURGERY	2266
ALLERGY (A)	2,115
ANATOMIC/CLINICAL PATHOLOGY (PTH)	2,000
SPORTS & INTERNAL MEDICINE (ISM)	1,413
GENERAL PREVENTATIVE MEDICINE (GPM)	1,254
NEUROPATHOLOGY (NP)	1,237
OSTEOPATHIC MANIPULATIVE MEDICINE (OMM)	1,210
CLINICAL GENETICS (CG)	1,149
LEGAL MEDICINE (LM)	1,072
PEDIATRIC OTOLARYNGOLOGY (PDO)	1,046
HEAD AND NECK SURGERY (HNS)	1,017
CLINICAL PHARMACOLOGY (PA)	905
CHIROPRACTICE (CRP)	780
FACIAL PLASTIC SURGERY (FPS)	744
NUCLEAR RADIOLOGY (NR)	579
DIABETES (DIA)	565
NEURORADIOLOGY (RNR)	522
PUBLIC HLTH & GEN?L PREV. MEDICINE (PHP)	508
DERMATOPATHOLOGY (DMP)	460
NUTRITION (NTR)	458
CLINICAL PATHOLOGY (CLP)	388
PSYCHOLOGIST, CLINICAL	256
CLINICAL NEUROPHYSIOLOGY (CN)	230
PEDIATRIC OPHTHALMOLOGY (PO)	220
PEDIATRIC RHEUMATOLOGY (PPR)	200
CERTIFIED CLINICAL NURSE SPECIALIST	174

ANATOMIC PATHOLOGY (ATP)	146
IMMUNOLOGY (IG)	146
CHEMICAL PATHOLOGY (PCH)	125
PHYSICAL THERAPY	18
OCCUPATIONAL THERAPY	14
PEDIATRIC RADIOLOGY (PDR)	14
OPTOMETRY	10

Table S4.4. Admission Type (where available).

Admission Type	# Admissions
EMERGENCY	22,237,533
ELECTIVE	8,868,840
URGENT	6,657,028
NEWBORN	4,352,538
TRAUMA CENTER	357,779

Table S4.5: GRU model hyperparameters

Hyperparameter	Value
Sequence Length	100
Embedding Shape	8 dimensions
GRU hidden size	128/64/32
GRU hidden dropout	0.1/0.1/0.1
Dense hidden size	32/16
Dense Hidden dropout	0.1
Early Stopping Patience	100

Supplementary Materials For Chapter 7

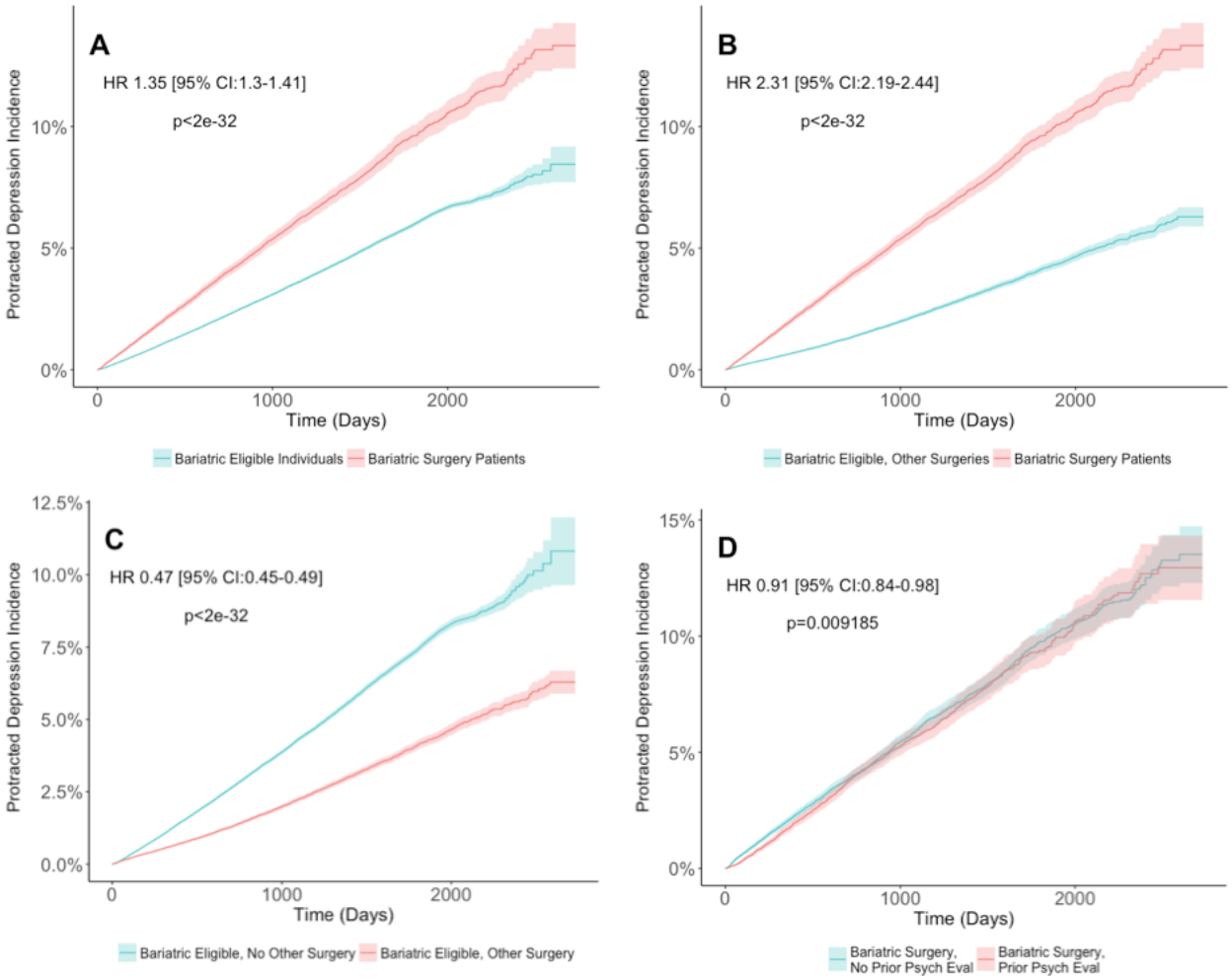


Figure S7.1: Incidence of Post-Surgical Protracted Depression
Protracted Depression was defined as 3 or more diagnoses of depression within a 6-month period. A) Time-to-protracted depression curve for bariatric surgery patients compared to bariatric eligible individuals. B) Time-to-protracted depression curve for bariatric surgery patients compared to bariatric eligible individuals with other abdominal surgeries. C) Time-to-protracted depression curve for bariatric eligible individuals with other abdominal surgeries compared to bariatric eligible individuals without other abdominal surgeries. D) Time-to-protracted depression curve for bariatric surgery patients with pre-surgical psychiatric evaluations compared to bariatric surgery patients without pre-surgical psychiatric evaluations- the p-value for this test was not found to be significant. All hazard ratios are adjusted for sex, age, and 6-month claim count.

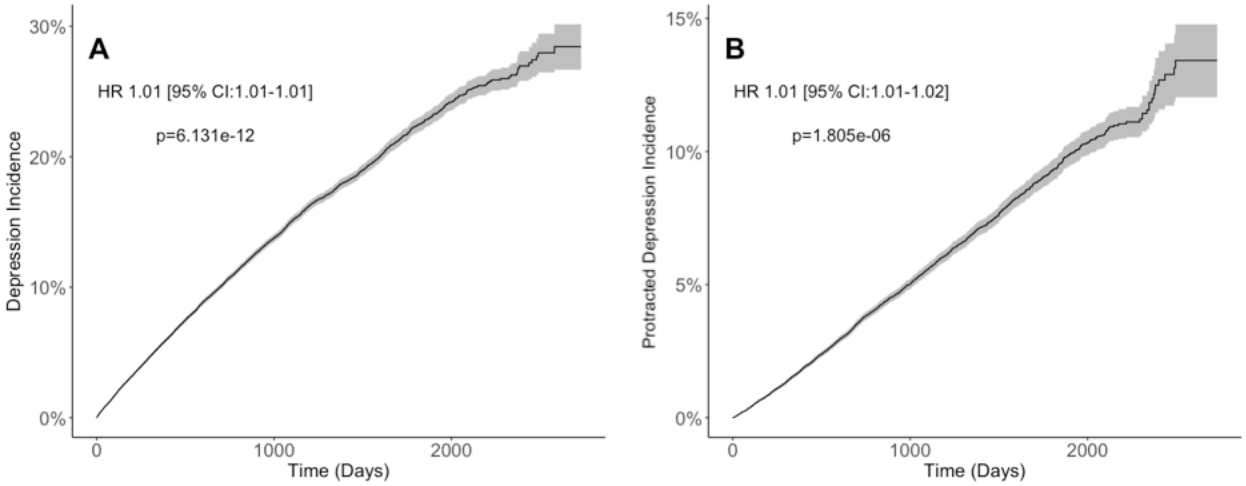


Figure S7.2: Influence of BMI on Depression Incidence
Individuals with BMI annotations were examined to identify a relationship between BMI and depression risk in bariatric eligible individuals and bariatric surgery patients. A) Time-to-depression curve for bariatric eligible individuals and bariatric surgery patients. B) Time-to-protracted depression curve for bariatric eligible individuals and bariatric surgery patients. All hazard ratios are for BMI with respect to depression/protracted depression.

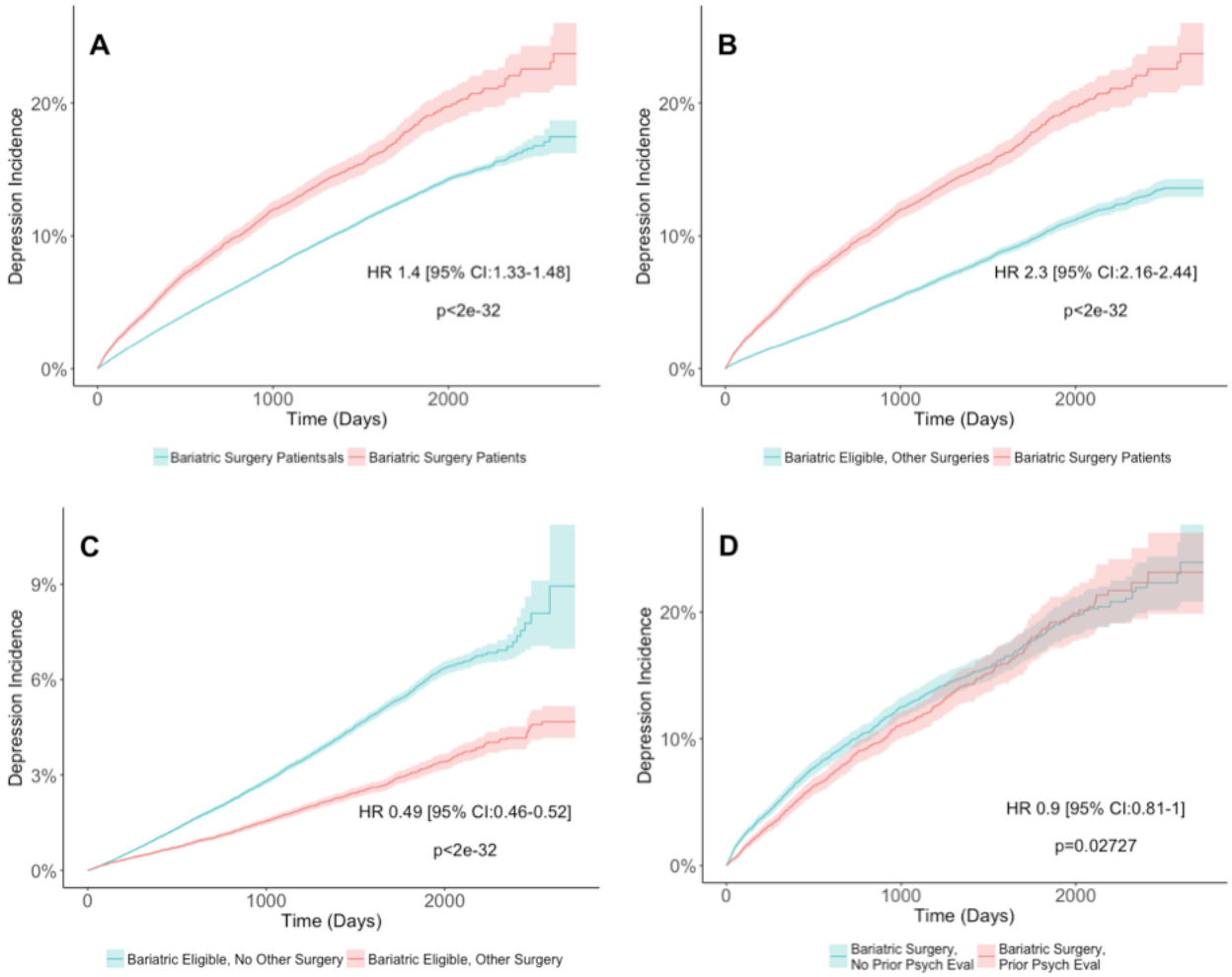


Figure S7.3: Incidence of Post-Surgical Depression in Men
A) Time-to-depression curve for male bariatric surgery patients compared to male bariatric eligible individuals. B) Time-to-depression curve for male bariatric surgery patients compared to male bariatric eligible individuals with other abdominal surgeries. C) Time-to-depression curve for male bariatric eligible individuals with other abdominal surgeries compared to male bariatric eligible individuals without other abdominal surgeries. D) Time-to-depression curve for male bariatric surgery patients with pre-surgical psychiatric evaluations compared to male bariatric surgery patients without pre-surgical psychiatric evaluations- the p-value for this test was not found to be significant. All hazard ratios are adjusted for age, and 6-month claim count.

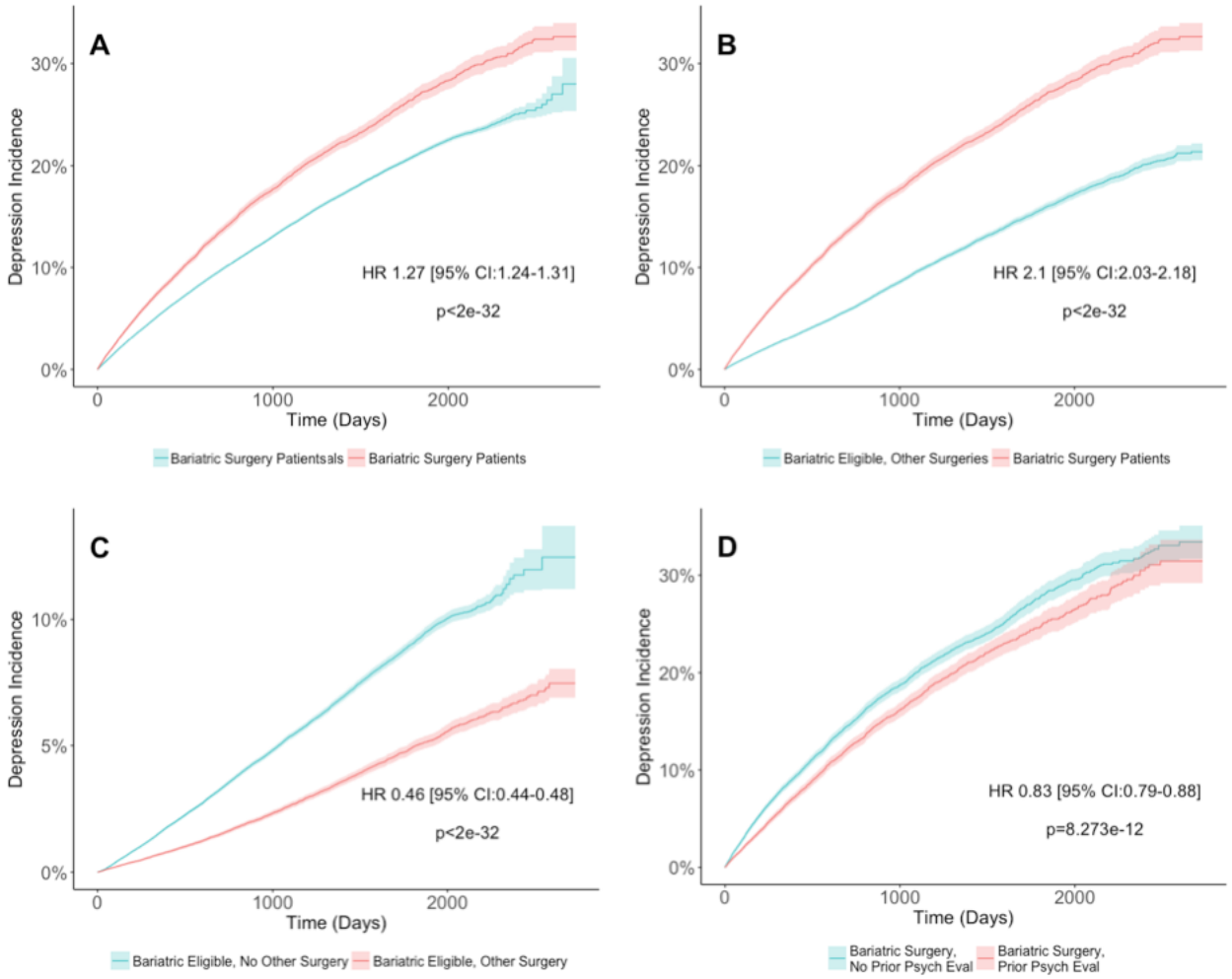


Figure S7.4: Incidence of Post-Surgical Depression in Women
A) Time-to-depression curve for female bariatric surgery patients compared to female bariatric eligible individuals. B) Time-to-depression curve for female bariatric surgery patients compared to female bariatric eligible individuals with other abdominal surgeries. C) Time-to-depression curve for female bariatric eligible individuals with other abdominal surgeries compared to female bariatric eligible individuals without other abdominal surgeries. D) Time-to-depression curve for female bariatric surgery patients with pre-surgical psychiatric evaluations compared to female bariatric surgery patients without pre-surgical psychiatric evaluations. All hazard ratios are adjusted for age, and 6-month claim count.

Table S7. 1: Procedure-stratified Cohort Characteristics (Table continues on next page)

	Number (%)		
	Revisions/Removals	Gastric Bypass	Vertical Band
Total	12319	18877	506
Men	3118 (25.31)	5240 (27.76)	137 (27.08)
Age, years, mean (SD)	57.69 (12.63)	45.52 (12.33)	45.52 (14.43)
Post-Surgical Depression Diagnosis (>=1)	1714 (13.91)	2734 (14.48)	77 (15.22)
Protracted Post-Surgical Depression Occurrences (>= 3 Diagnoses/6 Months)	566 (4.59)	974 (5.16)	26 (5.14)
BMI, mean (SD)	42.21 (7.08)	45.61 (6.84)	43.59 (6.06)
6 Month Code Count, mean (SD)	135.68 (160.28)	120.15 (102.46)	126.42 (136.15)
6 Month Diagnosis Count, mean (SD)	61.16 (70)	55.98 (46.02)	56.56 (58.99)
6 Month Procedure Count, mean (SD)	74.52 (93.35)	64.17 (59.46)	69.85 (80.14)
Follow-up Time, days, mean (SD)	706.94 (638.97)	858.76 (724.02)	912.75 (736.48)

	Number (%)		
	Adjustible Band	Duodenal Switch	Vertical Sleeve
Total	15799	6686	19852
Men	4022 (25.46)	2579 (38.57)	5673 (28.58)
Age, years, mean (SD)	43.67 (11.52)	50.94 (16.53)	44.51 (11.93)
Post-Surgical Depression Diagnosis (>=1)	2368 (14.99)	860 (12.86)	1549 (7.8)
Protracted Post-Surgical Depression Occurrences (>= 3 Diagnoses/6 Months)	739 (4.68)	311 (4.65)	520 (2.62)
BMI, mean (SD)	43.22 (5.82)	45.3 (7.36)	45.44 (7.58)
6 Month Code Count, mean (SD)	95.77 (67.1)	142.48 (189.48)	118.95 (91.84)
6 Month Diagnosis Count, mean (SD)	45.36 (30.57)	63.42 (81.23)	57.21 (42.49)
6 Month Procedure Count, mean (SD)	50.4 (39.21)	79.06 (111.84)	61.74 (52.03)
Follow-up Time, days, mean (SD)	1038.12 (772.13)	831.17 (670.27)	569.85 (503.56)

Table S7. 2: Laparoscopic Surgery Cohort Characteristics

	Number (%)	
	Laparoscopic Bariatric Surgeries	Bariatric Eligible, non-Bariatric Laparoscopic Surgeries of the Stomach and Esophagus
Total	58536	2681
Men	16745 (28.61)	712 (26.56)
Age, years, mean (SD)	45.74 (13.05)	52.15 (13.36)
Post-Surgical Depression Diagnosis (>=1)	6887 (11.77)	282 (10.52)
Protracted Post-Surgical Depression Occurrences (>= 3 Diagnoses/6 Months)	2347 (4.01)	94 (3.51)
BMI, mean (SD)	44.81 (7.04)	42.89 (7.91)
6 Month Code Count, mean (SD)	115.27 (114.59)	128.91 (121.88)
6 Month Diagnosis Count, mean (SD)	54.12 (50.65)	61.57 (56.4)
6 Month Procedure Count, mean (SD)	61.15 (66.72)	67.34 (68.39)
Follow-up Time, days, mean (SD)	766.47 (670.93)	921.15 (656.78)

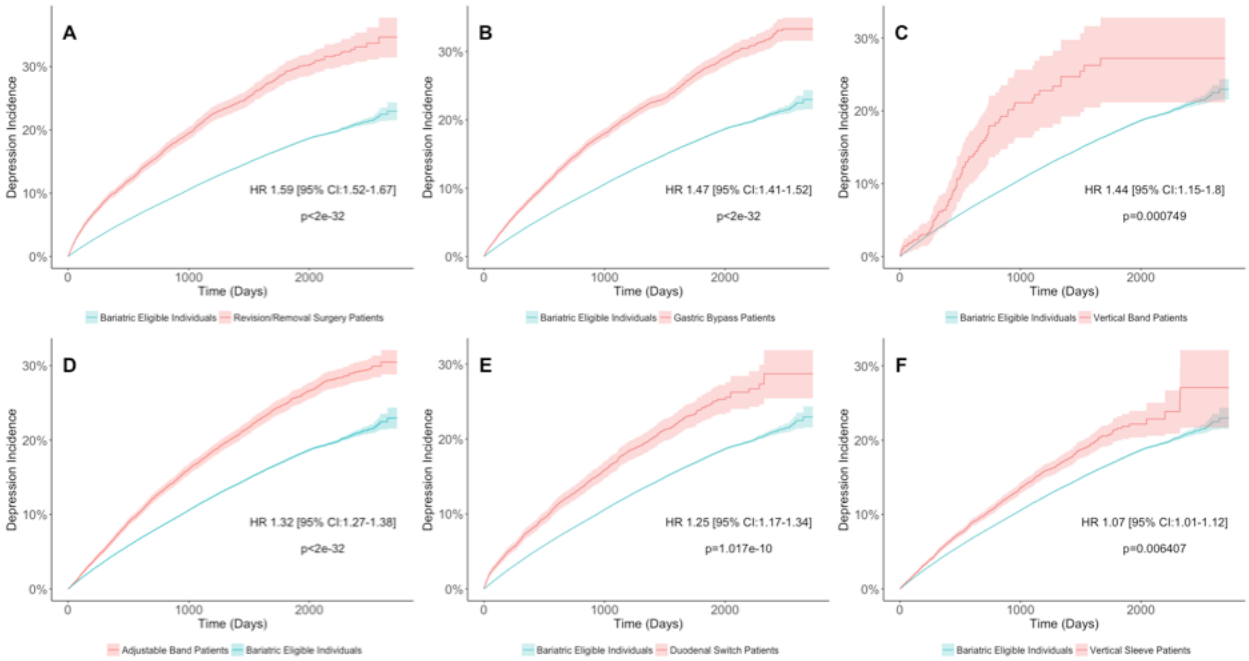


Figure S7.5: Incidence of Post-Surgical Depression Relative to Non-Surgical Controls, Stratified by Procedure

A) Time-to-depression curve for Revision/Removal surgery patients compared to bariatric eligible individuals. B) Time-to-depression curve for Gastric Bypass patients compared to bariatric eligible individuals. C) Time-to-depression curve for Vertical Band patients compared to bariatric eligible individuals. D) Time-to-depression curve for Adjustable Band patients compared to bariatric eligible individuals. E) Time-to-depression curve for Duodenal Switch patients compared to bariatric eligible individuals. F) Time-to-depression curve for Vertical Sleeve patients compared to bariatric eligible individuals. All hazard ratios are adjusted for sex, age, and 6-month claim count.

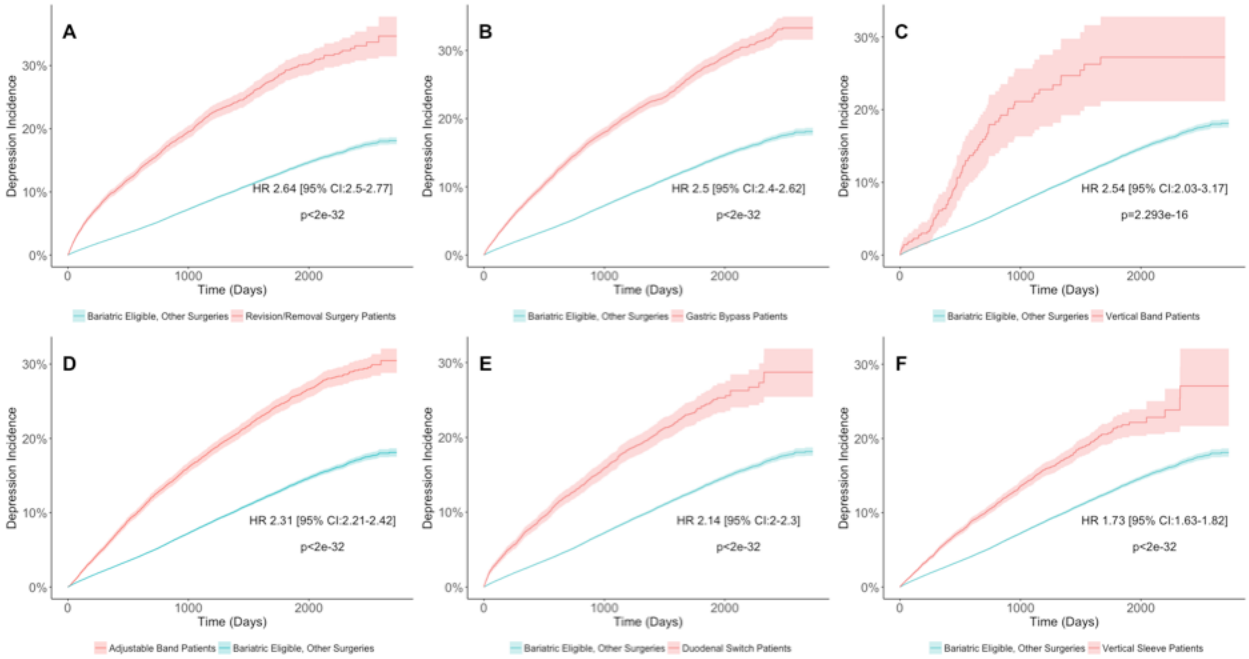


Figure S7.6: Incidence of Post-Surgical Depression Relative to Abdominal Surgical Controls, Stratified by Procedure

A) Time-to-depression curve for Revision/Removal surgery patients compared to bariatric eligible individuals with other abdominal surgeries. B) Time-to-depression curve for Gastric Bypass patients compared to bariatric eligible individuals with other abdominal surgeries. C) Time-to-depression curve for Vertical Band patients compared to bariatric eligible individuals with other abdominal surgeries. D) Time-to-depression curve for Adjustable Band patients compared to bariatric eligible individuals with other abdominal surgeries. E) Time-to-depression curve for Duodenal Switch patients compared to bariatric eligible individuals with other abdominal surgeries. F) Time-to-depression curve for Vertical Sleeve patients compared to bariatric eligible individuals with other abdominal surgeries. All hazard ratios are adjusted for sex, age, and 6-month claim count.

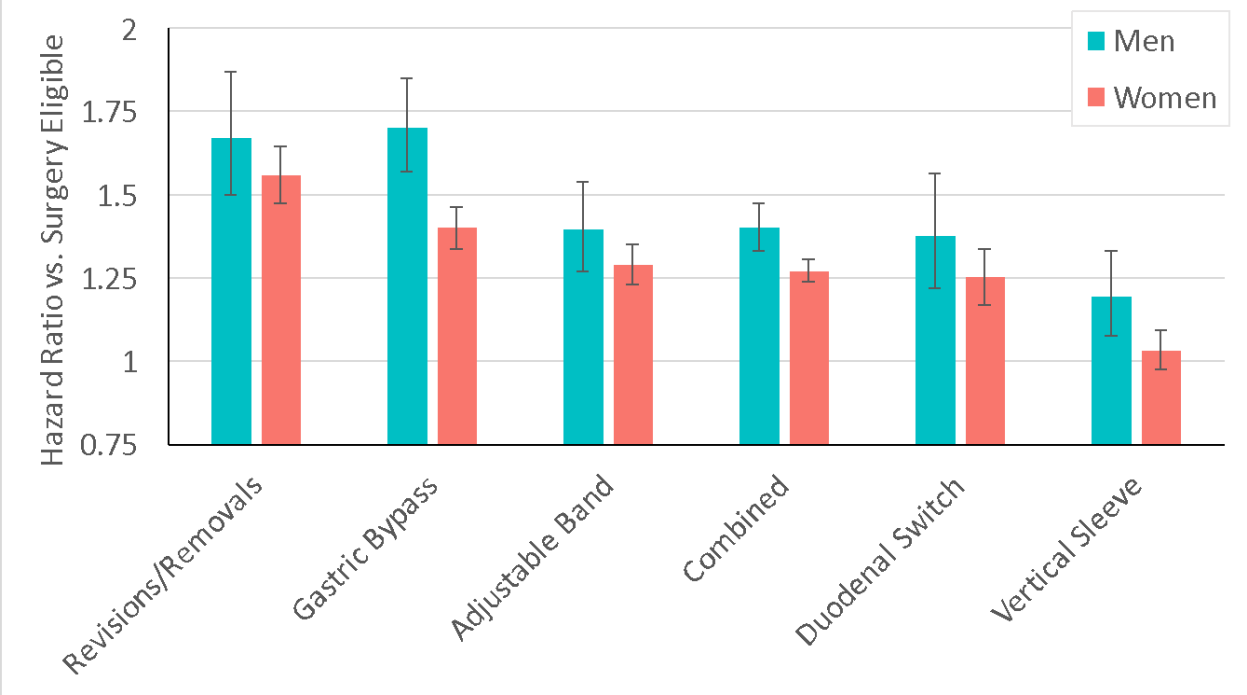


Figure S7.7: Incidence of Post-Surgical Depression Relative to Non-Surgical Controls, Stratified by Procedure and Sex
Error bars represent 95% confidence intervals of hazard ratios. Data for Vertical Band patients is not shown in this plot.

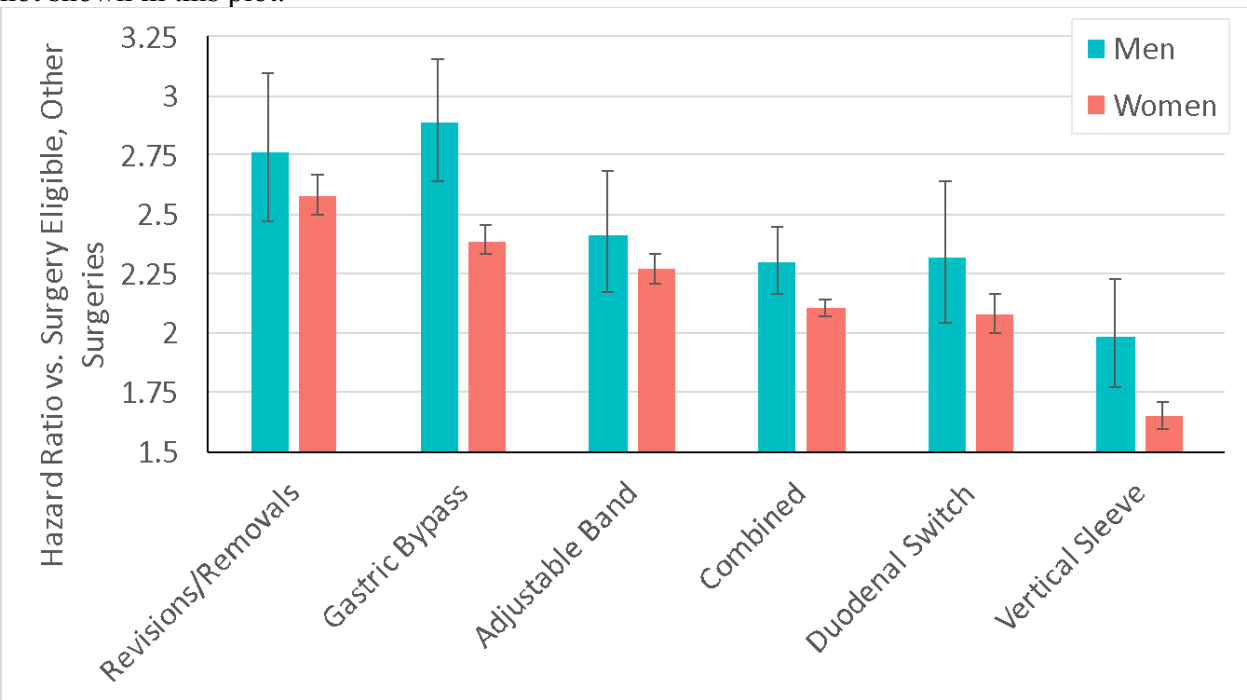


Figure S7.8: Incidence of Post-Surgical Depression Relative to Abdominal Surgical Controls, Stratified by Procedure and Sex
Error bars represent 95% confidence intervals of hazard ratios. Data for Vertical Band patients is not shown in this plot.

Table S7.3: Depression Diagnosis-Phenotype Codes (Table continues on next page)

ICD9	ICD9 String	PheCode	Phenotype
296.2	Major depressive disorder, single episode	296.22	Major depressive disorder
296.2	Major depressive disorder, single episode, unspecified degree	296.22	Major depressive disorder
296.21	Major depressive disorder, single episode, mild degree	296.2	Depression
296.22	Major depressive disorder, single episode, moderate degree	296.22	Major depressive disorder
296.23	Major depressive disorder, single episode, severe degree, without mention of psychotic behavior	296.22	Major depressive disorder
296.24	Major depressive disorder, single episode, severe degree, specified as with psychotic behavior	296.22	Major depressive disorder
296.25	Major depressive disorder, single episode, in partial or unspecified remission	296.22	Major depressive disorder
296.26	Major depressive disorder, single episode in full remission	296.22	Major depressive disorder
296.3	Major depressive disorder, recurrent episode	296.22	Major depressive disorder
296.3	Major depressive disorder, recurrent episode, unspecified degree	296.22	Major depressive disorder
296.31	Major depressive disorder, recurrent episode, mild degree	296.2	Depression
296.32	Major depressive disorder, recurrent episode, moderate degree	296.22	Major depressive disorder

296.33	Major depressive disorder, recurrent episode, severe degree, without mention of psychotic behavior	296.22	Major depressive disorder
296.34	Major depressive disorder, recurrent episode, severe degree, specified as with psychotic behavior	296.22	Major depressive disorder
296.35	Major depressive disorder, recurrent episode, in partial or unspecified remission	296.22	Major depressive disorder
296.36	Major depressive disorder, recurrent episode, in full remission	296.22	Major depressive disorder
311	Depressive disorder NEC	296.2	Depression

Bibliography

Abbott, R. D. *et al.* (2001) 'Frequency of bowel movements and the future risk of Parkinson's disease', *Neurology*, 57(3), pp. 456–462.

Abbott, R. D. *et al.* (2005) 'Excessive daytime sleepiness and subsequent development of Parkinson disease', *Neurology*, 65(9), pp. 1442–1446.

Agniel, D., Kohane, I. S. and Weber, G. M. (2018) 'Biases in electronic health record data due to processes within the healthcare system: retrospective observational study', *BMJ* , 361, p. k1479.

Alonso, A. *et al.* (2007) 'Gout and risk of Parkinson disease: a prospective study', *Neurology*, 69(17), pp. 1696–1700.

Amaducci, L. A. *et al.* (1987) 'Risk factors for clinically diagnosed patients of Alzheimer??s disease: A case-control study of an Italian population', *Alzheimer Disease & Associated Disorders*, p. 103. doi: 10.1097/00002093-198701020-00006.

Anderson, R. J. *et al.* (2001) 'The Prevalence of Comorbid Depression in Adults With Diabetes', *Diabetes care*. American Diabetes Association, 24(6), pp. 1069–1078.

ANSES (2020) *COVID-19 cannot be transmitted by either farm animals or domestic animals*, *French Agency for Food, Environmental and Occupational Health & Safety*. Available at: https://www.anses.fr/en/content/covid-19-cannot-be-transmitted-either-farm-animals-or-domestic-animals-0 (Accessed: 2020).

Bababekov, Y. J. *et al.* (2018) 'A Proposal to Mitigate the Consequences of Type 2 Error in Surgical Science', *Annals of surgery*, 267(4), pp. 621–622.

Baron, J. H. (2009) 'Sailors' scurvy before and after James Lind - a reassessment', *Nutrition Reviews*, pp. 315–332. doi: 10.1111/j.1753-4887.2009.00205.x.

Beam, A. L. *et al.* (2020) 'Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. World Scientific, 25, pp. 295–306.

Beam, A. L. and Kohane, I. S. (2016) 'Translating Artificial Intelligence Into Clinical Care', *JAMA: the journal of the American Medical Association*. jamanetwork.com, pp. 2368–2369.

Beaulieu-Jones, B. K. *et al.* (2018) 'Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis', *JMIR medical informatics*. medinform.jmir.org, 6(1), p. e11.

Beaulieu-Jones, B. K., Kohane, I. S. and Beam, A. L. (2019) 'Learning Contextual Hierarchical Structure of Medical Concepts with Poincairé Embeddings to Clarify Phenotypes', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 24, pp. 8–17.

Beaulieu-Jones, B. K., Kohane, I. S. and Beam, A. L. (2019) 'Learning Contextual Hierarchical

Structure of Medical Concepts with Poincaire Embeddings to Clarify Phenotypes', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 24, pp. 8–17.

Beaulieu-Jones, B. K., Orzechowski, P. and Moore, J. H. (2018) 'Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23, pp. 123–132.

Belle, A. *et al.* (2015) 'Big Data Analytics in Healthcare', *BioMed research international*, 2015, p. 370194.

Bender, D. and Sartipi, K. (2013) 'HL7 FHIR: An Agile and RESTful approach to healthcare information exchange', *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. doi: 10.1109/cbms.2013.6627810.

Berg, D. *et al.* (2015) 'MDS research criteria for prodromal Parkinson's disease', *Movement disorders: official journal of the Movement Disorder Society*, 30(12), pp. 1600–1611.

Bertuzzi, M. *et al.* (2002) 'Olive oil consumption and risk of non-fatal myocardial infarction in Italy', *International journal of epidemiology*, pp. 1274–7; author reply 1276–7.

van der Bij, S. *et al.* (2017) 'Improving the quality of EHR recording in primary care: a data quality feedback tool', *Journal of the American Medical Informatics Association: JAMIA*. academic.oup.com, 24(1), pp. 81–87.

Bjerkeset, O. *et al.* (2008) 'Association of adult body mass index and height with anxiety, depression, and suicide in the general population: the HUNT study', *American journal of epidemiology*, 167(2), pp. 193–202.

Botezatu, M. M. *et al.* (2016) 'Predicting Disk Replacement towards Reliable Data Centers', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. doi: 10.1145/2939672.2939699.

Botsis, T. *et al.* (2010) 'Secondary Use of EHR: Data Quality Issues and Informatics Opportunities', *Summit on translational bioinformatics*. ncbi.nlm.nih.gov, 2010, pp. 1–5.

Breen, D. P. *et al.* (2013) 'Determinants of delayed diagnosis in Parkinson's disease', *Journal of Neurology*, pp. 1978–1981. doi: 10.1007/s00415-013-6905-3.

Brolin, R. E. (1996) 'Gastrointestinal surgery for severe obesity', *Nutrition*, pp. 403–404. doi: 10.1016/s0899-9007(96)00154-2.

Burgmer, R. *et al.* (2014) 'Psychological outcome 4 years after restrictive bariatric surgery', *Obesity surgery*, 24(10), pp. 1670–1678.

Burns, P. B., Rohrich, R. J. and Chung, K. C. (2011) 'The levels of evidence and their role in evidence-based medicine', *Plastic and reconstructive surgery*, 128(1), pp. 305–310.

Ceglowski, M. (2010) *Scott And Scurvy*, *Idlewords*. Available at:

https://idlewords.com/2010/03/scott_and_scurvy.htm.

Centers for Disease Control and Prevention (no date) *Cholera – Vibrio cholerae infection*, *Information for Public Health & Medical Professionals*. Available at: https://www.cdc.gov/cholera/healthprofessionals.html (Accessed: 10 June 2019).

Chang, D. *et al.* (2017) 'A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci', *Nature genetics*, 49(10), pp. 1511–1516.

Cheung, L. K. *et al.* (2013) 'Racial disparity in short-term outcomes after gastric bypass surgery', *Obesity surgery*, 23(12), pp. 2096–2103.

Che, Z. *et al.* (2018) 'Recurrent Neural Networks for Multivariate Time Series with Missing Values', *Scientific Reports*. doi: 10.1038/s41598-018-24271-9.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., *et al.* (2018) 'Opportunities and obstacles for deep learning in biology and medicine', *Journal of the Royal Society, Interface / the Royal Society*, 15(141). doi: 10.1098/rsif.2017.0387.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M. and Others (2018) 'Opportunities and obstacles for deep learning in biology and medicine', *Journal of the Royal Society, Interface / the Royal Society*. The Royal Society, 15(141), p. 20170387.

Choi, E. *et al.* (2017) 'Using recurrent neural network models for early detection of heart failure onset', *Journal of the American Medical Informatics Association: JAMIA*, 24(2), pp. 361–370.

Choi, Y., Chiu, C. Y.-I. and Sontag, D. (2016) 'Learning Low-Dimensional Representations of Medical Concepts', *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2016, pp. 41–50.

Chou, R. C. *et al.* (2016) 'Treatment for Rheumatoid Arthritis and Risk of Alzheimer's Disease: A Nested Case-Control Analysis', *CNS drugs*, 30(11), pp. 1111–1120.

Crown, W. H. (2015) 'Potential application of machine learning in health outcomes research and some statistical cautions', *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 18(2), pp. 137–140.

Darweesh, S. K. L. *et al.* (2017) 'Trajectories of prediagnostic functioning in Parkinson's disease', *Brain: a journal of neurology*, 140(2), pp. 429–441.

Deng, H., Wang, P. and Jankovic, J. (2018) 'The genetics of Parkinson disease', *Ageing research reviews*, 42, pp. 72–85.

Denny, J. C. *et al.* (2010) 'PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations', *Bioinformatics* , 26(9), pp. 1205–1210.

English, W. J. *et al.* (2018) ‘American Society for Metabolic and Bariatric Surgery estimation of metabolic and bariatric procedures performed in the United States in 2016’, *Surgery for obesity and related diseases: official journal of the American Society for Bariatric Surgery*, 14(3), pp. 259–263.

Faustino, P. R. *et al.* (2019) ‘Risk of Developing Parkinson Disease in Bipolar Disorder: A Systematic Review and Meta-analysis’, *JAMA neurology*. doi: 10.1001/jamaneurol.2019.3446.

Fernández-Jarne, E. *et al.* (2002) ‘Risk of first non-fatal myocardial infarction negatively associated with olive oil consumption: a case-control study in Spain’, *International journal of epidemiology*, 31(2), pp. 474–480.

Flaherty, D. K. (2011) ‘The vaccine-autism connection: a public health crisis caused by unethical medical practices and fraudulent science’, *The Annals of pharmacotherapy*, 45(10), pp. 1302–1304.

Fleischmann, M. and Pons, S. (1989) ‘Electrochemically induced nuclear fusion of deuterium’, *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, pp. 301–308. doi: 10.1016/0022-0728(89)80006-3.

Futoma, Joseph and Hariharan, Sanjay and Heller, Katherine (2017) ‘Learning to detect sepsis with a multitask gaussian process rnn classifier’, *arXiv preprint arXiv:1706.04152*.

Gaig, C. and Tolosa, E. (2009) ‘When does Parkinson’s disease begin?’, *Movement Disorders*, pp. S656–S664. doi: 10.1002/mds.22672.

Garnotel, R. *et al.* (1998) ‘Long-term variability of serum lipoprotein(a) concentrations in healthy fertile women’, *Clinical chemistry and laboratory medicine: CCLM / FESCC*, 36(5), pp. 317–321.

Gelman, A. (2018) ‘Don’t Calculate Post-hoc Power Using Observed Estimate of Effect Size’, *Annals of Surgery*, p. 1. doi: 10.1097/00000658-900000000-95527.

Giles, C. L., Lawrence, S. and Tsoi, A. C. (2001) ‘10.1023/A:1010884214864’, *Machine Learning*, pp. 161–183. doi: 10.1023/A:1010884214864.

Glass, D. J. (2010) ‘A Critique of the Hypothesis, and a Defense of the Question, as a Framework for Experimentation’, *Clinical Chemistry*, pp. 1080–1085. doi: 10.1373/clinchem.2010.144477.

Glass, D. J. and Hall, N. (2008) ‘A Brief History of the Hypothesis’, *Cell*, pp. 378–381. doi: 10.1016/j.cell.2008.07.033.

Goldstone, J. A. *et al.* (2010) ‘A Global Model for Forecasting Political Instability’, *American Journal of Political Science*, pp. 190–208. doi: 10.1111/j.1540-5907.2009.00426.x.

Gonera, E. G. *et al.* (1997) ‘Symptoms and duration of the prodromal phase in Parkinson’s disease’, *Movement disorders: official journal of the Movement Disorder Society*, 12(6), pp.

871–876.

Graham, J. W. (2009) ‘Missing data analysis: making it work in the real world’, *Annual review of psychology*, 60, pp. 549–576.

Graves, A., Mohamed, A.-R. and Hinton, G. (2013) ‘Speech recognition with deep recurrent neural networks’, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. doi: 10.1109/icassp.2013.6638947.

Griggs, C. L. *et al.* (2018) ‘National Trends in the Use of Metabolic and Bariatric Surgery Among Pediatric Patients With Severe Obesity’, *JAMA pediatrics*, 172(12), pp. 1191–1192.

Guasch-Ferré, M. *et al.* (2015) ‘Olive oil consumption and risk of type 2 diabetes in US women’, *The American journal of clinical nutrition*, 102(2), pp. 479–486.

Haider, A. H., Bilimoria, K. Y. and Kibbe, M. R. (2018) ‘A Checklist to Elevate the Science of Surgical Database Research’, *JAMA surgery*, pp. 505–507.

Hardy, J. *et al.* (2009) ‘The genetics of Parkinson’s syndromes: a critical review’, *Current Opinion in Genetics & Development*, pp. 254–265. doi: 10.1016/j.gde.2009.03.008.

Hawthorne, J. (2018) ‘Inductive Logic’, in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*.

Haynes, W. (2013) ‘Bonferroni Correction’, *Encyclopedia of Systems Biology*, pp. 154–154. doi: 10.1007/978-1-4419-9863-7_1213.

Hedley, A. A. (2004) ‘Prevalence of Overweight and Obesity Among US Children, Adolescents, and Adults, 1999-2002’, *JAMA*, p. 2847. doi: 10.1001/jama.291.23.2847.

Heneghan, C. *et al.* (2009) ‘Diagnostic strategies used in primary care’, *BMJ*, pp. b946–b946. doi: 10.1136/bmj.b946.

Hernan, M. A. (2006) ‘Estimating causal effects from epidemiological data’, *Journal of Epidemiology & Community Health*, pp. 578–586. doi: 10.1136/jech.2004.029496.

Himes, B. E. *et al.* (2009) ‘Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records’, *Journal of the American Medical Informatics Association: JAMIA*, 16(3), pp. 371–379.

Hippe, D. S. *et al.* (2018) ‘Lp(a) (Lipoprotein(a)) Levels Predict Progression of Carotid Atherosclerosis in Subjects With Atherosclerotic Cardiovascular Disease on Intensive Lipid Therapy: An Analysis of the AIM-HIGH (Atherothrombosis Intervention in Metabolic Syndrome With Low HDL/High Triglycerides: Impact on Global Health Outcomes) Carotid Magnetic Resonance Imaging Substudy-Brief Report’, *Arteriosclerosis, thrombosis, and vascular biology*, 38(3), pp. 673–678.

Hughes, G. F. *et al.* (2002) ‘Improved disk-drive failure warnings’, *IEEE Transactions on*

Reliability, pp. 350–357. doi: 10.1109/tr.2002.802886.

Hume, D. (1739) 'A Treatise of Human Nature', *David Hume: A Treatise of Human Nature (Second Edition)*. doi: 10.1093/oseo/instance.00046221.

Husten, L. (2019) 'Beware the hype over the Apple Watch heart app. The device could do more harm than good', *STAT News*.

Inoue, J. *et al.* (2016) 'Resuscitative endovascular balloon occlusion of the aorta might be dangerous in patients with severe torso trauma: A propensity score analysis', *The journal of trauma and acute care surgery*, 80(4), pp. 559–66; discussion 566–7.

Integrative Analysis of Lung Cancer Etiology and Risk (INTEGRAL) Consortium for Early Detection of Lung Cancer *et al.* (2018) 'Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins', *JAMA oncology*, 4(10), p. e182078.

Ivezaj, V. and Grilo, C. M. (2015) 'When mood worsens after gastric bypass surgery: characterization of bariatric patients with increases in depressive symptoms following surgery', *Obesity surgery*, 25(3), pp. 423–429.

Iwaki, H. *et al.* (2019) 'Genetic risk of Parkinson disease and progression:: An analysis of 13 longitudinal cohorts', *Neurology. Genetics*, 5(4), p. e348.

Jacobson, T. A. (2013) 'Lipoprotein(a), Cardiovascular Disease, and Contemporary Management', *Mayo Clinic Proceedings*, pp. 1294–1311. doi: 10.1016/j.mayocp.2013.09.003.

Johnson-Mann, C. *et al.* (2019) 'Investigating racial disparities in bariatric surgery referrals', *Surgery for obesity and related diseases: official journal of the American Society for Bariatric Surgery*, 15(4), pp. 615–620.

Jones-Corneille, L. R., Wadden, T. A. and Sarwer, D. B. (2007) 'Risk of Depression and Suicide in Patients with Extreme Obesity Who Seek Bariatric Surgery', *Obesity Management*, pp. 255–260. doi: 10.1089/obe.2007.0114.

Joseph, B. *et al.* (2019) 'Nationwide Analysis of Resuscitative Endovascular Balloon Occlusion of the Aorta in Civilian Trauma', *JAMA surgery*, 154(6), pp. 500–508.

Keene, C Dirk and Montine, Thomas J and Kuller, Lewis H (2016) 'Epidemiology, pathology, and pathogenesis of Alzheimer disease', *UpToDate, Waltham, MA*.

Kennedy, L. (2008) 'Long-Term Mortality after Gastric Bypass Surgery', *Yearbook of Endocrinology*, pp. 4–6. doi: 10.1016/s0084-3741(08)79149-2.

King, G. and Zeng, L. (2001) 'Improving Forecasts of State Failure', *World Politics*, pp. 623–658. doi: 10.1353/wp.2001.0018.

Kyriacou, D. N. (2004) 'Evidence-based medical decision making: deductive versus inductive logical thinking', *Academic emergency medicine: official journal of the Society for Academic*

Emergency Medicine, 11(6), pp. 670–671.

Lang, A. E. and Espay, A. J. (2018) 'Disease Modification in Parkinson's Disease: Current Approaches, Challenges, and Future Considerations', *Movement disorders: official journal of the Movement Disorder Society*, 33(5), pp. 660–677.

Larson, B. A. O. (2016) 'Risk factors for cognitive decline and dementia', *UpToDate, Waltham, MA*.

Lerche, S. *et al.* (2014) 'Risk factors and prodromal markers and the development of Parkinson's disease', *Journal of neurology*, 261(1), pp. 180–187.

Lewallen, S. and Courtright, P. (1998) 'Epidemiology in practice: case-control studies', *Community eye health / International Centre for Eye Health*, 11(28), pp. 57–58.

Lewis, J. D. *et al.* (2005) 'The relationship between time since registration and measured incidence rates in the General Practice Research Database', *Pharmacoepidemiology and drug safety*, 14(7), pp. 443–451.

Lind, J. (1772) *A Treatise on the Scurvy: In Three Parts. Containing an Inquiry Into the Nature, Causes, and Cure, of that Disease. Together with a Critical and Chronological View of what Has Been Published on the Subject*.

Lin, Q. *et al.* (2018) 'Predicting Node failure in cloud service systems', *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018*. doi: 10.1145/3236024.3236060.

Loftus, P. (2019) 'Apple Watch Has Mixed Results in Big Heart Study', *Wall Street Journal*.

Luchsinger, J. A. *et al.* (2005) 'Aggregation of vascular risk factors and risk of incident Alzheimer disease', *Neurology*, 65(4), pp. 545–551.

Luppino, F. S. *et al.* (2010) 'Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies', *Archives of general psychiatry*, 67(3), pp. 220–229.

Mahlknecht, P., Seppi, K. and Poewe, W. (2015) 'The Concept of Prodromal Parkinson's Disease', *Journal of Parkinson's Disease*, pp. 681–697. doi: 10.3233/jpd-150685.

Marshall, T. (2004) 'What is a case-control study?', *International Journal of Epidemiology*, pp. 612–613. doi: 10.1093/ije/dyh055.

Martin-Fernandez, K. W., Heinberg, L. J. and Ben-Porath, Y. S. (2019) 'Using the preoperative psychological evaluation to determine psychosocial risk factors for CPAP nonadherence among bariatric surgery candidates', *Surgery for obesity and related diseases: official journal of the American Society for Bariatric Surgery*, 15(12), pp. 2115–2120.

Marzban, C. and Stumpf, G. J. (1996) 'A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes', *Journal of Applied Meteorology*, pp. 617–626. doi:

2.0.co;2">10.1175/1520-0450(1996)035<0617:annftp>2.0.co;2.

McGovern, A. *et al.* (2011) 'Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction', *Data Mining and Knowledge Discovery*, pp. 232–258. doi: 10.1007/s10618-010-0193-7.

McGuire, M. T. *et al.* (1999) 'What predicts weight regain in a group of successful weight losers?', *Journal of consulting and clinical psychology*, 67(2), pp. 177–185.

McIntosh, K. (2020) 'Coronavirus disease 2019 (COVID-19)', in Martin S Hirsch, A. B. (ed.) *UpToDate*.

Mechanick, J. I. *et al.* (2013) 'Clinical practice guidelines for the perioperative nutritional, metabolic, and nonsurgical support of the bariatric surgery patient--2013 update: cosponsored by American Association of Clinical Endocrinologists, The Obesity Society, and American Society for Metabolic & Bariatric Surgery', *Obesity* , 21 Suppl 1, pp. S1–27.

Mikolov, Tomas and Karafiat, Martin and Burget, Lukas and Cernocky, Jan and Khudanpur, Sanjeev (2010) 'Recurrent neural network based language model', *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff (2013) 'Distributed representations of words and phrases and their compositionality', *Advances in neural information processing systems*.

Mill, J. S. (1843) *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*.

Mintz, E., Slayton, R. and Walters, M. (2015) 'Typhoid Fever and Paratyphoid Fever', *Control of Communicable Diseases Manual*. doi: 10.2105/ccdm.2745.149.

Miskelly, G. M. *et al.* (1989) 'Analysis of the published calorimetric evidence for electrochemical fusion of deuterium in palladium', *Science*, 246(4931), pp. 793–796.

Mitchell, J. E. *et al.* (2014) 'Course of depressive symptoms and treatment in the longitudinal assessment of bariatric surgery (LABS-2) study', *Obesity* , 22(8), pp. 1799–1806.

Murray, Joseph F, Hughes, Gordon F and Kreutz-Delgado, Kenneth (2005) 'Machine learning methods for predicting failures in hard drives: A multiple-instance application', *Journal of machine learning research: JMLR*, 6, pp. 783–816.

Nalls, M. A. *et al.* (2014) 'Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease', *Nature genetics*, 46(9), pp. 989–993.

Nazir, D. J. *et al.* (1999) 'Monthly intra-individual variation in lipids over a 1-year period in 22 normal subjects', *Clinical biochemistry*, 32(5), pp. 381–389.

Newsom, S. W. B. (2006) 'Pioneers in infection control: John Snow, Henry Whitehead, the

Broad Street pump, and the beginnings of geographical epidemiology', *Journal of Hospital Infection*, pp. 210–216. doi: 10.1016/j.jhin.2006.05.020.

Nguyen, N. T. and Varela, J. E. (2017) 'Bariatric surgery for obesity and metabolic disorders: state of the art', *Nature reviews. Gastroenterology & hepatology*, 14(3), pp. 160–169.

Norgeot, B. *et al.* (2019) 'Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis', *JAMA network open*, 2(3), p. e190606.

Norii, T., Crandall, C. and Terasaka, Y. (2015) 'Survival of severe blunt trauma patients treated with resuscitative endovascular balloon occlusion of the aorta compared with propensity score-adjusted untreated patients', *The journal of trauma and acute care surgery*, 78(4), pp. 721–728.

O'Gara, P. T., Kushner, F. G. and Ascheim, D. D. (2013) '2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: a report of the American College of Cardiology Foundation/American Heart …', *Journal of the*. onlinejacc.org. Available at: http://www.onlinejacc.org/content/61/4/e78?_ga=2.173241763.1693228286.1553184676-1221175421.1553184676.

Otsuka, H. *et al.* (2018) 'Effect of resuscitative endovascular balloon occlusion of the aorta in hemodynamically unstable patients with multiple severe torso trauma: a retrospective study', *World journal of emergency surgery: WJES*, 13, p. 49.

Paneth, N., Susser, E. and Susser, M. (2002) 'Origins and early development of the case-control study: Part 1, Early evolution', *Sozial- und Praventivmedizin*, 47(5), pp. 282–288.

Paré, G. *et al.* (2019) 'Lipoprotein(a) Levels and the Risk of Myocardial Infarction Among 7 Ethnic Groups', *Circulation*, 139(12), pp. 1472–1482.

Perez, M. V. *et al.* (2019) 'Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation', *The New England journal of medicine*, 381(20), pp. 1909–1917.

Platnick, N. I. and Popper, K. R. (1977) 'Conjectures and Refutations: The Growth of Scientific Knowledge', *Systematic Zoology*, p. 363. doi: 10.2307/2412688.

Poewe, W. *et al.* (2017) 'Parkinson disease', *Nature Reviews Disease Primers*. Nature Publishing Group, 3(1), pp. 1–21.

Postuma, R. B. *et al.* (2012) 'Identifying prodromal Parkinson's disease: pre-motor disorders in Parkinson's disease', *Movement disorders: official journal of the Movement Disorder Society*, 27(5), pp. 617–626.

Premier Applied Sciences (2019) *Premier Healthcare Database*. https://products.premierinc.com/downloads/PremierHealthcareDatabaseWhitepaper.pdf.

Pringsheim, T. *et al.* (2014) 'The prevalence of Parkinson's disease: A systematic review and meta-analysis', *Movement Disorders*, pp. 1583–1590. doi: 10.1002/mds.25945.

Queiroz, L. P. *et al.* (2017) 'A Fault Detection Method for Hard Disk Drives Based on Mixture of Gaussians and Nonparametric Statistics', *IEEE Transactions on Industrial Informatics*, pp. 542–550. doi: 10.1109/tii.2016.2619180.

Rajkomar, A. *et al.* (2018) 'Scalable and accurate deep learning with electronic health records', *npj Digital Medicine*, 1(1), p. 18.

Rajkomar, A., Dean, J. and Kohane, I. (2019) 'Machine Learning in Medicine', *The New England journal of medicine*, 380(14), pp. 1347–1358.

Rand, L. I. *et al.* (1985) 'Multiple factors in the prediction of risk of proliferative diabetic retinopathy', *The New England journal of medicine*, 313(23), pp. 1433–1438.

Ranganath, R. *et al.* (2016) 'Deep Survival Analysis', *arXiv preprint arXiv:1608.02158*.

Razavian, Narges and Sontag, David (2015) 'Temporal convolutional neural networks for diagnosis from lab tests', *arXiv preprint arXiv:1511.07938*.

Roberts, R. E. *et al.* (2000) 'Are the Obese at Greater Risk for Depression?', *American Journal of Epidemiology*, pp. 163–170. doi: 10.1093/aje/152.2.163.

Ross, G. W. *et al.* (2008) 'Association of olfactory dysfunction with risk for future Parkinson's disease', *Annals of neurology*, 63(2), pp. 167–173.

Rothman, K. J. (2012) *Epidemiology: An Introduction*. Oxford University Press.

Rydén, A. and Torgerson, J. S. (2006) 'The Swedish Obese Subjects Study—what has been accomplished to date?', *Surgery for Obesity and Related Diseases*, pp. 549–560. doi: 10.1016/j.soard.2006.07.006.

Sardi, S. P. and Simuni, T. (2019) 'New Era in disease modification in Parkinson's disease: Review of genetically targeted therapeutics', *Parkinsonism & related disorders*, 59, pp. 32–38.

Schlessinger, D. I. *et al.* (2019) 'Artificial intelligence and dermatology: opportunities, challenges, and future directions', *Seminars in cutaneous medicine and surgery*, 38(1), pp. E31–37.

Schrag, A. *et al.* (2015) 'Prediagnostic presentations of Parkinson's disease in primary care: a case-control study', *Lancet neurology*, 14(1), pp. 57–64.

Schrag, A. *et al.* (2019) 'Predicting diagnosis of Parkinson's disease: A risk algorithm based on primary care presentations', *Movement disorders: official journal of the Movement Disorder Society*, 34(4), pp. 480–486.

Shadlen, Marie-Florence and Larson, Eric B (2010) 'Evaluation of cognitive impairment and dementia', *Waltham, MA: UpToDate*.

Sheets, C. S. *et al.* (2015) 'Post-operative psychosocial predictors of outcome in bariatric

surgery', *Obesity surgery*, 25(2), pp. 330–345.

Shickel, B. *et al.* (2018) 'Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis', *IEEE journal of biomedical and health informatics*. ieeexplore.ieee.org, 22(5), pp. 1589–1604.

Simpson, S. M. *et al.* (2007) 'Racial disparities in diagnosis and treatment of depression: a literature review', *The Psychiatric quarterly*, 78(1), pp. 3–14.

Slee, V. N. (1978) 'The International Classification of Diseases: Ninth Revision (ICD-9)', *Annals of Internal Medicine*, p. 424. doi: 10.7326/0003-4819-88-3-424.

Snow, J. (1856) 'On the Mode of Communication of Cholera', *Edinburgh medical journal*, 1(7), pp. 668–670.

Song, J. W. and Chung, K. C. (2010) 'Observational Studies: Cohort and Case-Control Studies', *Plastic and Reconstructive Surgery*, pp. 2234–2242. doi: 10.1097/prs.0b013e3181f44abc.

Speck, C. E. *et al.* (1995) 'History of depression as a risk factor for Alzheimer's disease', *Epidemiology* , 6(4), pp. 366–369.

Stannard, A., Eliason, J. L. and Rasmussen, T. E. (2011) 'Resuscitative Endovascular Balloon Occlusion of the Aorta (REBOA) as an Adjunct for Hemorrhagic Shock', *The Journal of Trauma: Injury, Infection, and Critical Care*, pp. 1869–1872. doi: 10.1097/ta.0b013e31823fe90c.

Steere, A. C. *et al.* (1977) 'Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three connecticut communities', *Arthritis and rheumatism*, 20(1), pp. 7–17.

Steere, A. C. *et al.* (2016) 'Lyme borreliosis', *Nature reviews. Disease primers*, 2, p. 16090.

Sudlow, C. *et al.* (2015) 'UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age', *PLoS medicine*, 12(3), p. e1001779.

Sundermeyer, Martin and Schluter, Ralf and Ney, Hermann (2012) 'LSTM neural networks for language modeling', *Thirteenth annual conference of the international speech communication association*.

Tabbal, S. D. *et al.* (2012) 'Low nigrostriatal reserve for motor parkinsonism in nonhuman primates', *Experimental neurology*, 237(2), pp. 355–362.

Tierney, L. (2012) 'The R Statistical Computing Environment', *Lecture Notes in Statistics*, pp. 435–447. doi: 10.1007/978-1-4614-3520-4_41.

Tindle, H. A. *et al.* (2010) 'Risk of suicide after long-term follow-up from bariatric surgery', *The American journal of medicine*, 123(11), pp. 1036–1042.

Topol, E. J. (2019) 'High-performance medicine: the convergence of human and artificial

intelligence', *Nature medicine*, 25(1), pp. 44–56.

UnitedHealthcare (2018) 'Bariatric Surgery, Policy Number 2018T0362AA'.

Vandenbroucke, J. P. *et al.* (2014) 'Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration', *International journal of surgery* , 12(12), pp. 1500–1524.

Wakefield, A. J. *et al.* (1998) 'Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children', *The Lancet*, 351(9103), pp. 637–641.

Wang, X. *et al.* (2014) 'Exploring joint disease risk prediction', *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014, pp. 1180–1187.

Wang, Y. *et al.* (2014) 'A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives', *IEEE Transactions on Industrial Informatics*, pp. 419–430. doi: 10.1109/tii.2013.2264060.

Weiskopf, N. G. and Weng, C. (2013) 'Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research', *Journal of the American Medical Informatics Association: JAMIA*. academic.oup.com, 20(1), pp. 144–151.

Weiss, J. C. *et al.* (2012) 'Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records', *AI Magazine*, p. 33. doi: 10.1609/aimag.v33i4.2438.

Wheeler, E. (2008) 'Adherence to outpatient program postoperative appointments after bariatric surgery', *Surgery for obesity and related diseases: official journal of the American Society for Bariatric Surgery*. Elsevier, 4(4), pp. 515–520.

Whitehead, H. (1865) 'THE BROAD STREET PUMP: AN EPISODE IN THE CHOLERA EPIDEMIC OF 1854', *Macmillan's magazine*, pp. 113–122.

Whitmer, R. A. *et al.* (2005) 'Midlife cardiovascular risk factors and risk of dementia in late life', *Neurology*, pp. 277–281. doi: 10.1212/01.wnl.0000149519.47454.f2.

de Wit, L. M. *et al.* (2009) 'Depression and body mass index, a u-shaped association', *BMC public health*, 9, p. 14.

Wolk, David A and Dickerson, Bradford C (2016) 'Clinical features and diagnosis of Alzheimer disease', *UpToDate, Waltham, MA*.

Wright, A. (1900) 'ON THE PATHOLOGY AND THERAPEUTICS OF SCURVY', *The Lancet*, pp. 565–567. doi: 10.1016/s0140-6736(01)99819-8.

Yokota J. (2016) '[Japan Trauma Data Bank (JTDB) managed by Japan Trauma Care and Research (JTCR)]', *Nihon rinsho. Japanese journal of clinical medicine*, 74(2), pp. 329–336.

Yusuf, S. *et al.* (2004) ‘Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study’, *The Lancet*, 364(9438), pp. 937–952.