Essays in Health and Behavioral Economics

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citable link</th>
<th><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:40050125">http://nrs.harvard.edu/urn-3:HUL.InstRepos:40050125</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Essays in Health and Behavioral Economics

A dissertation presented
by

Stephen Coussens

to
The Department of Public Policy

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Public Policy

Harvard University
Cambridge, Massachusetts
April 2018
Abstract

This dissertation consists of three chapters, each of which is an independent essay in the fields of health or behavioral economics. The first chapter explores the use of heuristics among highly-trained physicians diagnosing heart disease in the emergency department (ED), a common task with life-or-death consequences. Using data from a large private-payer claims database, I find compelling evidence of heuristic thinking in this setting: patients arriving in the ED just after their 40th birthday are roughly 10% more likely to be tested for and 20% more likely to be diagnosed with ischemic heart disease (IHD) than patients arriving just before this date, despite the fact that the incidence of heart disease increases smoothly with age. Moreover, I show that this shock to diagnostic intensity has meaningful implications for patient health, as it reduces the number of missed IHD diagnoses among patients arriving in the emergency department just after their 40th birthday, thereby preventing future heart attacks. I then develop a model that ties this behavior to an existing literature on representativeness heuristics, and discuss the implications of this class of heuristics for diagnostic decision-making.

The second chapter examines the admitting decisions of ED physicians. Roughly half of all hospital admissions in the US flow through EDs, making emergency
physicians important gatekeepers for expensive, high-intensity inpatient care. However, despite the high costs associated with hospital admissions, even physicians working in the same ED exhibit wide variation in their tendencies to admit patients to the hospital. This begs the question: are the additional admissions made by high-admitting physicians conferring benefits that are sufficient to outweigh their substantial costs? By exploiting quasi-random assignment of patients to physicians in a large Boston-area ED, I find that physicians with above-median admission rates are on average at least 20% more likely to admit a given patient than those with below-median admission rates. Although patients assigned to high-admitting physicians receive significantly greater treatment intensity, I find that this makes them no less likely to experience adverse health outcomes in the future. This suggests that these marginal admissions are of low value.

The final chapter proposes a methodological contribution to the design of randomized control trials (RCTs), which are ubiquitous in the health sciences. Statistical power increases in the compliance rate of an experiment, so selecting participants with the highest likelihoods of compliance can provide the experiment with more power than if participants were chosen randomly. In this paper, I explore how data from prior experiments or quasi-experiments can allow researchers to systematically select the potential participants that are most likely to be compliers, and discuss the potential benefits and drawbacks of incorporating such an approach into RCT design. Using publicly available data from the Oregon Health Insurance Experiment, I empirically demonstrate the feasibility of these methods.
Contents

Abstract ................................................................. iii
Acknowledgments ....................................................... x

1 Behaving Discretely: Heuristic Thinking in the Emergency Department 1
1.1 Introduction .......................................................... 1
1.2 The Clinical Decision ................................................ 4
1.3 Data ................................................................. 6
1.4 Empirical Framework ................................................ 9
1.5 Results .............................................................. 12
1.6 Theoretical Framework ............................................. 37
1.7 Policy Remedies ..................................................... 45
1.8 Conclusion .......................................................... 47

2 Worth the Price of Admission? Evidence from Emergency Department Admissions 49
2.1 Introduction .......................................................... 49
2.2 The Admission Decision ............................................ 51
2.3 Data ................................................................. 53
2.4 Empirical Framework .............................................. 55
2.5 Results .............................................................. 58
2.6 Discussion ........................................................ 73
2.7 Conclusion .......................................................... 75

3 Fight the Power [Calculation]: Incorporating Compliance Prediction into RCT Design 77
3.1 Introduction .......................................................... 77
3.2 Treatment Effects Framework ..................................... 79
3.3 Estimation of Treatment Effects ................................. 83
List of Tables

1.1 Heart attack testing rates ................................................. 14
1.2 IHD diagnosis rates ...................................................... 17
1.3 IHD admission rates ..................................................... 18
1.4 Heart attack testing rates, with and without chest pain (CP) .... 18
1.5 Prevalence of relevant characteristics above and below the age-40 threshold ......................................................... 19
1.6 Subsequent heart attack diagnosis rates .............................. 24
1.7 Subsequent heart attack diagnosis rates among female patients with and without chest pain (CP) ................................. 28
1.8 Subsequent heart attack diagnosis rates among male patients with and without chest pain (CP) ................................. 28
1.9 Heart attack testing rates, Boston-area ED .......................... 36

2.1 Patient Characteristics .................................................... 54
2.2 Characteristics of patients assigned to high- and low-admitting physicians ................................................................. 59
2.3 Relationship with Patient Predicted Mortality ....................... 60
2.4 Effect on Likelihood of Patient Admission to Hospital ............ 61
2.5 Effect on ED Charges (in Standard Deviations) ....................... 61
2.6 Effect on Likelihood of Patient Being Held for Observation ......... 62
2.7 Effect on ED Length of Stay (in Hours) .............................. 63
2.8 Effect on Subsequent ED Visit Rates .................................. 64
2.9 Effect on Subsequent Admission Rates ............................... 65
2.10 Effect on Subsequent Mortality Rates ............................... 65
2.11 Effect on Admission Rates for Patients with Positive Troponin Test ................................................................. 67
2.12 Effect on Admission Rates for Patients with Negative Troponin Test ................................................................. 68
2.13 Effect on Admission Rates for Patients not given Troponin Test ................................................................. 69
2.14 Effect on Troponin Testing Rates ..................................... 69
2.15 Relationship between Admission Rates and Physician Case Volume 73

3.1 Application #1 first stage estimates .......................... 93
3.2 Application #1 LATE estimates ............................. 94
3.3 Application #2 LATE estimates ............................. 96

A.1 Heart attack testing rates, by bandwidth ....................... 110
A.2 IHD diagnosis rates, by bandwidth .......................... 111
A.3 IHD admission rates, by bandwidth .......................... 111
A.4 Subsequent heart attack diagnosis rates ....................... 112
A.5 Average aggregate IHD testing cost ......................... 114
List of Figures

1.1 Proportion of ED patients tested for heart attack ................. 13
1.2 Proportion of ED patients diagnosed with IHD .................. 15
1.3 Proportion of ED patients admitted to hospital with IHD diagnosis . 16
1.4 Number of ED visits, by month-age bin .......................... 20
1.5 Subsequent heart attack diagnosis rate, by patient sex .......... 25
1.6 Subsequent heart attack diagnosis rate, among female patients with and without chest pain ................................. 26
1.7 Subsequent heart attack diagnosis rate, among male patients with and without chest pain ..................................... 27
1.8 Proportion of patients tested for heart attack in Boston-area ED .. 33
1.9 Proportion of patients tested for heart attack in Boston-area ED, by patient volume .............................................. 34
1.10 Proportion of patients tested for heart attack in Boston-area ED, related to attending physician’s shift ......................... 35
1.11 Annual heart attack deaths per capita ................................. 38
1.12 Piecewise linear fit - annual heart attack deaths per capita ...... 39
1.13 Proportion of ED patients tested for heart attack ................. 40

2.1 Relating physician admission rates to experience .................. 70
2.2 Relating physician admission rates to caseload .................... 72

3.1 Estimated power using participants most likely to comply .......... 95
3.2 Standardized LATE as a function of predicted compliance .......... 97
Acknowledgments

First and foremost, I thank my advisors Brigitte Madrian, David Cutler, and Ziad Obermeyer for their guidance and support throughout the past several years. I am honored to have had the opportunity to learn from them firsthand. I am also grateful for the financial support of my graduate studies provided by Harvard University, the Harvard Kennedy School, and the Taubman Center for State and Local Government.

The research in this dissertation has also benefited from conversations with a number of faculty at Harvard, including John Beshears, Katherine Coffman, Nathan Hendren, Sendhil Mullainathan, and Daniel Shoag. I also thank the participants at the Harvard Economics Department Labor Workshop, Labor Lunch, and Harvard Medical School Health Economics Seminar for their helpful comments and suggestions.

The past five years would not have been nearly as enjoyable without my classmates, both as colleagues and companions. Most people don’t think of doctoral programs as being synonymous with fun, but thanks to my cohort, I do.
To Phuong
Chapter 1

Behaving Discretely: Heuristic Thinking in the Emergency Department

1.1 Introduction

Most economic models make the simplifying assumption that the decision-maker is able to precisely optimize choice using all available information, both continuous and discrete in nature. Yet there also exists a well-established literature in both economics and psychology demonstrating the important role of heuristic decision-making (and the biases it can generate) in a variety of contexts (Gilovich et al., 2002; Kahneman et al., 1982). Evidence in both experimental (List, 2003, 2004; Alevy et al., 2015) and non-experimental (Lacetera et al., 2012) settings indicates that the effects of "nonstandard decision-making" (DellaVigna, 2009) on market transactions tend to be concentrated among naive agents with limited experience making the decision at hand. This suggests that anomalous behavior may not have
a meaningful impact in settings in which the decision-makers are highly trained or experienced.

This paper provides evidence of an important environment in which this notion is violated. I explore the use of heuristics in a high-stakes setting in which the agents are highly trained professionals making familiar decisions. Specifically, I examine physician treatment patterns in the emergency department (ED) involving the diagnosis of ischemic heart disease (IHD), a common yet life-threatening condition. Given that age is a strong predictor of heart disease, the manner in which physicians incorporate this continuous attribute into their assessment of a patient’s risk for IHD is likely to meaningfully influence treatment decisions. If physicians employ a cognitive shortcut that discretizes age into coarse categories, patients falling on either side of a category boundary may receive substantially different treatment, despite being otherwise similar.

Using private-payer health insurance claims for over 5 million ED visits in the U.S. between 2005 and 2013, I find evidence that emergency physicians use a heuristic that classifies a subset of these patients as higher-risk for IHD if age 40 years or older, and lower-risk otherwise. This rule of thumb results in a sharp discontinuity in treatment patterns: patients entering the ED are roughly 10% more likely to be tested for a heart attack (a severe form of IHD) if they arrive shortly after their 40th birthday than if they had arrived just before this date. I argue that this is consistent with a representativeness-based heuristic (Kahneman and Tversky, 1972) whereby patients in their 30s are less representative of the prototypical heart attack patient than patients in their 40s. Further, I find that these discontinuities are most pronounced among ED patients possessing fewer of the characteristics most commonly associated with IHD, making their diagnosis less clear-cut.
What mechanisms appear to be driving this heuristic? There are a number of intertemporal factors that could influence physicians’ reliance on cognitive shortcuts like this one. Two of the most intuitive are fatigue and cognitive load. I examine the importance of these possibilities using ED records from a large Boston-area hospital. I find that the age-40 heuristic persists even when the emergency department patient volume is lower than usual (when cognitive load is below-average), as well as when patients arrive near the beginning of the physician’s shift (when the physician is less likely to be fatigued). This suggests that the observed behavior is a relatively stable feature of physician decision-making in this setting.

As in Almond et al. (2010) and Almond and Doyle (2011), I also measure the effect of this discontinuous behavior on patients’ health outcomes. Specifically, I examine the rate of heart attacks subsequent to patients’ ED visits, which serves as a proxy for IHD diagnoses that were likely missed during these visits. I find that the shock to diagnostic intensity generated by the age-40 heuristic substantially reduces the likelihood of missed diagnoses within a group of IHD patients known by the medical community to have an elevated risk of misdiagnosis: women, particularly those presenting without chest pain. This suggests that at least for this subpopulation, the marginal return to IHD testing is high.

Despite this physiologically unjustified discontinuity in treatment intensity at age 40, it is important to note that in a general equilibrium sense, this heuristic is not necessarily suboptimal, as the cognitive effort it saves the physician may yield higher performance along other dimensions. Therefore, I do not draw welfare conclusions about heuristics more generally. Instead, I argue that undesirable features of heuristics represent a potential opportunity to leverage machine-learning prediction in decision-support systems. Heuristics generally provide an efficient means of arriving at near-optimal solutions, so there is likely little reason to dis-
courage their use in general. But if machines can identify narrow instances in which heuristics tend to lead physicians astray, automated interventions in these scenarios might provide a welfare improvement.

The remainder of this paper is organized as follows. Section 1.2 provides a discussion of the nature of decision-making in the ED with respect to patients who may be suffering from heart disease. Section 1.3 describes the data used in my analysis. Section 1.4 discusses the empirical framework used to estimate the discontinuities generated by the age-40 heuristic. Section 1.5 presents the results. Section 1.6 discusses the role of the representativeness heuristic in clinical decision-making, and presents a representativeness-driven model of diagnostic behavior. Section 1.7 discusses the policy implications of heuristic thinking in medicine, and makes a case for an expanded role for electronic decision support systems in automatically identifying and attempting to correct problematic behavior. Section 1.8 concludes.

1.2 The Clinical Decision

In many medical specialties, physicians are able to accrue knowledge of their patients’ health status over the course of several visits in which they observe, diagnose, treat, and follow-up with their patients. However, physicians employed in a hospital’s emergency department are practicing in a uniquely low-information setting. Often the only evidence that they may use to assess a patient’s risk for a particular condition is the patient’s vital signs, demographic characteristics, medical history (as imperfectly recalled by the patient), and reported symptoms (Groopman and Prichard, 2007). Moreover, patient volume and limited hospital resources often require that this assessment be made in a matter of minutes.
Despite these limitations on the quantity and quality of the information readily available to emergency physicians, they are tasked with rapidly assessing the nature and severity of their patient’s condition, determining the extent of testing necessary, making a diagnosis, administering treatment, and deciding whether the patient should be admitted to the hospital or discharged. One of the most common diagnostic dilemmas facing ED physicians is the diagnosis of IHD. Given the severe health implications of missed IHD diagnoses, when assessing a patient presenting with symptoms that are consistent with this condition, a primary goal of the physician is to assess the likelihood of IHD as the underlying cause, despite the fact that the vast majority of such patients do not have a heart condition. The physician must therefore rule out an IHD diagnosis in a large number of low-risk patients in order to focus the hospital’s limited resources on those with the greatest risk.

When an ED physician considers the possibility that her patient has experienced a heart attack, she has the option to order first-line tests to aid in the diagnosis. These tests are commonly used and inexpensive.\(^1\) While the tests are not monetarily costly, they require time to be processed – time during which the patient occupies an ED bed. However, EDs frequently operate at (or near) capacity, both in terms of bed and physician availability, so physicians in this setting are therefore unable to test patients indiscriminately. Among patients who appear to be relatively low-risk, how does the physician determine which patients will be tested, and which will be discharged without testing? As mentioned above, the patient history available to the emergency physician is limited, as is the amount of time spent with the

\(^1\)The two most common tests are the troponin blood test and the electrocardiogram (often abbreviated ECG or EKG). While the ECG is frequently used to detect a wide array of conditions, the troponin test is specifically used to detect heart attacks. Therefore, my analyses regarding physician testing decisions will focus on the use of the troponin test.
patient. Given the steep age gradient in the incidence of heart attacks, age is almost
certainly included as a component of this decision. Moreover, among younger
patients, the incidence is at least twice as high for males as for females, making
the sex of the patient a salient predictor as well (Mozaffarian et al., 2016). These
facts, and their relationship to the representativeness heuristic, are discussed in
greater detail in Section 1.6.

When interpreting a continuous characteristic, individuals may discretize it
into coarse categories to which a decision rule can be more easily applied (Lacetera
et al., 2012). Suppose that physicians were to exhibit such a tendency with respect
to patient age while considering a diagnosis, lumping patients into one of two
groups: patients are "young" if under $\alpha$ years of age, and "old" otherwise. Were
this the case, it would likely result in discrete treatment patterns – patients whose
age falls just on either side of $\alpha$ may receive substantially different treatment,
despite being otherwise comparable. Using the data described in Section 1.3, I
identify a pattern consistent with this hypothesis, which motivates the regression
discontinuity (RD) design that I outline in Section 1.4.

1.3 Data

The empirical analysis in this paper makes use of two data sources: private-payer
health insurance claims from the Truven Commercial Claims and Encounters
database, and records from the emergency department of a large Boston-area
hospital, hereafter referred to as the "Truven claims," and "ED records," respectively.
My analysis primarily focuses on the Truven claims, as it contains a far larger
sample of emergency department visits, yielding superior statistical power. I
supplement this approach with insights gleaned from the ED records, which offer a richer description of ED visits than the Truven claims.

Data Description - Truven Claims

The Truven data provide private-payer health insurance claims information (both inpatient and outpatient) for millions of individuals in the US. Truven obtains this information from a number of large employers, governments, and health plans, so although individuals in this data are not a representative sample of the general population, geographic coverage is fairly broad. All claims records contain a unique individual-level identifier that allows for longitudinal tracking of all plan enrollees. This data does not, however, include provider or geographic identifiers. This limits my ability to explore of heterogeneous effects.

Throughout my analysis, I make use of the service dates, ICD-9 diagnosis codes and CPT procedure codes relating to each claim.\(^2\) These allow me to determine the diagnoses made and procedures performed during every patient interaction with a health care provider. Year of birth and sex (but not race or ethnicity) are available for each individual in the database. I am able to determine the month of birth for roughly 95% of individuals in my analysis sample, which provides me with a finer measure of patient age during each visit than if I were to use the patient’s integer age. For the remaining 5% of the sample, I impute the month of birth.\(^3\)

\(^2\)ICD (International Classification of Diseases) codes are the "international standard for reporting diseases and health conditions" (WHO, 2016); only 9th revision codes (i.e. ICD-9) appear in this data. All diagnoses are mapped to ICD codes prior to submitting claims. CPT (Current Procedural Terminology) codes convey the medical procedures and services performed by healthcare providers, and are submitted with all claims.

\(^3\)My results are not sensitive to the exclusion of the 5% of individuals with an imputed month of birth. See Appendix A.1 for details regarding the process I use to determine month of birth.
Analysis Sample - Truven Claims

I impose a number of restrictions on this data to reach the sample used in my analysis. I first restrict the sample to those in the Truven claims who visit a hospital ED between 2005 and 2013. However, I am only able to identify ED heart attack testing claims for the period 2010 through 2013. Therefore, estimates of effects on testing rates are restricted to this period. Estimates for all other outcomes are generated using the full sample (2005 through 2013), which yields very similar point estimates, but provides much greater statistical power. If a patient visits an ED more than once during this period, I include only the patient’s first visit in my sample, as physician behavior in subsequent ED visits may be endogenous to that of prior visits. All patients with a diagnosis of IHD prior to their first ED visit are also excluded. As the focus of this paper pertains to differences in physician treatment decisions around age 40, I further restrict the sample to individuals who were within 5 years of their 40th birthday at the time of their visit. I also exclude patients whose ED visit takes place during the month of their 40th birthday, as I cannot determine these patients’ integer age on the day of their visit. After imposing these restrictions, my analysis sample consists of roughly 5.6 million patients, with more than one million patients arriving in the ED within one year of their 40th birthday.

Data Description - ED Records

While the Truven claims provide me with a large sample from which I can generate precise estimates, it lacks information that might allow me to test a wider range of hypotheses. To supplement my primary analysis of the Truven claims, I conduct an analysis of all ED visits at a large Boston-area hospital between January 2010
and May 2015. In addition to the types of data present in the Truven claims, these records also include timestamps for patient arrival and discharge, as well as attending physician shift schedules.

Analysis Sample - ED Records

The sample restrictions that I impose on the ED records are consistent with those of the Truven data: If a patient visits the ED more than once during this period, I keep only the patient’s first visit in my sample, and include only those who were within 5 years of their 40th birthday at the time of their visit. However, unlike in the Truven claims, since I observe the patients’ exact date of birth, there is no need to exclude patients visiting the ED during their month of birth. After imposing these restrictions, my analysis sample consists of approximately 20,000 patients.

1.4 Empirical Framework

Estimating Equation

Regression discontinuity designs are frequently employed in settings in which individuals’ "treatment" status is at least partially determined by an arbitrary threshold in a continuous characteristic ("running variable") possessed by the population of interest. In this paper, my empirical strategy treats the patients’ age as the running variable. I posit that while patients arriving in the ED shortly before and after their 40th birthday are physiologically very similar, crossing this arbitrary threshold can change the way that they are perceived by the physician, which can in turn affect the manner in which the patient is treated. So while the impact of the age-40 threshold on a physician’s assessment of her patient’s IHD
risk is unobserved, it can be inferred from the subsequent impact of this heuristic on the physician’s actions, such as testing and diagnosis.

In order to measure the discontinuity in physician behavior that is generated by the age-40 heuristic, I estimate a local linear regression using only patients visiting the ED near their 40th birthday, following standard methods such as those outlined in Lee and Lemieux (2010). Specifically, for patients whose age on the date of their ED visit falls within $h$ months of their 40th birthday, I estimate the following model:

$$Y_i = \beta_0 + \beta_1 Turned40 + \beta_2 (Age_i - 40) + \beta_3 Turned40 \times (Age_i - 40) + \epsilon_i \quad (1.1)$$

where $Y_i$ represents a binary decision by a physician (e.g. whether to order a test or make a diagnosis), or a subsequent outcome (such as future heart attacks) for patient $i$, and $Turned40_i = 1\{Age_i \geq 40\}$. The parameter of interest in this equation is $\beta_1$, which captures the discontinuous jump in the average value of $Y_i$ in patients arriving in the ED just after their 40th birthday relative to those arriving just before. This estimate of the discontinuity is unbiased if the underlying functional form of $Y$ is piecewise linear in age within $h$ months of the threshold. The methodology for determining the appropriate bandwidth $h$ is discussed in detail below.

Equation (1.1) is estimated using a triangular kernel centered at the age-40 threshold, so that the weight placed on each observation decays as the difference between the patient’s age and the threshold increases (Cheng et al., 1997). In the Truven claims, the running variable is discrete since patient age is only available in month-age bins. To account for the implications of this discreteness for inference, I compute heteroscedasticity-robust standard errors that are clustered at the month-
age level, following Lee and Card (2008).\footnote{However, my results are insensitive to whether or not the standard errors are clustered.} I select the bandwidth $h$ using the mean squared error-optimal procedure developed by Imbens and Kalyanaraman (2012) and implemented in Calonico et al. (2015). This method yields an optimal bandwidth of 1.24 years. However, since patient age is only observable in month-age bins in the Truven claims data, I round this number to the nearest month, resulting in a bandwidth of $h = 15$ months.\footnote{This optimal bandwidth is estimated using the full analysis sample, with IHD diagnosis as the outcome. Re-estimating the optimal bandwidth for each outcome of interest yields very similar results. Therefore, for ease of interpretation, I use the same 15-month bandwidth for estimation of all discontinuities in the Truven claims. See Appendix A.2 for the robustness of these estimates to alternative bandwidths.}

As alluded to in the introduction to this paper, my empirical strategy is similar to that of the instrumental variables framework applied in Almond et al. (2010): I estimate the effect of a shock on physician behavior, and then estimate the effect of the shock on subsequent health outcomes. While these estimates are intuitively akin to first stage and reduced form estimates, I do not interpret the ratio of the two as the local average treatment effect (LATE). I am able to demonstrate that the age-40 heuristic is a plausibly random shock to the treatment of patients near the threshold. However, there are likely many unobservable pertinent physician decisions that are also affected by the heuristic; these may impact subsequent health outcomes as well. So while the instrument is exogenous, the exclusion restriction does not hold with respect to any one observable action by the physician.

In other cases, such as Almond et al. (2010), this problem might be reasonably side-stepped by using hospital charges as a summary measure of treatment intensity; this would not be feasible in my setting. In Almond et al. (2010), crossing the "very low birthweight" threshold triggered a monotonic increase in treatment
intensity and hospital costs for newborns. In my paper, when the age-40 heuristic makes a physician more likely to expend resources pursuing a heart attack diagnosis, it might also make her less likely to expend resources considering an alternative diagnosis. This violates the monotonicity assumption that is required for a valid LATE estimate.

1.5 Results

Impact on Physician Decisions

Figure 1.1 presents the percentage of all ED patients that are tested for heart attacks. This, like all other regression discontinuity plots that follow it, is presented as a function of patient age on the ED visit date. Raw means are plotted in quarter-age bins, with a local-linear smoother fitted separately to the underlying data on either side of the age-40 threshold.

The figure shows an approximately linear relationship between testing and age on both sides of the threshold, with a sharp discontinuity at age 40. The slopes of these lines imply that the average increase in the probability of testing for each additional quarter of age is approximately 0.1 percentage points, while the increase in the probability of testing from the quarter before to the quarter after the patients’ 40th birthday is nearly an order of magnitude greater than this. Local linear estimation (as described in the previous section) yields an estimated 0.89 percentage point discontinuity in testing rates, which corresponds to a relative increase of 9.5% (Table 1.1).

Figure 1.2 shows a similar result with respect to the percentage of ED patients who are diagnosed with IHD: a fairly steep linear upward trend in the rate of
Figure 1.1: Proportion of ED patients tested for heart attack

Note: The dots in the above figure represent the mean heart attack testing rate among ED patients by quarter-age bin. The blue lines are local-linear smoothing functions fitted to the underlying month-age data. Source: Truven Health MarketScan Database, 2010 - 2013.
Table 1.1: Heart attack testing rates

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Tested for Heart Attack x100</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Female Male</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Turned40</td>
<td>0.887*** 0.998*** 0.752**</td>
<td>(0.150)</td>
<td>(0.219)</td>
</tr>
<tr>
<td>Intercept</td>
<td>9.328*** 8.404*** 10.463***</td>
<td>(0.119)</td>
<td>(0.164)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>9.5% 11.9% 7.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>622,783 343,577 279,206</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01. The coefficient on "Turned40" is the RD estimate. The "Relative Change" is the quotient of this coefficient and the intercept. See Equation (1.1). The dependent variable is binary, and all coefficients and standard errors are scaled by a multiple of 100 for ease of interpretation (e.g. a coefficient of 1.000 on "Turned40" represents a one percentage point discontinuity). Cluster-robust standard standard errors in parentheses. Source: Truven Health MarketScan Database, 2010 - 2013 for heart attack testing rates, 2005 - 2013 for all others.

diagnosis on either side of the threshold, with a sharp discontinuity at age 40. The local linear regression estimate of this discontinuity is approximately 0.13 percentage points, representing an even larger relative increase of 19.3% (Table 1.2). Likewise, Figure 1.3 and Table 1.3 show that the same pattern exists with respect to the proportion of ED patients that are admitted to the hospital with an IHD diagnosis.

Notable in Tables 1.1 through 1.3 is the disparity in the impact of the heuristic by patient sex. The baseline rate of testing and diagnosis for IHD in males is substantially higher than for females, which is to be expected, given the higher prevalence of IHD among males in this age group. However, the relative effects of crossing the age-40 threshold are substantially higher for women for each of the three measures: testing rates, IHD diagnosis rates, and IHD admission rates.
Figure 1.2: Proportion of ED patients diagnosed with IHD

Note: The dots in the above figure represent the mean IHD diagnosis (defined as visits including ICD-9 diagnosis codes 410-414) rate among ED patients by quarter-age bin. The blue lines are local-linear smoothing functions fitted to the underlying month-age data. Source: Truven Health MarketScan Database, 2005 - 2013.
Figure 1.3: Proportion of ED patients admitted to hospital with IHD diagnosis

Note: The dots in the above figure represent the mean IHD admission (defined as visits including ICD-9 diagnosis codes 410-414 that result in admission to the hospital) rate among ED patients by quarter-age bin. The blue lines are local-linear smoothing functions fitted to the underlying month-age data. Source: Truven Health MarketScan Database, 2005 - 2013.
Interestingly, a similar pattern is present with respect to whether the patient presented with chest pain, the most common symptom of IHD: the relative impact of the heuristic is nearly three times as large among patients presenting without chest pain as among patients presenting with it (Table 1.4). This suggests that less typical IHD patients are relatively more likely to have their treatment affected by the heuristic.

Table 1.2: IHD diagnosis rates

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Diagnosed with IHD x100</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Turned40</td>
<td>0.131***</td>
<td>0.087***</td>
<td>0.191***</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.021)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.679***</td>
<td>0.370***</td>
<td>1.061***</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.013)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>19.3%</td>
<td>23.4%</td>
<td>18%</td>
</tr>
<tr>
<td>Observations</td>
<td>1,325,771</td>
<td>735,228</td>
<td>590,543</td>
</tr>
</tbody>
</table>

Addressing Alternative Hypotheses

Throughout this paper I argue that the differential treatment patterns around the age-40 threshold are driven by physician reliance upon a heuristic. But one could reasonably posit that this pattern is driven by patient choice instead – that is, suppose that it is the patient and not the physician for whom the age-40 threshold is salient. If there exists some subset of the population that, upon experiencing distressing symptoms, is significantly more likely to visit an ED as a result of turning 40, a difference in mean testing rates might emerge without any difference
### Table 1.3: IHD admission rates

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Admitted with IHD x100</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All (1)</td>
<td>Female (2)</td>
<td>Male (3)</td>
</tr>
<tr>
<td>Turned40</td>
<td>0.068***</td>
<td>0.052**</td>
<td>0.091**</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.021)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.382***</td>
<td>0.188***</td>
<td>0.621***</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.013)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>17.9%</td>
<td>27.9%</td>
<td>14.6%</td>
</tr>
<tr>
<td>Observations</td>
<td>1,325,771</td>
<td>735,228</td>
<td>590,543</td>
</tr>
</tbody>
</table>

### Table 1.4: Heart attack testing rates, with and without chest pain (CP)

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Tested for Heart Attack x100</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With CP (1)</td>
<td>No CP (2)</td>
</tr>
<tr>
<td>Turned40</td>
<td>1.996**</td>
<td>0.414***</td>
</tr>
<tr>
<td></td>
<td>(0.817)</td>
<td>(0.097)</td>
</tr>
<tr>
<td>Intercept</td>
<td>49.089***</td>
<td>3.542***</td>
</tr>
<tr>
<td></td>
<td>(0.776)</td>
<td>(0.064)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>4.1%</td>
<td>11.7%</td>
</tr>
<tr>
<td>Observations</td>
<td>79,597</td>
<td>543,186</td>
</tr>
</tbody>
</table>
in physician behavior. But if this were the case, it would likely result in an observable shift in patient composition at age 40, generating discontinuities in the baseline prevalence of risk factors for IHD at this age. However, Table 1.5 demonstrates that patients’ relevant risk factors are well-balanced at the threshold, casting doubt on this hypothesis. Additionally, if such an influx of patients at age 40 were the primary driver of this pattern, one would expect to see a discontinuous increase in the number of patients arriving in the ED at age 40. Visual inspection of the distribution of ED visits by patient age (Figure 1.4) shows no such pattern. More formally, using the method developed in Frandsen (2017) to test for the presence of manipulation around a regression discontinuity cutoff when the running variable is discrete, no statistically significant difference across the threshold can be detected.\textsuperscript{6,7}

\begin{table}[h]
\centering
\begin{tabular}{lccccc}
\hline
Covariates & Below & Above & Diff & p & \\
\hline
Female & 0.552 & 0.555 & 0.003 & 0.110 & \\
Prior Diabetes DX & 0.048 & 0.047 & -0.000 & 0.809 & \\
Prior High Blood Pressure DX & 0.134 & 0.133 & -0.001 & 0.341 & \\
Prior High Cholesterol DX & 0.044 & 0.045 & 0.001 & 0.343 & \\
Prior Obesity DX & 0.045 & 0.044 & -0.001 & 0.121 & \\
Days Enrolled Before Visit & 535.652 & 533.891 & -1.761 & 0.511 & \\
\hline
\end{tabular}
\caption{Prevalence of relevant characteristics above and below the age-40 threshold}
\end{table}

Note: All differences and p-values in this table are computed using the same model and bandwidth used to estimate the discontinuities presented in the tables above. Source: Truven Health MarketScan Database, 2005 - 2013.

\textsuperscript{6}Even under the most stringent parameterization (i.e. \(k = 0\)) of the Frandsen (2017) test, \(p > 0.6\).

\textsuperscript{7}I thank Brigham Frandsen for making the code for this test readily available.
Figure 1.4: Number of ED visits, by month-age bin

One might also wonder whether the discontinuity in testing rates is driven by formal decision rules grounded in the medical literature or hospital policy. However, risk assessment algorithms that ED physicians might use to aid their clinical decisions such as HEART (Six et al., 2008) or TIMI (Antman et al., 2000) do not include any guidance for patients specifically around age 40, and such algorithms are generally intended to be referenced after initial tests are conducted. These diagnostic aides are therefore unlikely to be driving the observed behavior.

Given the inexpensive nature of first-line tests to detect heart attacks, it also seems unlikely that hospitals would adopt a formal policy to actively avoid testing patients under age 40. This is particularly improbable since younger patients are known to be at a higher than average risk for missed IHD diagnoses. Moreover, the number of heart attacks among younger patients is non-trivial: as many as 10% of heart attacks are experienced by individuals under the age of 45 (Hals and LoVecchio, 2009).

As noted previously, the discontinuous jump in the IHD diagnosis rate at age 40 is substantially larger than that found in the IHD testing rate (in relative terms). Any alternative hypothesis that predicts a jump in testing that is independent of physicians’ perception of patient risk is inconsistent with this fact. Patients on the IHD testing margin should on average be lower risk for IHD than the infra-marginal tested patients. Therefore, anything that causes an exogenous positive shock to physicians’ testing rates, but not their perceptions of patient risk, should instead result in a smaller relative discontinuity in the diagnosis rate.

---

8 A piece of anecdotal evidence that is consistent with my argument: no such policy exists at the Boston-area hospital from which I obtained records, yet I nonetheless find the effect of the age-40 heuristic on physician behavior to be strong there.
Impact on Subsequent Health Outcomes

While the large number of observations in this data allows for precise estimation of the impact of the age-40 heuristic on physician behavior, the number of patients affected by the heuristic – the "compliers" (Angrist et al., 1996) – relative to the number of all ED patients is small. Although these compliers make up approximately 10% of 40-year-old patients tested for a heart attack, they represent only about 1% of all ED visits. This poses issues for precise estimation of casual effects on broad subsequent outcomes of interest, such as hospital readmission or charges. One way to address this problem would be to exclude irrelevant portions of the population from the analysis to reduce the statistical noise in the outcomes. However, identifying such irrelevant portions of this population is particularly difficult: the patients who present in the ED without the most common characteristics of IHD are the ones who are most likely to be affected by the heuristic (as discussed earlier in this section). For instance, restricting the sample to patients presenting with chest pain (the most common symptom of IHD) excludes the majority of the population of interest (Table 1.4).

Because the relevant population cannot be readily identified ex-ante, the impacts of this heuristic on broad measures of health outcomes are likely to be swamped by statistical noise generated by the remaining 99% of the sample. I therefore focus instead on a subsequent health outcome that is specifically responsive to an ED physician’s decision to pursue an IHD diagnosis: subsequent heart attacks. Among all patients not diagnosed with IHD during their ED visit in my sample, the rate of heart attack diagnosis within 90 days following their visit is low – less than 1 in 1,000. However, conditional on being diagnosed with a heart attack in this time frame, it is very likely that the patient’s initial ED visit pertained to
undiagnosed IHD.\textsuperscript{9} The frequency of subsequent heart attack diagnoses among patients not diagnosed with IHD can therefore serve as a proxy for missed IHD diagnoses in the ED.

There are a few caveats to bear in mind with respect to this proxy. On one hand, it will inevitably include some subsequent heart attacks that were unrelated to the patient’s ED visit, thus overstating the number of likely missed diagnoses. This bias increases in the length of the observation period following the visit. On the other hand, the shorter this period is, the less likely it is to capture further contacts with health care providers, mechanically reducing the likelihood of observing a subsequent heart attack diagnosis. The net effect of these two forces is ambiguous, but given that at least one-third of heart attacks go undiagnosed altogether (Hals and LoVecchio, 2009), it is likely that this proxy is a conservative measure of missed diagnoses.

Figure 1.5 presents the rate at which ED patients (who were not diagnosed with IHD during their visit) are subsequently diagnosed with a heart attack within 90 days.\textsuperscript{10} Using the logic outlined above, a negative discontinuity in this rate at the age-40 threshold would indicate a reduction in the number of missed IHD diagnoses among those 40 years and older, a result of the jump in physicians’ diagnostic intensity at this threshold. Empirically, I find that while this rate does not differ significantly across the age-40 threshold among men, there is a large and statistically significant negative discontinuity among women (Table 1.6). Moreover, I find that this effect appears to be concentrated among women.

\textsuperscript{9}This assertion is supported by the data: roughly half of these patients had chest pain symptoms recorded during their ED visit, suggesting that their visit was indeed related to their subsequent heart attack diagnosis.

\textsuperscript{10}For robustness, I also conducted my analysis using periods of 30 and 180 days – the resulting estimates are consistent with those presented here (Appendix A.2).
presenting without chest pain (Figure 1.6).\textsuperscript{11} Using the same local linear regression framework described in Section 1.4, I measure the discontinuity in the number of patients diagnosed with a heart attack within 90 days of their ED visit. Among women presenting without chest pain, I estimate this discontinuity to be -21.3 per 100,000 ED visits, a relative decrease of 53.9\% (Table 1.7).\textsuperscript{12}

\begin{table}[h]
\centering
\small
\begin{tabular}{lccc}
\hline
\textbf{Dependent variable:} & \multicolumn{3}{c}{Missed IHD, 90 Days (per 100,000)} \\
 & All & Female & Male \\
\hline
\textit{Turned40} & -0.552 & -24.537*** & 29.977 \\
 & (9.206) & (7.199) & (22.908) \\
\textit{Intercept} & 82.345*** & 57.020*** & 113.809*** \\
 & (3.776) & (6.224) & (10.733) \\
\hline
\textbf{Relative Change} & -0.7\% & -43\% & 26.3\% \\
\textbf{Observations} & 1,315,741 & 732,085 & 583,656 \\
\hline
\end{tabular}
\caption{Subsequent heart attack diagnosis rates}
\end{table}

However, I qualify this point estimate with a caveat. Recall that patient age in this data is observed only at the month-age level. As discussed in Section 1.3 above, patients arriving in the ED in the month of their 40th birthday are therefore excluded from the analysis sample. As a result, my local linear regression estimates represent a "donut" RD design, extrapolating from the linear functions’ support to the threshold. When the underlying functional form of interest is linear, this practice yields unbiased estimates of the discontinuity. In my estimates of the effect of the age-40 heuristic on physician testing and diagnosis patterns, this

\textsuperscript{11}However, I have insufficient power to rule out a similar effect among women with chest pain.

\textsuperscript{12}To further demonstrate that this discontinuity is unlikely to be spurious, I also conducted a falsification test in which I estimated the discontinuity at every feasible month-age in the analysis sample ($p < 0.05$).
Figure 1.5: Subsequent heart attack diagnosis rate, by patient sex

(a) Female

(b) Male

Note: This figure displays the rate at which patients are subsequently diagnosed with a heart attack (ICD-9 diagnosis code 410) within 90 days of their initial ED visit, during which they were not diagnosed with IHD. As discussed in Section 1.5, I use this as a proxy for missed IHD diagnoses. The dots represent mean rates by quarter-age bin, while the blue lines are local-linear smoothing functions fitted to the underlying month-age data. Source: Truven Health MarketScan Database, 2005 - 2013.
Figure 1.6: Subsequent heart attack diagnosis rate, among female patients with and without chest pain

(a) With chest pain

(b) Without chest pain
Figure 1.7: *Subsequent heart attack diagnosis rate, among male patients with and without chest pain*

(a) *With chest pain*

(b) *Without chest pain*
Table 1.7: Subsequent heart attack diagnosis rates among female patients with and without chest pain (CP)

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Missed IHD, 90 Days (per 100,000)</td>
</tr>
<tr>
<td></td>
<td>With CP</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Turned40</td>
<td>-55.509</td>
</tr>
<tr>
<td></td>
<td>(64.529)</td>
</tr>
<tr>
<td>Intercept</td>
<td>200.837***</td>
</tr>
<tr>
<td></td>
<td>(40.541)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>-27.6%</td>
</tr>
<tr>
<td>Observations</td>
<td>79,264</td>
</tr>
</tbody>
</table>

Table 1.8: Subsequent heart attack diagnosis rates among male patients with and without chest pain (CP)

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Missed IHD, 90 Days (per 100,000)</td>
</tr>
<tr>
<td></td>
<td>With CP</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Turned40</td>
<td>153.889</td>
</tr>
<tr>
<td></td>
<td>(107.974)</td>
</tr>
<tr>
<td>Intercept</td>
<td>465.730***</td>
</tr>
<tr>
<td></td>
<td>(50.883)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>33%</td>
</tr>
<tr>
<td>Observations</td>
<td>70,805</td>
</tr>
</tbody>
</table>

seems to be a reasonable assumption, as the global trends (at least between age 35 and 45) in these measures are nearly linear and very precisely estimated. Visual inspection of Figure 1.6 suggests that the same cannot be said for the relationship between age and subsequent heart attack diagnoses to the right of the threshold. Moreover, due to the steep slope of the regression line at 40 years and one month,
linearly extrapolating to the age-40 threshold substantially increases the estimated magnitude of the discontinuity.

A more conservative estimate of this discontinuity that does not rely upon functional form assumptions can be obtained by comparing mean misdiagnosis rates above and below age 40 within a very narrow neighborhood around the threshold (Cattaneo et al., 2015). Given that IHD risk increases in age, the larger this neighborhood is, the more likely this estimate is to understate the true discontinuity. Using a neighborhood of 6 months, this yields an estimated drop in the subsequent heart attack rate among women without chest pain of 16.8 per 100,000 ED visits, or 42.5%. This estimate is statistically significant ($p < 0.01$), under both asymptotic and randomization inference procedures described in Cattaneo et al. (2015).

The reason that I observe this effect among female and not male patients is likely twofold. First, as discussed earlier in this section, the heuristic itself appears to generate a larger discontinuity in physicians’ behavior when treating female patients. Second, it is well-documented in the medical literature that women with IHD are under-diagnosed and under-treated relative to men with the same condition (Pope et al., 2000; Kim et al., 2016). Both of these facts suggest that women under 40 would be relatively more likely than men to benefit from the increase in diagnostic intensity that the heuristic generates at age 40. These results also imply that the marginal returns to additional diagnostic effort in screening women of this age for IHD are likely to be positive, an implication that I discuss in greater detail below.

---

Note that this estimate is not very sensitive to the choice of neighborhood: considering all candidate neighborhoods 1 through 6 months from the age-40 threshold yields point estimates ranging from -17.0 to -11.8, with a median of -15.2.
Are Women Under-Tested for IHD?

Given the estimated effect of the age-40 heuristic on subsequent heart attacks among women, it is straightforward to conduct a back-of-the-envelope cost-benefit analysis to assess whether women under the age of 40 are under-tested for IHD. Doing so first necessitates assigning a monetary value to the cost of testing for IHD, as well as to the benefits of preventing a heart attack. In Appendix A.3, I estimate the average cost of testing this population for IHD to be approximately $250 per patient tested, or $250,000 per 100,000 ED visits.\footnote{Recall that in Table 1.1, I estimate that an additional 1\% of women visiting the ED are tested for IHD upon turning 40. Therefore, for 100,000 ED visits, only 1,000 additional patients are tested, at a total cost of $250,000 per 100,000 ED visits.}

On the other hand, identifying patients with previously undiagnosed IHD has potentially large benefits: patients can be placed on an inexpensive prophylactic treatment regimen which greatly reduces heart attack risk. Roughly 15\% of heart attacks are fatal Mozaffarian et al. (2016), while Mahoney et al. (2002) estimate that among patients with IHD, a non-fatal heart attack reduces life expectancy by about one-eighth. If patients are assumed to have a relatively normal lifespan (e.g. roughly 80 years) if their IHD is diagnosed and treated, the average heart attack for someone around 40 years of age results in a loss of approximately 11 life-years.\footnote{15\% \times (80 - 40) + \frac{1}{8} \times (80 - 40) = 11 \text{ life-years.}} Using the suggested benchmark valuation of $100,000 per life-year (Neumann et al., 2014), this implies that the value of preventing a heart attack for a patient of this age is roughly $1.1 million.\footnote{This likely to be a conservative valuation, as it implies that that a fatal heart attack at age 40 is valued at $4 million, which is less than half of the value of a statistical life used in US Department of Health and Human Services regulatory impact analyses (HHS, 2016).} Based on my more conservative estimate of the effect of the age-40 heuristic on subsequent heart attacks among
women, this implies a benefit of $18.48 million per 100,000 female ED patients.\textsuperscript{17} This is almost two orders of magnitude greater than the estimated costs of testing this marginal population for IHD. Thus it appears that under any reasonable set of assumptions, women at this age are under-tested for IHD.

**Considering the Role of Physician Cognitive Load and Fatigue**

Most leading hypotheses attempting to address *when* individuals employ heuristics invoke the notion of bounded rationality (Shah and Oppenheimer, 2008). In many settings, decision-makers have insufficient time or cognitive resources to make choices via exhaustive search or formal probabilistic reasoning, forcing them to rely upon heuristics instead. Kahneman (2011) discusses the tendency of individuals to switch from deliberative reasoning ("System 2") to heuristic thinking ("System 1") when available cognitive resources are low – that is, during or following particularly cognitively taxing tasks. This begs the question: is the age-40 heuristic only relevant to decision-making during especially demanding periods in the ED?

To explore this hypothesis, I analyze detailed records from the ED of a large Boston-area hospital, as described in Section 1.3. Using the ED arrival timestamps recorded for each patient, I am able to generate proxies for physician cognitive load (as measured by ED patient volume) and physician fatigue (as measured by the order in which patients arrive during the physician’s shift). Since more physicians are scheduled to work during the typically busy times of the day, it may not be the case that the burden on any individual physician is actually higher during periods of peak patient flow. Therefore, I define an hour in which an above-median (for that hour of the day) number of patients arrive as a "high volume" period, and

\textsuperscript{17}16.8 \times $1.1\ million = $18.48\ million
"low volume" otherwise. This measure of patient volume thus reflects deviations from typical patient flows for any given time of the day. Then, after merging the attending physicians’ shift schedules to these patient-level records, I can also determine the order in which patients were assigned to attending physicians within a given shift.\(^\text{18}\)

Using these records, I am able to replicate the pattern found in the Truven claims data. Employing the same regression discontinuity methodology presented above, I find a large and statistically significant jump in the heart attack testing rate at age 40 (Figure 1.8). The point estimate of this discontinuity is substantially larger than in the claims data (although less precisely estimated), implying an 51.7% increase in the testing rate at the threshold (Table 1.9, column 1). In order to examine whether use of the age-40 heuristic is primarily driven by periods in which cognitive resources are depleted, I re-estimate the discontinuity separately among patients arriving in the ED during high vs. low-volume hours (Figure 1.9, Table 1.9, columns 2 & 3), and again for patients arriving during the second vs. first half of their physician’s shift (Figure 1.10, Table 1.9, columns 4 & 5). In each case, a large discontinuity in testing rates is present at age 40.

Counter-intuitively, these point estimates turn out to be *larger* when patient volume is lower and when patients are seen during the first half of the physician’s shift; these estimates are statistically significant (*p* < .05).\(^\text{19}\) Though noisy, these results demonstrate that the discontinuity in testing rates at age 40 appears to persist even in settings in which available cognitive resources are presumably

\(^{18}\) However, the shift schedule that I obtained does not cover the entire sample; roughly 75% of patient visits could be linked directly to scheduled shifts.

\(^{19}\) However, the limited number of observations in these ED records preclude me from finding a statistically significant difference between these two sets of point estimates.
Figure 1.8: Proportion of patients tested for heart attack in Boston-area ED

Note: The dots in the above figure represent the mean heart attack testing rate among ED patients by quarter-age bin. The blue lines are local-linear smoothing functions fitted to the underlying data. Source: Emergency department records from a Boston-area hospital, January 2010 - May 2015.
Figure 1.9: Proportion of patients tested for heart attack in Boston-area ED, by patient volume

Note: The solid blue line (dashed red line) is a local linear fit among patients arriving during an hour in which there were an above-median (below-median) number of patients arriving in the ED, conditional on the hour of day. The blue o’s (red x’s) represent the mean testing rate among patients within each quarter-age bin in the above-median (below-median) group. Source: Emergency department records from a Boston-area hospital, January 2010 - May 2015.
Figure 1.10: Proportion of patients tested for heart attack in Boston-area ED, related to attending physician’s shift

Note: The solid blue line (dashed red line) is a local linear fit among patients treated during the second (first) half of an attending physician’s shift. The blue o’s (red x’s) represent the mean testing rate among patients within each quarter-age bin in second half (first half) group. Source: Emergency department records from a Boston-area hospital, January 2010 - May 2015.
Table 1.9: Heart attack testing rates, Boston-area ED

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Tested for Heart Attack x100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td>Turned40</td>
<td>4.033** (1.607)</td>
</tr>
<tr>
<td>Intercept</td>
<td>7.807*** (1.021)</td>
</tr>
</tbody>
</table>

Relative Change 51.7% 30.3% 70.5% 24.2% 76.8%
Observations 6,271 2,654 3,617 2,299 2,338

Note: *p<0.1; **p<0.05; ***p<0.01. The coefficient on "Turned40" is the RD estimate. The "Relative Change" is the quotient of this coefficient and the intercept. See Equation (1.1). The dependent variable is binary, and all coefficients and standard errors are scaled by a multiple of 100 for ease of interpretation (e.g. a coefficient of 1.000 on "Turned40" represents a one percentage point discontinuity). Heteroscedasticity-robust standard errors in parentheses. Optimal bandwidth of 1.52 years selected by procedure described in Section 1.4. Source: Emergency department records from a Boston-area hospital, January 2010 - May 2015.

higher than average. Thus it seems that reliance upon this heuristic is not merely limited to unusually taxing conditions, but is instead a more integral part of physicians’ diagnostic process in the ED.

Why Age 40?

It seems intuitive that a physician might informally categorize patients into coarse age bins to simplify diagnostic decisions involving patient age. However, it is not obvious why 40 years of age might be a particularly salient number for establishing a cutoff between "young" and "old" patients for the purpose of diagnosing IHD. Literature in both psychology and economics demonstrates that individuals seem
to have a preference for round numbers (Lynn et al., 2013), and in some cases appear to anchor their decisions to them (Lacetera et al., 2012; Pope and Simonsohn, 2011), so perhaps it is not surprising that the cutoff age is round.

Among integers ending in a "0," a few objective criteria seem to suggest that age 40 might be a reasonable choice. First, 40 years is approximately the midway point of the average lifespan in the US, making it a relatively natural (if naive) point at which to bifurcate age. Alternatively, suppose that the physician’s choice of threshold is driven by the relationship between age and IHD risk. Figure 1.11 shows the US annual per-capita heart attack mortality rate as a smooth, continuous function of age. If such a non-linear relationship is too complex to fully internalize, it could be simply approximated as a piece-wise linear function. Using the same data, the MSE-minimizing split point for such a function is approximately 40 years (Figure 1.12). Moreover, age 40 is roughly the inflection point in the heart attack testing rate curve (Figure 1.13), further supporting the notion of this age being a pivotal point with respect to testing decisions.

1.6 Theoretical Framework

The Role of Representativeness in Clinical Decision-Making

The process by which a physician learns to practice medicine is through an on-the-job training environment. Medical students and residents observe more senior physicians in action, and begin making diagnoses and performing procedures under their supervision. It is through this process that they learn what it looks like for a patient to be suffering from a heart attack, angina, acid reflux, or pneumonia. In economics, a common way to model the diagnostic process is to assume that
Figure 1.11: Annual heart attack deaths per capita

Note: Each dot represents the per-capita acute myocardial infarction (AMI, colloquially known as a "heart attack") death rate for all individuals in the US at a given year of age. The blue line is a local-linear smoothing function fitted to these points. Source: 2010 - 2015 CDC NVSS; 2010 - 2015 Census Population Estimates.
Figure 1.12: Piecewise linear fit - annual heart attack deaths per capita

Note: Each dot represents the per-capita acute myocardial infarction (AMI, colloquially known as a "heart attack") death rate for all individuals in the US at a given year of age. The blue lines represent the MSE-minimizing piecewise linear fit for these points. Source: 2010 - 2015 CDC NVSS; 2010 - 2015 Census Population Estimates.
Figure 1.13: Proportion of ED patients tested for heart attack

Note: The dots in the above figure represent the mean heart attack testing rate among ED patients by quarter-age bin. The blue lines are local-linear smoothing functions fitted to the underlying month-age data. Source: Truven Health MarketScan Database, 2010 - 2013.
the physician is a Bayesian agent. This implies that the physician arrives at a diagnosis by computing the posterior probability that a patient suffers from a medical condition given a set of observed characteristics. In reality, despite having completed a long and arduous education both in and outside of the classroom, physicians (like most people) struggle with the estimation of conditional probabilities even in situations that are familiar to them (Casscells et al., 1978; Eddy, 1982; Manrai et al., 2014). Yet on the whole, they are reasonably effective diagnosticians; how do we reconcile these two facts? I posit that the problem they are solving is not governed directly by objective probability, but by a form of reasoning that is nonetheless correlated with it: representativeness.

Kahneman and Tversky (1972) define representativeness as a commonly used heuristic in which a person judges the subjective probability of an event by the extent to which it is "(i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated." I argue that this form of reasoning seems particularly relevant in the context of emergency department diagnoses, where algorithmic approaches to medicine are commonplace, and often amount to a matching exercise: How closely does a patient’s characteristics resemble the prototypical presentation of a particular condition? If "very close," make this diagnosis. If not close at all, rule out this diagnosis. If "close enough," collect additional information and then re-evaluate.

To formalize this process, I propose a representativeness-based model of diagnostic decision-making which bears similarity to the "drift diffusion" model of Krajbich et al. (2014). Patient $i$ with vector of characteristics $X_i$ arrives in the emergency department, and is considered for a diagnosis of medical condition $c$. $X^c$ represents the prototypical presentation of characteristics consistent with condition $c$. The subjective risk of patient $i$ for condition $c$ is inversely related to
the distance $d$ between $X_i$ and $X^c$:

$$R_i^c \equiv 1 - \frac{d_c(X_i, X^c)}{d_c^{\text{MAX} c}}; \quad R_i^c \in [0, 1] \quad (1.2)$$

Distance $d$ is normalized by $d_c^{\text{MAX} c}$ so that a patient perfectly matching the prototypical presentation of condition $c$ has a risk of 1, while a patient bearing no resemblance to the prototypical presentation has a risk of 0. The physician pursues a diagnosis of condition $c$ if the following are true:

1. $R_i^c < R_i^c < \overline{R}^c$, and

2. $R_i^c > \frac{K_t}{B_t^c}$

where $K_t$ is the cost of test $t$, and $B_t^c$ is the benefit of running the test on a patient with $c$. That is, diagnostic tests are conducted (the results of which are incorporated into patient characteristics $X_i$) until subjective risk $R_i^c$ falls below $\overline{R}^c$, at which point the diagnosis is ruled out, or rises above $\overline{R}^c$, at which point a diagnosis of condition $c$ is made, or until the cost of any remaining test is greater than its expected benefits.

Note that in equation (1.2) I make no restrictions on the functional form of the distance function $d$, thereby allowing for the possibility that physicians may misweight the relevance of characteristics in determining the patient’s risk. In what ways might physicians systematically (mis)weight the relative importance of patient characteristics when making a diagnosis? Consider the following example. In Casscells et al. (1978), physicians employed at Harvard Medical School teaching hospitals were asked this question:

"If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person’s symptoms or signs?"
Applying Bayes’ theorem yields a conditional probability of less than 2%, yet the modal response from those surveyed was 95%\(^{20}\), a finding that has been replicated by others, including recently in Manrai et al. (2014). As Eddy (1982) points out, this pattern is consistent with physicians inadvertently substituting the attribute of interest \(P(disease|positive)\) with the more easily recalled \(P(positive|disease)\).\(^{21}\) This suggests that when physicians are presented with a characteristic \(X_j\) from which to make a diagnosis, they tend to draw upon the likelihood that someone with the condition possesses that characteristic \(P(X_j|c)\) rather than the posterior probability that the patient has condition \(c\) given the characteristic \(P(c|X_j)\). Note how this tendency is consistent with the Kahneman and Tversky (1972) definition of representativeness quoted above.

This proclivity also fits naturally with the research on stereotypes found in Bordalo et al. (2016a,b), which formally model representativeness in terms of these likelihoods. Following their model and adapting it to the specific question at hand, I define the representativeness of a binary characteristic \(X_j\) for a patient with condition \(c\) relative to a patient without \(c\) to be:

\[
\pi_j^c = \frac{P(X_j = 1|c)}{P(X_j = 1|\bar{c})}.
\] (1.3)

I call \(X_j\) a "representative characteristic" for condition \(c\) if \(\pi_j^c > 1\) (that is, when \(X_j\) is more prevalent among those with \(c\) than among those without \(c\)). It is a set of these characteristics that comprise \(X^c\), the prototypical presentation of condition \(c\).

\(^{20}\)The mean estimate given by the respondents was 55.9%.

\(^{21}\)\(P(positive|disease) = 0.95\) assuming that the test’s false negative rate is also 5%.
Implications of Representativeness for Missed Diagnoses

A fundamental attribute of representativeness is that the decision weights placed on representative characteristics are exaggerated relative to their actual predictive power (Bordalo et al., 2016a,b). This feature of representative thinking is plausibly responsible for the heterogeneity in the empirical results discussed in Section 1.5. Recall that the discontinuous drop in missed IHD diagnoses upon crossing the age-40 threshold is concentrated among females presenting in the ED without chest pain, indicating that this group is under-diagnosed before age 40. The attributes that define this group (female and without chest pain) are simply the absence of two representative characteristics for IHD, which would lead physicians to underestimate this group’s risk. This decreases the intensity with which they are screened for IHD, thereby increasing the likelihood that those with IHD will be misdiagnosed.

It is also important to keep in mind that representative characteristics are not universal, but context-dependent – what is representative for one condition is not necessarily representative for another. For example, consider the results of Abaluck et al. (2016), which demonstrate that substantial welfare losses are generated by physicians’ misweighting of patient characteristics when making testing decisions for another potentially fatal acute condition, pulmonary embolism (PE). Yet patient sex is not one of these misweighted characteristics: the authors find that testing for PE appears to be reasonably balanced across males and females according to their relative risks. But this is in fact consistent with my model, as unlike the male-skewed incidence of heart attacks around age 40, the prevalence of PE is relatively similar among males and females (Stein et al., 1999), implying that being male is not a representative characteristic for PE.
1.7 Policy Remedies

Given the importance of clinical decision-making in health care, there are potentially large welfare gains to be made by implementing policies to address the types of systematic biases discussed in this paper. When policymakers are confronted with the errors in physicians’ probabilistic judgment as illustrated in Casscells et al. (1978) (as described in Section 1.6 above), a natural response is to call for additional training in statistics during medical school. Certainly, greater statistical sophistication is a worthy goal, as it might not only improve clinical decisions directly, but also physicians’ ability to interpret important findings in the medical literature. However, it is unreasonable to assume that enhancing medical education in this manner is a panacea, as this solution fails to address three key issues.

First, it is not clear that further education would substantially reduce the types of biases described in this paper. Indeed, Manrai et al. (2014) suggests that physician behavior has changed little in this respect in the decades since Casscells et al. (1978), despite changes made to medical education curriculum in the interim. This should not come as a surprise – these biases are the result of heuristics that are endemic to human judgment under uncertainty more generally, not because individuals do not possess the ability to compute probabilities.

Second, it is conceivable that an attempt to compel physicians to behave in a more Bayesian manner could actually result in a deterioration rather than an improvement in the quality of care. Heuristics are employed to quickly solve problems while conserving cognitive effort, and in most cases, they result in near-optimal choices. By forcing physicians to expend more effort estimating probabilities, less effort may be expended on other aspects of their jobs, a trade-off that may very well result in a welfare loss.
Third, even if improvements in this area can be made through training, any effort to improve physician behavior through medical education will likely take decades to come to fruition, as after physicians complete their medical training, many fail to stay abreast of recent developments. In the words of a prominent cardiologist, "it’s a job, and they’re trying to earn money, and they don’t necessarily keep up. So really major changes have to be generational" (Epstein, 2017). Thus, a new generation of physicians must be trained before they can join the profession and eventually influence common practices.

Implicit in arguments for more training is the desire to somehow make physicians more like computers. But why expend great effort training individuals to think in a way that is unnatural to them when machines can perform these tasks at exceptionally low cost? It is likely more productive to encourage physicians to focus on improving high-value aspects of their jobs that machines cannot readily handle, such as eliciting and interpreting the patient’s symptoms and patient history, conducting careful physical exams, noticing subtle details about the patient that are not easily captured in a patient chart, or improving bedside manner.

To the extent that research can identify scenarios in which physicians’ heuristics tend to lead them astray, electronic decision support systems (DSS) can be developed to "nudge" (Thaler and Sunstein, 2009) them away from undesirable actions. Clearly, the benefits to these systems will depend heavily on successful design and implementation. Such efforts should be guided by a principle of minimal intrusion in the diagnostic process – the decision-maker should only be nudged in the limited settings in which biased behavior is prevalent and meaningful, thereby minimizing DSS "alert fatigue" (Kesselheim et al., 2011) and maintaining the efficiency that heuristic thinking typically offers. Just as Kleinberg et al. (2017) demonstrates the potential to improve decisions in the judicial system by harness-
ing machine learning prediction techniques, a similar approach could be taken in the field of medicine.

### 1.8 Conclusion

Most of the evidence regarding heuristics and biases in human decisions has been demonstrated among relatively unsophisticated or inexperienced agents operating in low-stakes experimental settings. In this paper, I provide evidence that such behavior is in fact present even among extensively trained professionals facing common dilemmas in a high-stakes environment. If heuristic behavior is prevalent in this setting, it is likely to have meaningful impacts in many domains, making this an important area of research going forward.

Empirically, I find that physicians in the ED are roughly 10% more likely to test and 20% more likely to diagnose patients with IHD just after the patients’ 40th birthday than just beforehand, despite the fact that the underlying incidence of IHD increases smoothly with age. This shock to diagnostic intensity appears to substantially reduce the likelihood of missed diagnoses among women, a population that is known to be at elevated risk of misdiagnosis. This suggests that the marginal returns to screening for IHD in female ED patients at this age are relatively large. I also demonstrate a more general pattern in the misdiagnosis of IHD that is also consistent with heuristic thinking: patients presenting without characteristics that are representative of IHD appear more likely to be misdiagnosed than can be justified by their objective risk.

It is clear that the quality of diagnostic decisions in the ED can have large impacts on patient outcomes. And from a policy perspective, the role of EDs in medical spending has expanded greatly in recent years, as the share of hospital
admissions that originate in the ED has risen sharply in recent decades (Schuur and Venkatesh, 2012; Kocher et al., 2013), making emergency physicians increasingly important gatekeepers for expensive inpatient care. However, the solution to the issues described in this paper is not to somehow purge decision-makers of heuristic behavior. Decision-makers have good reason to think heuristically – in general, it serves them well, efficiently generating near-optimal conclusions with minimal cognitive effort. Instead, effort should be made to accommodate heuristic thinking, and only implement safeguards in settings where it is likely to lead decision-makers astray. As technological advances in data quality and computation continue, machine-learning algorithms provide the opportunity to systematically identify these settings.
Chapter 2

Worth the Price of Admission?
Evidence from Emergency Department Admissions

2.1 Introduction

Aggregate health care spending in the US has grown to more than $3.3 trillion, or roughly 18% of GDP (Hartman et al., 2017), well above that of other OECD countries whose citizens enjoy better health (Anderson and Frogner, 2008). Inpatient hospital stays are responsible for roughly one-third of these costs (Weiss et al., 2014) despite the fact that they comprise only 3% of all health care visits (Torio and Moore, 2016), making them a particularly expensive form of patient care. In recent decades, the source of these costly inpatient stays has shifted: roughly half now originate in hospital emergency departments (EDs), up from one-third in the 1990s. This trend has made emergency physicians increasingly important gatekeepers for
expensive, high-intensity inpatient care (Schuur and Venkatesh, 2012; Morganti et al., 2013).

Yet despite the high costs of hospital admissions from EDs, we know little about their benefits. This is partly because it is difficult to generate estimates of their causal effects on patient outcomes using data that is widely available to researchers. For instance, there is substantial variation across EDs in the rate at which patients are admitted to the hospital (Pines et al., 2013). However, patient characteristics, both observable and unobservable to the researcher, can vary substantially across hospitals, so attempting to estimate the causal effect of admission by comparing patient outcomes for high- and low-admitting EDs would likely yield biased results. Yet within any given ED, patient assignment to individual physicians is often quasi-random, as patients cannot choose their physicians, and EDs are legally obligated to treat all patients (Chan, 2016). Since the practice styles of physicians can vary widely even within a given emergency department (Van Parys, 2016), this variation combined with the quasi-random assignment of patients to physicians can be exploited to recover the causal effects of their choices on their patients.

This framework has been utilized frequently in research on physician practice styles in EDs. Recently, Van Parys (2016) demonstrated the substantial variation in ED spending across physicians treating patients with minor injuries, and notes that physician characteristics appear to explain little of this variation; Barnett et al. (2017) highlighted the impact of emergency physician prescribing tendencies on patients’ long-term opioid use; and Gowrisankaran et al. (2017) explored the relationships between physicians’ resource use within the emergency department, diagnostic skill, and patient outcomes.

The primary contributions of this paper are two-fold. First, using records from a large Boston-area hospital pertaining to patients presenting in the ED with
chest pain or shortness of breath, I demonstrate that there is substantial within-institution heterogeneity in the rate at which emergency physicians admit patients. I find that physicians who admit their patients at an above-median rate are at least 20% more likely to admit a given patient than those who admit at a below-median rate, despite the fact that on average, their patients possess similar characteristics. Second, I show that patients quasi-randomly assigned to physicians who admit at a high rate fare no better (with respect to subsequent ED visits, hospital admissions, or mortality) than those assigned to physicians who admit at a low rate, and in some cases fare worse. This suggests that the additional admissions generated by high-admitting physicians are likely of low value, and possibly harmful.

The remainder of this paper is organized as follows. Section 2.2 provides a discussion of the nature of the admission decision in the ED, particularly with respect to patients presenting with chest pain or shortness of breath. Section 2.3 describes the data used in my analysis. Section 2.4 discusses the empirical framework used to estimate the causal effects of patients being assigned to high-versus low-admitting physicians in the ED. Section 2.5 presents the results of my analysis. Section 2.6 discusses the policy implications of these results, and Section 2.7 concludes.

2.2 The Admission Decision

In most fields of medicine, physicians have a relatively long-term relationship with their patients, allowing them to establish a comprehensive medical history. Emergency physicians, on the other hand, typically have no prior interactions with their patients, and their knowledge of the patient’s medical history is limited to whatever the patient tells them (Groopman and Prichard, 2007). The initial
diagnostic information at their disposal is typical scarce, limited to the patient’s vital signs, observable demographic characteristics, and symptoms.

Two of the most common symptoms that bring patients to EDs in the US are chest pain and shortness of breath (Rui et al., 2013). Emergency physicians therefore evaluate patients with these complaints on a daily basis. Like most cases handled in the emergency department, the majority of these cases are not attributable to an imminently life-threatening condition – most are relatively benign in nature. At the same time, these two symptoms are also associated with ischemic heart disease (IHD), a serious yet relatively common set of conditions that includes heart attacks. Given that a missed IHD diagnosis in the ED can have dire consequences for the patient, these cases must be taken seriously, particularly among middle-aged and elderly adults (Pope et al., 2000). Emergency physicians are therefore tasked with assessing the condition of these patients through physical examination and performing diagnostic tests, and then stratifying patients by risk. In practice, patients typically fall into one of three categories:

1) **Clearly low-risk cases.** The cause of the complaints has been determined to be non life-threatening. These patients can be treated and safely discharged.

2) **Clearly high-risk cases.** These patients require immediate and intensive medical care. These patients are admitted to the hospital.

3) **Intermediate cases.** Diagnostic tests do not provide sufficient evidence to assign the patient to categories 1) or 2).

Patients in category 3) above fall into the "gray area" of emergency medicine (Chandra et al., 2015), in which the disposition of the patient is driven in large part by the attending physician’s discretion. I posit that it is primarily these cases
that allow for the substantial variation in practice style observed across emergency physicians, a hypothesis supported by my analysis in Section 2.5. Given the high cost of admitting patients to the hospital, it is worthwhile to determine whether these "gray area" patients fare sufficiently better when being treated by a physician who is more or less likely to admit them.

2.3 Data

The empirical analysis in this paper is performed using data obtained from the ED of a large Boston-area hospital. The level of detail contained in this data is substantially greater than that of commonly used research databases, and is ultimately what enables me to carry out this analysis. Hospital data (such as that provided by the Healthcare Cost and Utilization Project (HCUP)) typically record inpatient and outpatient hospital visits in separate datasets. Therefore, if a patient’s ED visit results in admission, the visit will appear in the inpatient database, whereas if the patient is discharged from the ED, the visit will appear in the outpatient dataset. Unfortunately, the inpatient database only identifies the inpatient attending physician, rather than the admitting ED physician, making it impossible to determine physician-specific admission rates (Van Parys, 2016). However, the hospital records used in this paper identify the attending emergency physician for each ED visit, regardless of whether the patient is admitted or discharged.

These records pertain to all patients presenting in the ED with a chief complaint of chest pain or shortness of breath between January 2010 and May 2015, and are recorded at the patient-visit level. This data contains basic demographic characteristics of the patient (age, race, sex), the patient’s chief complaint, the date
and time of the arrival at and disposition from the ED, the type of disposition (discharge, admission, or held for observation), unique identifiers for both the patient and attending physician, as well any diagnoses or procedures performed in the emergency department.

I impose a number of sample restrictions on this data prior to conducting any analysis: first, if patients have more than one visit to this ED with complaints of chest pain or shortness of breath, only their first visit is included in the analysis sample; second, any visits lacking demographic information or timestamps are excluded; third, visits assigned to attending physicians with less than 100 visits during this period are excluded, as it would be difficult to precisely estimate these physicians’ admission behavior. After these exclusions, the sample consists of 19,005 ED visits.

**Table 2.1: Patient Characteristics**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.560</td>
</tr>
<tr>
<td>Age</td>
<td>54.166</td>
</tr>
<tr>
<td>White</td>
<td>0.532</td>
</tr>
<tr>
<td>Black</td>
<td>0.216</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.169</td>
</tr>
<tr>
<td>Private Insurance</td>
<td>0.441</td>
</tr>
<tr>
<td>Prior Hypertension DX</td>
<td>0.472</td>
</tr>
<tr>
<td>Prior Diabetes DX</td>
<td>0.226</td>
</tr>
<tr>
<td>One-Year Mortality</td>
<td>0.103</td>
</tr>
<tr>
<td>Held for Observation</td>
<td>0.198</td>
</tr>
<tr>
<td>Admitted</td>
<td>0.378</td>
</tr>
<tr>
<td>AMI DX</td>
<td>0.027</td>
</tr>
<tr>
<td>IHD DX</td>
<td>0.202</td>
</tr>
</tbody>
</table>

N = 19,005

1Only approximately 5% of all patients were assigned to these low-volume physicians.
Table 2.1 presents summary statistics relating to this sample, showing that on average, members of this population are roughly 54 years of age and in fairly poor health, with a one-year mortality rate (as determined by Social Security death records) of roughly 10%. Many of these patients suffer from conditions placing them at elevated risk of IHD: nearly one-half had a prior diagnosis of hypertension, and almost one-quarter had a prior diagnosis of diabetes. While less than 3% are diagnosed with a heart attack (AMI) during their visit, over 20% are diagnosed with some form of IHD, which puts them at risk of heart attacks in the future. More than one-third of these patients are admitted to the hospital as a result of their ED visit, and almost one-fifth are held for observation in the ED prior to being discharged.

2.4 Empirical Framework

Identifying High-Admitting Physicians

There are a number of ways in which to model the effects of attending physicians’ propensity to admit their patients. In this paper, my analysis is guided by the following thought experiment:

Every patient arriving in the ED will be seen by an attending physician who admits patients at an above-average rate, or one who admits patients at a below-average rate. How much does a patient’s treatment and subsequent outcomes depend on which of these two physicians are assigned to her?

In order to proceed in addressing this question, I must first define what it means for a physician to have an "above-average" admission rate. When making this

---

2 As prior diagnoses could only be determined for individuals previously diagnosed within the health care system with which this hospital is affiliated, these rates likely understate the disease prevalence in this population.
determination, I control for the fact that the timing of physicians’ shifts in the ED is unlikely to be completely randomly assigned, while the types of patients visiting the ED may vary from shift to shift. I also control for the fact that differences in admission rates across physicians in part reflect average differences in their patients’ characteristics.\(^3\).

In order to characterize the attending physicians in this ED according to their propensity to admit their patients to the hospital, I estimate the following equation:

$$Admitted_{ij} = \hat{a}_0 + T_{ij}\hat{\delta} + P_{ij}\hat{\gamma} + \hat{u}_{ij}$$ \hspace{1cm} (2.1)

where patient \(i\) is assigned to attending physician \(j\). \(Admitted\) is a binary variable indicating whether patient \(i\) was admitted from the ED to this hospital. \(T_{ij}\) represents a vector of time fixed-effects, including hour of day, day of week, month, year, and month-by-year interactions for each visit. \(P_{ij}\) represents a vector of variables including a quadratic polynomial of the patient’s age, sex, race, chief complaint and condition severity.\(^4\) The residual \(\hat{u}_{ij}\) therefore represents the part of the admission decision that is unexplained by the patient’s demographic characteristics or the timing of the visit. Then, I compute the average residual for each physician \(j\):

$$\bar{u}_j = \frac{1}{N_j} \sum_{i \in I_j} \hat{u}_{ij}$$ \hspace{1cm} (2.2)

\(^3\)This variation in patient characteristics across physicians need not be driven by physicians selecting their patients – this will also arise from random chance

\(^4\)Condition severity is assessed by a triage nurse prior to physician assignment, and is described in Section 2.5.
Next, I define "high admitting" physicians as those whose average residual ranks in the top half of all physicians:

\[
HighAdmitting_j = \begin{cases} 
1 & \bar{\bar{u}}_j \geq \bar{\bar{u}}_{\text{Median}} \\ 
0 & \bar{\bar{u}}_j < \bar{\bar{u}}_{\text{Median}} 
\end{cases} \quad (2.3)
\]

The relationship between this indicator and patient treatment and outcomes will be the focal point of the analysis.

**Estimating Causal Effects on Patient Treatment and Outcomes**

Having categorized all physicians according to their admission propensity, I estimate the causal effect of a patient being assigned a high-admitting type rather than a low-admitting type via the following equation:

\[
Y_{ij} = \hat{\beta}_0 + \hat{\beta}_1 HighAdmitting_{ij} + T_{ij}\delta + \epsilon_{ij} \quad (2.4)
\]

where \(Y_{ij}\) represents a treatment (e.g. admission to the hospital) or patient outcome of interest (e.g. subsequent readmission or mortality rates), and \(\hat{\beta}_1\) is the estimated causal effect of interest. Note that this estimate is unbiased only if, conditional on the timing of the patient’s arrival in the ED, the physician’s type is uncorrelated with the characteristics of the patient – this is necessarily true if patients are in fact randomly assigned to their physicians in the ED.\(^5\) To the extent to which higher (lower) risk patients are assigned to high-admitting physicians, \(\hat{\beta}_1\) will underestimate (overestimate) any health benefits of being assigned to a high-admitting physician.

---

\(^5\)In Section 2.5, I perform tests assessing the similarity of patients assigned to high- and low-admitting physicians.
2.5 Results

Is Assignment of Physicians to Patients Completely Random?

When patients arrive in this ED, they are first examined by a triage nurse who elicits their chief complaint, preliminarily assesses the severity of their condition on a 1 to 5 scale,\(^6\) and assigns them to a bed. After being assigned to a bed, an available attending physician must then accept responsibility for treating the patient. Therefore, while the physicians have no control over which patients arrive in the ED, there is some room for non-random assignment if there are multiple attending physicians working at a given point in time (Chan, 2015). This has clear implications for the extent to which the estimates in this paper can be interpreted as causal. However, the only patient characteristics that are known to the physician prior to accepting the patient for treatment are the patient’s age, sex, last name, chief complaint, and severity. To the extent that physicians are able to select patients based on their characteristics, any such selection should be limited only to these observables.

Table 2.2 displays the mean baseline characteristics of patients assigned to high- and low-admitting physicians (controlling for arrival date/time fixed-effects), as well as the p-value under the null hypothesis that the mean is the same in both groups. These results show that patients assigned to high-admitting physicians are two percentage points more likely to be white, and one and one-half percentage points more likely to be male than those assigned to low-admitting physicians. Although relatively small, these differences are statistically significant at the 5% level, implying that patient assignment is unlikely to be completely random.

---

\(^6\) The severity of the patient’s condition is determined in accordance with the widely-used Emergency Severity Index (ESI) triage algorithm (Tanabe et al., 2004).
However, the average differences between these two groups of patients for all other baseline characteristics are quantitatively and statistically insignificant.

**Table 2.2: Characteristics of patients assigned to high- and low-admitting physicians**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Low</th>
<th>High</th>
<th>Diff</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.567</td>
<td>0.552</td>
<td>-0.015</td>
<td>0.042</td>
</tr>
<tr>
<td>Age</td>
<td>54.054</td>
<td>54.451</td>
<td>0.397</td>
<td>0.148</td>
</tr>
<tr>
<td>White</td>
<td>0.521</td>
<td>0.541</td>
<td>0.020</td>
<td>0.006</td>
</tr>
<tr>
<td>Black</td>
<td>0.224</td>
<td>0.209</td>
<td>-0.015</td>
<td>0.014</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.171</td>
<td>0.167</td>
<td>-0.004</td>
<td>0.509</td>
</tr>
<tr>
<td>Chest Pain</td>
<td>0.626</td>
<td>0.632</td>
<td>0.006</td>
<td>0.400</td>
</tr>
<tr>
<td>ESI</td>
<td>2.377</td>
<td>2.363</td>
<td>-0.014</td>
<td>0.114</td>
</tr>
<tr>
<td>Private Insurance</td>
<td>0.438</td>
<td>0.443</td>
<td>0.005</td>
<td>0.485</td>
</tr>
<tr>
<td>Prior Hypertension DX</td>
<td>0.469</td>
<td>0.479</td>
<td>0.010</td>
<td>0.186</td>
</tr>
<tr>
<td>Prior Diabetes DX</td>
<td>0.222</td>
<td>0.234</td>
<td>0.011</td>
<td>0.061</td>
</tr>
</tbody>
</table>

To quantify the extent to which high-admitting physicians are assigned higher- or lower-risk patients, I predict patients’ one-year mortality as a logit function of the vector of patient observables $P_{ij}$, then use the fitted values from this prediction as $Y_{ij}$ to estimate Equation (2.4). Table 2.3 presents the results, which indicate that on average, the predicted mortality rate of patients assigned to high-admitting physicians is only 0.002 higher than that of patients assigned to low-admitting physicians, who have a mean predicted mortality rate of 0.102. The difference is not large or statistically significant, suggesting that on average, high- and low-admitting physicians are assigned patients with similar risk profiles.

**Effects on Patient Treatment in the ED**

Table 2.4 displays the results of estimating Equation (2.4), where $Y_{ij}$ represents a dummy variable indicating whether or not ED patient $i$ was admitted to the hospital by physician $j$. Given that both types of physicians appear to be treating
similar patients, the difference in their admission rates is large: on average, high-admitting physicians are roughly seven percentage points more likely to admit a given patient than are low-admitting physicians, who admit roughly one-third of their patients. This represents a relative difference in admission rates of at least 20%, implying that all else equal, substantial variation in treatment intensity results from what is equivalent to a coin-toss between two physicians in the ED.

One potential explanation for this pattern is that high-admitting physicians might merely be demonstrating a greater willingness to trade-off between inpatient treatment intensity and outpatient ED treatment intensity. That is, rather than conducting additional procedures in the ED, the high-admitting physician is more likely to admit the patient to the hospital instead. If this were the case, the variation in admission rates might substantially overstate the extent to which aggregate treatment intensity varies. However, I find no evidence of this. Table 2.5 shows that on average, there is no difference in ED charges per patient across high- and
Table 2.4: Effect on Likelihood of Patient Admission to Hospital

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admitted</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td></td>
</tr>
<tr>
<td>HighAdmitter</td>
<td>0.076***</td>
<td>0.069***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.343</td>
<td>0.343</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>22.2%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>19,005</td>
<td>19,005</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

low-admitting physicians, suggesting that ED treatment intensity is similar for both physicians.\(^7\)

Table 2.5: Effect on ED Charges (in Standard Deviations)

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observation</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td></td>
</tr>
<tr>
<td>HighAdmitter</td>
<td>0.018</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.638</td>
<td>0.638</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>2.8%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>17,176</td>
<td>17,176</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

\(^7\)Note in Table 2.5 that ED charges were missing for roughly 10% of ED visits. However, the missing charges were evenly balanced across high- and low-admitting physicians.
Often, for patients near the admission margin, physicians will opt to hold the patient in the ED for further observation before making the decision to admit or discharge. With this in mind, it is worth considering whether high-admitting physicians are simply less likely to hold a patient for observation, and are therefore more likely to admit the patient to the hospital instead. However, Table 2.6 shows that these two types of physicians exhibit no difference in the rate at which they hold patients in the ED for observation. In fact, I find that patients treated by high-admitting physicians on average spend roughly 20 minutes longer in the ED than those treated by low-admitting physicians (see Table 2.7). This, combined with the similarity in ED charges, suggests that the variation in admission rates represents a large marginal increase in overall treatment intensity, and not merely a displacement of ED procedures in favor of inpatient ones.

Table 2.6: Effect on Likelihood of Patient Being Held for Observation

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>HighAdmitter</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
</tr>
</tbody>
</table>

"Control" Mean 0.198 0.198
Relative Effect 0.3% -0.1%
Time Fixed Effects X X
Patient Controls X X
Observations 19,005 19,005

Note: *p<0.1; **p<0.05; ***p<0.01
Table 2.7: Effect on ED Length of Stay (in Hours)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HighAdmitter</strong></td>
<td>0.345</td>
<td>0.334</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.045)</td>
</tr>
<tr>
<td><strong>&quot;Control&quot; Mean</strong></td>
<td>4.346</td>
<td>4.346</td>
</tr>
<tr>
<td><strong>Relative Effect</strong></td>
<td>7.9%</td>
<td>7.7%</td>
</tr>
<tr>
<td><strong>Time Fixed Effects</strong></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Patient Controls</strong></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>19,005</td>
<td>19,005</td>
</tr>
</tbody>
</table>

*Note: ᵇp<0.1; **p<0.05; ***p<0.01

Effects on Subsequent Patient Outcomes

Since patients being treated by high- and low-admitting physicians are similar, yet the former receive substantially greater treatment intensity than the latter, this provides a unique opportunity to estimate the causal impact of this marginal treatment on patient health. Given the costly nature of hospital admissions, appreciable improvements in patient health would be necessary for these admissions to be cost-effective. A common means of evaluating the efficacy of a treatment is to measure its impact on the rate of re-hospitalization – if a treatment is effective, it should reduce the likelihood that the treated patient needs urgent medical care in the near future.

Using the same methodological framework outlined above, I estimate the effect of being assigned to a high-admitting physician on ED patients’ rate of subsequent hospital visits. The results in Table 2.8 show no difference in patients’ likelihood of returning to the ED within a week, month, or year of their initial visit,
suggesting that on average the marginal admissions did not prevent the recurrence of emergency medical episodes. In fact, as displayed in Table 2.9, patients assigned to high-admitting physicians are almost 10% more likely than those assigned to low-admitting physicians to be admitted to the hospital from the ED in the year following their initial visit. This pattern implies that admission to the hospital is itself causally linked to poorer health outcomes among this population. Given the high prevalence of hospitalization-associated disability among elderly patients (Covinsky et al., 2011), it is reasonable to suspect that these apparently low-value admissions are in fact harmful to patient health. However, I find no statistically significant impact on patient mortality rates over the same time period (see Table 2.10).

**Table 2.8: Effect on Subsequent ED Visit Rates**

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>7-Day (1)</th>
<th>30-Day (2)</th>
<th>1-Year (3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighAdmitter</td>
<td>-0.001</td>
<td>-0.001</td>
<td>-0.004</td>
<td>-0.004</td>
<td>0.012</td>
<td>0.014*</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.008)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.037</td>
<td>0.037</td>
<td>0.119</td>
<td>0.119</td>
<td>0.377</td>
<td>0.377</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>-3.5%</td>
<td>-3%</td>
<td>-3.7%</td>
<td>-3.4%</td>
<td>3%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Observations</td>
<td>18,957</td>
<td>18,957</td>
<td>18,778</td>
<td>18,778</td>
<td>16,097</td>
<td>16,097</td>
</tr>
</tbody>
</table>

*Note:* p<0.1; **p<0.05; ***p<0.01

**Physician Practice Style Variation under Uncertainty**

The results discussed above demonstrate that there is substantial variation in the rate at which emergency physicians admit their patients to the hospital that
Table 2.9: Effect on Subsequent Admission Rates

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>7-Day (1)</th>
<th>30-Day (2)</th>
<th>1-Year (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighAdmitter</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.014</td>
<td>0.053</td>
<td>0.53</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>-5.9%</td>
<td>4.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Observations</td>
<td>18,957</td>
<td>18,778</td>
<td>16,097</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2.10: Effect on Subsequent Mortality Rates

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>7-Day (1)</th>
<th>30-Day (2)</th>
<th>1-Year (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighAdmitter</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.011</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>15.7%</td>
<td>6.5%</td>
<td>4.7%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Observations</td>
<td>19,005</td>
<td>19,005</td>
<td>19,005</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01
cannot be explained by observable patient characteristics, and that the marginal admissions generated by this variation appear to be of questionable value. This suggests that the low-admitting physicians at this hospital are not generally under-admitting their patients, but that high-admitting physicians may be over-admitting. In which scenarios is this undesirable practice most likely to arise? As discussed in Section 2.2, the decision to admit or discharge a patient from the ED is in large part a synthesis of test results and physician judgment. To a degree, these two factors can be thought of as substitutes in a physician’s decision-making process (Doyle Jr et al., 2010). When the results of diagnostic tests largely resolve the uncertainty regarding their patients’ condition, physicians’ subsequent decisions should be fairly homogeneous. On the other hand, when test results are less informative, physicians’ professional judgment should play a greater role in deciding the appropriate course of action, leading to more heterogeneity in the treatment received by their patients.

As mentioned in Section 2.2, one of the primary goals of an emergency physician treating a patient presenting with chest pain or shortness of breath is to rule out the possibility that the patient has suffered a heart attack. The preferred first-line test for this condition is the troponin blood test, a highly sensitive lab assay that detects cardiac tissue damage. When this test is positive, it is likely that the patient’s condition is serious, and so physicians will admit the vast majority of these patients. On the other hand, when the troponin test is negative, a heart attack can generally be ruled out. However, since the patient could be suffering from any number of other diseases, a negative result tells the physician much less about the severity of the patient’s condition than a positive result.

---

8A positive troponin test is defined by a test value ≥ 0.01ng/ml.
In order to explore the extent to which diagnostic uncertainty accounts for the variation in physicians’ admitting practices, I obtained troponin test results relating to the ED visits in my analysis. Approximately 47% of patients in my sample received a troponin test, and of those tested, approximately 26% had a positive result. Among patients testing positive, roughly 86% were admitted to the hospital, and those assigned to low-admitting physicians were admitted at nearly the same rate as those assigned to high-admitting physicians (see Table 2.11). But among patients testing negative, of whom roughly 40% were admitted, the rate of the difference in admission rates was large: those assigned to high-admitting physicians were about 26% more likely to be admitted than those assigned to low-admitting physicians (see Table 2.12).

Table 2.11: Effect on Admission Rates for Patients with Positive Troponin Test

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Admission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>HighAdmitter</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
</tr>
</tbody>
</table>

"Control" Mean       0.856    0.856
Relative Effect       2.1%     2.1%
Time Fixed Effects    X        X
Patient Controls      X        
Observations          2,347    2,347

Note: *p<0.1; **p<0.05; ***p<0.01

This gap is comparable to the relative difference in admission rates among patients who were not administered the troponin test (see Table 2.13). However, high-admitting physicians are roughly 8% more likely to administer the troponin test to their patients (see Table 2.14) than low-admitting physicians. Assuming
that the marginal tested patients are lower-risk than the infra-marginal tested patients, this implies that on average, those receiving the test from high-admitting physicians are lower risk, and should therefore be less likely to be admitted. In light of this, the estimated 26% gap in admission rates among patients testing negative is likely to understate the effect of physician assignment on admission within this population. These stark patterns in admissions rates are consistent with the aforementioned hypothesis that high degrees of diagnostic uncertainty lead to physician practice style becoming a pivotal factor in the admission decision.

**Relationship with Physician Characteristics**

Given the substantial variation in emergency physician admitting practices, what factors might contribute to physicians’ tendency to admit their patients? Both the number of physicians included in my sample and the richness of their observable characteristics are limited, so this paper cannot provide strong causal evidence to answer this question. Nonetheless, I present some descriptive analyses below.
### Table 2.13: Effect on Admission Rates for Patients not given Troponin Test

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>HighAdmitter</td>
<td>0.055***</td>
<td>0.055***</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.224</td>
<td>0.224</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>24.5%</td>
<td>24.6%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>10,065</td>
<td>10,065</td>
</tr>
</tbody>
</table>

*Note:* *p*<0.1; **p*<0.05; ***p*<0.01

### Table 2.14: Effect on Troponin Testing Rates

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Troponin Testing</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>HighAdmitter</td>
<td>0.039***</td>
<td>0.033***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>&quot;Control&quot; Mean</td>
<td>0.455</td>
<td>0.455</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>8.6%</td>
<td>7.3%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>19,005</td>
<td>19,005</td>
</tr>
</tbody>
</table>

*Note:* *p*<0.1; **p*<0.05; ***p*<0.01
to lay a foundation for future research on the relationship between physician characteristics and their admitting behavior. In particular, I explore the role that experience and patient volume plays in physicians’ admission rates.

Whether physicians become more efficient in making admission decisions over time is an open question. There exists a related and growing literature which examines the role that physicians’ human capital plays in the cost-effectiveness of medical care. Notably, Doyle Jr et al. (2010) demonstrates in an inpatient setting that physicians possessing greater human capital provide care that is less expen-
sive, while achieving the same patient outcomes as those possessing less human capital. On the other hand, Van Parys (2016) explores heterogeneity in practice styles among physicians treating patients presenting in the ED with relatively minor injuries, and finds that physicians do not appear to reduce treatment costs as they gain additional years of experience. Yet it stands to reason that as physicians gain experience, the fraction of cases in which they face diagnostic uncertainty might decrease. Given that the results in this paper indicate that high-admitting physicians are likely admitting too many patients, this would suggest that physicians’ admission rates would fall over time. However, in the ED studied in this paper, there appears to be no relationship between physicians’ experience and their admission rate. Figure 2.1 plots the admission rate (adjusted for the same patient and time controls discussed in Section 2.4) for each physician (N = 58) as a function of their medical school graduation year, and shows a near-zero correlation between the two.

There is also a large literature demonstrating the strong and positive relationship between physicians’ patient volume and their patients’ health outcomes (Halm et al., 2002; Chowdhury et al., 2007), particularly among surgeons. The primary rationale for this pattern is intuitive: the more practice a physician receives in performing a given procedure, the better she performs. If the same logic holds for emergency physicians, then perhaps higher-volume physicians would exhibit a tendency to admit fewer low-risk patients than lower-volume physicians. On average, the physicians studied in this paper handled 85 cases per year (recall that this refers to only the cases included in the analysis sample – patients presenting with chest pain or shortness of breath for the first time). Variation in patient volume across physicians is substantial: the physicians in the 25th and 75th percentile averaged 66 and 106 cases, respectively. Figure 2.2 plots physician
Note: "Cases per Year" refers to the total number of visits within the analysis sample relating to each physician, divided by the number of years between the first and last visit observed for each physician.
admission rates as a function of annual patient volume, and consistent with this volume-quality hypothesis, there is a clear negative relationship between these two factors. Regressing the patient admission indicator on the physician’s case volume, I estimate that each additional case is associated with a 0.12% to 0.17% relative decrease in the likelihood of admission (see Table 2.15).

**Table 2.15: Relationship between Admission Rates and Physician Case Volume**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases per Year</td>
<td>-0.0006***</td>
<td>-0.0005***</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Mean Admission Rate</td>
<td>0.378</td>
<td>0.378</td>
</tr>
<tr>
<td>Relative Effect</td>
<td>-0.17%</td>
<td>-0.12%</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patient Controls</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>19,005</td>
<td>19,005</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

2.6 Discussion

**Welfare Implications**

The estimates in Section 2.5 show that for every 100 patients presenting in the ED with chest pain or shortness of breath, high-admitting physicians will on average admit 7 more patients than low-admitting physicians. If the costs per admission are assumed to be roughly $20,000 (Sabbatini et al., 2018), this implies an additional cost per patient treated by the high-admitting physician of approximately $1,400.
Moreover, the immediate treatment costs associated with these admissions is only part of the aggregate social cost: hospitalization disrupts patients’ lives, discomforts them, and introduces them to an environment which increases their likelihood of acquiring a new illness or disability (Covinsky et al., 2011).

However, the results in this paper show that these costs do not appear to be offset by a commensurate improvement in patient health outcomes. Patients treated by high-admitting physicians are no less likely to visit the ED in the future, and are in fact more likely to be admitted to the hospital in the year following their initial visit. Similarly, there appears to be no beneficial effects with regards to patient mortality, although the estimates in Table 2.10 lack the precision necessary to rule out small but meaningful reductions in mortality rates with a high degree of confidence. But given the detrimental impact on long-term readmission rates shown in Table 2.9, any such mortality improvements seem unlikely.

Limitations

As is the case with all studies conducted within a single institution, the extent to which the findings of this paper generalize to EDs throughout the country is unknown. For instance, my results suggest that the marginal admissions generated by high-admitting physicians are of low value, but the ED studied in this paper has admission rates that are substantially higher than the national average. Assuming that there are decreasing returns to admitting patients, it is plausible that high-admitting physicians working in EDs with lower admission rates may generate benefits that clearly outweigh the costs. Only by incorporating data from a diverse set of institutions can such concerns be addressed.

Another limitation of this paper is scope of the analysis, which only included patients presenting with chest pain or shortness of breath. I chose to focus on
this group because it represents one of the most common case types faced by emergency physicians, and because their treatment is relatively high-stakes in nature. Restricting the sample in this manner also allowed for a straightforward examination of the role that diagnostic uncertainty plays in admission rate variation across physicians. Nonetheless, results based on patients presenting with these symptoms may differ from those based on other populations.

2.7 Conclusion

This paper makes use of the quasi-random assignment of patients to physicians that takes place within an ED. This mechanism allows for average differences in patient treatment and subsequent outcomes to be causally attributed to the emergency physician responsible for the patient’s care. Specifically, I focus on the physician’s decision to admit a patient to the hospital, an event that triggers substantially greater costs and treatment intensity. Using records pertaining to ED patients presenting with chest pain or shortness of breath, I find that there is substantial variation in the rate at which physicians admit patients to the hospital that cannot be explained by patient characteristics. Namely, I note that physicians with above-median admission rates are on average roughly 20% more likely to admit a given patient than physicians with below-median admission rates. Despite the higher degree of treatment intensity received by patients of high-admitting physicians relative to those of low-admitting physicians, I find no evidence that this reduces their mortality risk or likelihood of returning to the ED, but rather increases their likelihood of being admitted to the hospital in the future. This suggests that these marginal admissions are likely to be low-value.
I go on to explore mechanisms that could lead to the variation in admission rates across physicians within a given hospital. First, I find that the variation in admission rates primarily arises through differences in the manner in which physicians handle cases characterized by a high degree of diagnostic uncertainty. It is in these scenarios that emergency physicians must rely more heavily on their professional judgment in deciding whether to admit or discharge a patient. Despite this, I find that the length of physicians’ professional experience does not appear to be correlated with their propensity to admit patients. However, I do find that physicians’ admission rates correlate negatively with their patient volume, suggesting that more frequent practice yields more discerning admission decisions.

Future research might seek to extend this analysis to a broader set EDs. Doing so would not only provide additional statistical power to more precisely estimate the relationship between admission and mortality, but would also shed light on the external validity of the findings of this paper. Given the high cost of inpatient medical care, and the increasing share of inpatient stays that originate in the emergency department, identifying low-value admissions and developing strategies to prevent them is a worthwhile health policy goal.
Chapter 3

Fight the Power [Calculation]:
Incorporating Compliance Prediction into RCT Design

3.1 Introduction

Empirical researchers frequently struggle to obtain sufficient statistical power to estimate parameters of interest with precision. When using preexisting observational data, the sample size is generally fixed, so the researcher has limited options to improve statistical power. On the other hand, when designing a randomized control trial (RCT), which has become an increasingly important empirical tool for applied microeconomists, the researcher selects the number of participants, thereby exercising direct control over the study’s statistical power. Yet the choice of sample size is typically subject to numerous constraints, which can be financial or bureaucratic in nature. Participant recruitment, provision of treatment, and the
procurement of follow-up data can all be factors that limit the researcher’s ability to obtain sufficient power.

In most randomized control trials (RCTs) of interest to social scientists, particularly those conducted outside of a lab setting, the experimenter cannot exercise perfect control over which participants will receive treatment, and which will not. Some of those assigned to the control group may obtain the treatment nonetheless, while some assigned to the treatment group may not in fact obtain the treatment. Moreover, these forms of noncompliance are not generally random – they are in most cases plausibly correlated with participant characteristics including treatment effects. Therefore, estimates of average treatment effects will suffer from selection bias if obtained by naively comparing the outcomes of those who received the treatment and those who did not. To avoid this problem, it is therefore common practice to use random assignment as a proxy or an instrument for treatment receipt, which respectively yield estimates of the intent-to-treat (ITT) or the local average treatment effect (LATE).

Both the ITT and the LATE are determined only by the treatment effects experienced by the participants who comply with treatment assignment. The inclusion of non-compliers in an experiment with a fixed number of participants therefore attenuates the experiment’s statistical power. Given this fact, the choice of sample size may not the only margin that the researcher can control when performing a power calculation. What if the researcher could screen out non-compliers from a pool of potential participants? This would undoubtedly increase the power of the experiment. Yet the compliance of a given participant cannot be directly observed, making any such effort a non-trivial exercise. In this paper, I propose straightforward compliance prediction procedures based upon data from existing experiments or quasi-experiments to assess the likelihood of a
potential participant’s compliance. Armed with this information, the researcher can selectively enroll the participants that are most likely to comply.

The remainder of this paper is organized as follows. Section 3.2 reviews the existing treatment effects framework that is widely used throughout the economics literature. Section 3.3 discusses the estimation of treatment effects and the role that compliance plays in the power calculation when designing an RCT. Section 3.4 describes how compliance prediction can be used to inform RCT design, and Section 3.5 lays out the framework for estimating a compliance prediction function. Section 3.6 implements two empirical applications for this function using publicly available data from the Oregon Health Insurance Experiment (Finkelstein et al., 2012). Section 3.7 discusses some of the limitations of compliance prediction and how to address them, while Section 3.8 concludes.

3.2 Treatment Effects Framework

Potential Outcomes

Throughout this paper I will make use of the potential outcomes framework developed by Rubin (1977) and employed throughout the instrumental variables treatment effect literature (Angrist et al., 1996). Specifically, the notation in this section closely follows that employed in Angrist and Fernandez-Val (2013). A participant in an experiment has two potential outcomes, \( Y_0 \) and \( Y_1 \), only one of which will be realized. The observed outcome \( Y \) is defined as:

\[
Y = Y_0 + (Y_1 - Y_0)D
\]
where binary indicator $D$ represents the participant’s treatment status. Expressed in regression notation, this is equivalent to:

$$Y = \alpha + rD + \eta$$

where $\alpha = E[Y_0]$, $\eta = Y_0 - \alpha$, and $r = Y_1 - Y_0$, the individual-level causal treatment effect. However, in many randomized control trials, compliance with treatment assignment is imperfect. That is, the experimenter cannot exercise perfect control over which participants will receive treatment, and which will not. It is therefore common to model an individual’s potential receipt of treatment, $D_0$ and $D_1$, as a function of the random treatment assignment, binary indicator $Z$:

$$D^Z = D^Z_0 + (D^Z_1 - D^Z_0)Z$$

or in regression notation:

$$D^Z = \gamma + pZ + \nu$$

where $\gamma = E[D^Z_0]$, $\nu = D^Z_0 - \gamma$, and $p = D^Z_1 - D^Z_0$. Under this framework, all participants can be characterized as falling into one of four mutually-exclusive groups:

1) "Compliers" – $(D^Z_1 = 1, \text{ and } D^Z_0 = 0)$ – those who "comply" with their assignment to the treatment or control group;

2) "Always-takers" – $(D^Z_1 = 1, \text{ and } D^Z_0 = 1)$ – those who would always take-up treatment regardless of their assignment;

3) "Never-takers" – $(D^Z_1 = 0, \text{ and } D^Z_0 = 0)$ – those who would never take-up treatment regardless of their assignment;
4) "Defiers" – \((D_1^Z = 0, \text{ and } D_0^Z = 1)\) — those who "defy" their assignment to the treatment or control group;

In this paper, as is generally the case in the treatment effects literature, I will assume that treatment takeup is monotonic in treatment assignment. That is, \(D_1^Z \leq D_0^Z\) for all individuals (i.e. there are no "defiers").

Note that individual compliers cannot be directly identified, as the counterfactual treatment assignment is never observed for any given individual. Despite this, Abadie (2003) demonstrates that when baseline covariates are available for experiment participants, treatment assignment and takeup can be used to estimate the average characteristics of compliers. The method that I propose in Section 3.5 is based on a related concept: using the characteristics of observable non-compliers, a function that estimates an individual’s likelihood of compliance can be constructed.

**Intent-to-Treat (ITT)**

As alluded to in the introduction, the ITT is simply the average causal effect of being assigned to the treatment arm rather than the control arm:

\[
\text{ITT} = E(Y|Z = 1) - E(Y|Z = 0).
\]

Substituting the above potential outcomes notation into this equation yields:

\[
\text{ITT} = E[Y_0 + (Y_1 - Y_0)D_1^Z] - E[Y_0 + (Y_1 - Y_0)D_0^Z] = E[Y_0 + (Y_1 - Y_0)D_1^Z] - E[Y_0 + (Y_1 - Y_0)D_0^Z].
\]

Given that treatment assignment is random, the proportion of always-takers, never-takers, and compliers in each arm will be equal in expectation. Therefore the above expectations can be decomposed into these four groups. However, since
always-takers only realize \( Y_1 \) and never-takers only realize \( Y_0 \), while defiers are assumed not to exist, only the compliers remain:

\[
= [E(Y_1 - Y_0|D^Z_1 > D^Z_0)P(D^Z_1 > D^Z_0)]
\]

\[
+ E(Y_1 - Y_1|D^Z_1 = D^Z_0 = 1)P(D^Z_1 = D^Z_0 = 1)
\]

\[
+ E(Y_0 - Y_0|D^Z_1 = D^Z_0 = 0)P(D^Z_1 = D^Z_0 = 0)]
\]

\[
+ [E(Y_0 - Y_1|D^Z_1 < D^Z_0)P(D^Z_1 < D^Z_0)]
\]

\[
= [E(Y_1 - Y_0|D^Z_1 > D^Z_0)P(D^Z_1 > D^Z_0)]
\]

\[
= E(Y_1 - Y_0|D^Z_1 > D^Z_0)[E(D^Z_1) - E(D^Z_0)]
\]

In other words, the ITT is equal to the average treatment effect for compliers \( (E(Y_1 - Y_0|D^Z_1 > D^Z_0)) \) multiplied by the proportion of participants that are compliers \( (E(D^Z_1) - E(D^Z_0)) \), which I refer to as the compliance rate. This result reflects the fact that the ITT only captures the treatment effects on the compliers. This is intuitive, as always- and never-takers’ treatment takeup is not influenced by their treatment assignment; treatment effects are for these groups are therefore not identified by the experiment.

**Local Average Treatment Effect (LATE)**

While the ITT is the average effect of a participant being assigned to the treatment group, the LATE represents the average effect of treatment assignment on the compliers:

\[
LATE = E(Y_1 - Y_0|D^Z_1 > D^Z_0) = \frac{E(Y_1 - Y_0|D^Z_1 > D^Z_0)[E(D^Z_1) - E(D^Z_0)]}{E(D^Z_1) - E(D^Z_0)} = \frac{ITT}{E(D^Z_1) - E(D^Z_0)}
\]

The LATE is therefore equal to the ITT divided by the compliance rate.
3.3 Estimation of Treatment Effects

The ITT can be estimated via the following OLS regression:

\[
Y_i = \hat{\phi}_0 + \hat{\beta}_{ITT}Z_i + \hat{\epsilon}_i
\]

The LATE can be estimated as via the following two-stage least squares (2SLS) regression:

\[
Y_i = \hat{\beta}_0 + \hat{\beta}_{LATE}\hat{D}_i + \hat{\epsilon}_i
\]

where random treatment assignment is used as an instrument for takeup indicator \(D_i\) according to the following first stage regression:

\[
\hat{D}_i = \hat{\xi}_0 + \hat{\beta}_{FS}Z_i
\]

In Appendix B.1, I show that the formula for the estimated standard error of \(\hat{\beta}_{LATE}\) can be written as a function of the estimated compliance rate \(\hat{\beta}_{FS}\):

\[
\frac{\hat{\omega}}{\sqrt{N\hat{\beta}_{FS}}}
\]

where \(N\) represents the number of observations in the sample.

Role of Compliance in the Power Calculation

When designing an RCT, the researcher performs a power calculation to determine the sample size necessary to detect a hypothetical treatment effect with high probability (often 0.8). The test statistic associated with the LATE estimate of \(\hat{\beta}_{LATE}\) given the null hypothesis that \(\hat{\beta}_{LATE} = 0\) is:

\[
t_0 = \frac{\hat{\beta}_{LATE}}{\hat{\sigma}_{LATE}}
\]
Prior to treatment assignment, given a desired significance level \( \alpha \), a hypothetical LATE value of \( \delta_0 \), and a sufficiently large \( N \) such that the \( t \)-distribution approximates the normal distribution, the formula for statistical power can be represented as follows:

\[
\text{Power} = P(t_0 > t^a|\beta_{LATE} = \delta_0) \\
= P\left(\frac{\hat{\beta}_{LATE}}{\hat{\sigma}_{LATE}} > t^a|\beta_{LATE} = \delta_0\right) \\
= 1 - P\left(\frac{\hat{\beta}_{LATE}}{\hat{\sigma}_{LATE}} < t^a|\beta_{LATE} = \delta_0\right) \\
\approx 1 - \Phi(t^a - \frac{\delta_0}{\hat{\sigma}_{LATE}}) = 1 - \Phi(t^a - \frac{\sqrt{N\hat{\beta}_{FS}\delta_0}}{\hat{\omega}}).
\]

where in the last line, \( \hat{\sigma}_{LATE} = \frac{\hat{\omega}}{\sqrt{N\hat{\beta}_{FS}}} \) is substituted. Given a desired minimum power level \( \rho_0 \), an \( N_0 \) must be chosen such that:

\[
1 - \Phi(t^a - \frac{\sqrt{N\hat{\beta}_{FS}\delta_0}}{\hat{\omega}}) > \rho_0 \\
\Rightarrow \Phi(t^a - \frac{\sqrt{N\hat{\beta}_{FS}\delta_0}}{\hat{\omega}}) < 1 - \rho_0 \\
\Rightarrow \frac{\sqrt{N\hat{\beta}_{FS}\delta_0}}{\hat{\omega}} > t^a - z_{(1-\rho_0)} \\
\Rightarrow N_0 > \left(\frac{t^a - z_{(1-\rho_0)}}{\hat{\beta}_{FS}\delta_0}\right)^2 \hat{\omega}^2.
\]

This highlights the trade-off between the magnitude of the compliance rate and the minimum number of observations necessary to conduct an RCT with sufficient power: an increase in the compliance rate implies that fewer participants are necessary.

Now suppose that it were possible to increase the compliance rate in an experiment from \( \hat{\beta}_{FS} \) to \( \gamma \hat{\beta}_{FS} \), where \( \gamma > 1 \). Then holding all else constant, the
sample size required to maintain the same level of statistical power is

\[ N_1 > \left( \frac{t^2 - z(1-\rho_0)}{\gamma \beta_{FS} \delta_0} \right)^2 \omega^2 \]

making the ratio of the minimum sufficient sample sizes \( N_0^* \) and \( N_1^* \):

\[ \frac{N_1^*}{N_0^*} = \frac{1}{\gamma^2} \]

The minimum sample size is thus decreasing in the square of \( \gamma \). Alternatively, this can be represented as an elasticity between \( N_1 \) and \( \gamma \) by taking the logarithm of the equation and differentiating with respect to \( \gamma \):

\[
\ln N_1 = \ln N_0 - 2 \ln(\gamma) \\
\epsilon_{N_1, \gamma} = \frac{dN_1}{d\gamma} \times \frac{\gamma}{N_1} = -2
\]

Thus for each one percent increase in the compliance rate, the researcher can reduce the sample size by two percent while maintaining the same level of power, making it clear that even modest gains in compliance rates can drive down minimum sample sizes substantially. This is especially true when compliance rates are relatively low. For instance, boosting compliance from 25% to 35% would allow the researcher to cut her sample size nearly in half while maintaining the same statistical power, holding all else constant.¹

### 3.4 Compliance Prediction Use Case

When designing an RCT, in addition to baseline characteristics \( X \) for all potential participants, suppose that a researcher also possessed a function \( c(X) \in [0,1] \)

\[ \frac{N_1}{N_0} = \frac{1}{(1/35)^2} = 1.36 \]

1. ^1
such that \( P(D_{1i} > D_{0i}) = c(X_i) + \epsilon_i \) – that is, an imperfect (yet informative) ex-ante predictor of compliance for a given participant. How might this additional information be used to inform the trial’s design, and what are the potential benefits and drawbacks in doing so?

Using the existing baseline characteristics and the compliance prediction function \( c(X) \), the researcher could generate predicted compliance probabilities for potential participants then recruit them for participation in descending order of predicted compliance. By enrolling the top \( N \) individuals then randomizing treatment among them, the researcher can boost the compliance rate (and therefore the statistical power) of the RCT relative to a scenario in which enrollees are enrolled randomly from within a given population. Alternatively, the researcher could conduct the study with fewer participants, while maintaining the same level of statistical power. However, if the primary parameter of interest is the ITT – that is, the experimenter is drawing a random sample from a specific population in part because the compliance rate itself is of importance, and not merely a nuisance parameter – this is unlikely to be a wise choice. I will therefore focus on the case in which the primary parameter of interest is the LATE.

Using such a method to enroll participants into an RCT does not harm the internal validity of the estimated treatment effects that it yields, as treatment assignment is still random within the selected sample. However, it’s clear that the generalizability of the results may suffer in some circumstances, since the participants are no longer representative of the population from which they are being drawn. The extent to which the external validity of the experiment is diminished is determined by the distribution of individual-level treatment effects within this population. Recall that the LATE is the average treatment effect for the compliers. To further simplify notation from the previous section, let \( \pi_i = Y_{i1} - Y_{i0} \).
the unobserved individual-level treatment effect, and let $C_i = 1$ if the individual is a complier, and $C_i = 0$ otherwise. The LATE can therefore be defined in the following manner:

$$LATE = E(Y_1 - Y_0 | D_1 > D_0) = \frac{E(C_i \pi_i)}{E(C_i)} = E(\pi_i) + \frac{cov(C_i, \pi_i)}{E(C_i)}$$

That is, the LATE can be decomposed into two parts: the population average treatment effect (ATE) and the covariance between treatment effect and compliance (normalized by the compliance rate).

Given this fact, the researcher should first carefully consider the likely relationship between compliance and treatment effects prior to undertaking an effort to select participants based on predicted compliance:

1° If uncorrelated, this procedure is particularly appealing, as it improves statistical power while recovering an unbiased estimate of the population ATE.

2° If positively correlated, the sample LATE will overstate the population LATE. However, if the primary goal of the experiment is to conduct a minimum viable RCT to determine whether a treatment effect exists for some subpopulation, this could be helpful for procuring additional resources for future studies with a larger, more representative sample.

3° If negatively correlated, the sample LATE will understate the population LATE. In this case, selecting a sample based on predicted compliance would be undesirable. However, if compliance is in fact negatively correlated with treatment effects, it may be difficult for an RCT to detect a statistically significant treatment effect regardless of how participants are selected.

Empirical strategies for exploring this relationship are discussed in Section 3.7.
3.5 Identifying Likely Compliers

Given the gains in statistical power that can be obtained by increasing the number of compliers in a study, it may be a goal worth pursuing. In this section, I lay out a straightforward framework for identifying likely compliers.

Framework

Using the same notation employed above, the probability that a participant is an always-taker given observable characteristics $X$ is:

$$Pr(D_0 = 1) = E(D_0|X)$$

and similarly the probability of that she is a never-taker is:

$$Pr(D_1 = 0) = 1 - E(D_1|X)$$

thus (again assuming the non-existence of defiers) the probability of being a complier is simply the probability of not being an always-taker or never-taker:

$$Pr(D_1 > D_0) = 1 - Pr(D_0 = 1) - Pr(D_1 = 0)$$

$$= E(D_1|X) - E(D_0|X)$$

This last expression can be estimated piece-wise, since clearly $D_1$ is observable within the treatment group and $D_0$ is observable within the control group. Yet since treatment assignment is random (implying $Z \perp X, D_0, D_1$), the baseline characteristics of these identifiable non-compliers are on average the same as those of unidentifiable non-compliers in the population from which the sample was drawn. We can therefore generate unbiased predicted always-taker (never-taker)
probabilities for individuals in the entire sample using functions estimated within the control (treatment) group.

**Estimation**

Consistent with previous literature on the subject (such as Heckman and Vytlacil (1999)), an individual’s decision to obtain treatment can be described using a random utility model so that treatment takeup for is a function of observable characteristics \( X_i \), treatment assignment \( Z_i \), and idiosyncratic error term \( \varepsilon_i \):

\[
D_i = \begin{cases} 
1 & h(X_i, Z_i) + \varepsilon_i > 0 \\
0 & h(X_i, Z_i) + \varepsilon_i < 0 
\end{cases}
\]

where

\[
E(D|X, Z) = Pr(h(X, Z) + \varepsilon > 0).
\]

In principle, the function \( h(X, Z) \) could be estimated empirically using a limited dependent variable regression model. Yet simply modeling take-up would not enable us to specifically identify likely compliers, since both always-takers and compliers in the treatment arm will take up treatment. But as mentioned above, compliance can be expressed as a function of an individual’s likelihood of being an always-taker or never-taker, since always-takers are identified in the control arm, and never-takers are identified in the treatment arm:

\[
D_{0i} = \begin{cases} 
1 & f(X_i|Z_i = 0) + \varepsilon_{0i} > 0 \\
0 & f(X_i|Z_i = 0) + \varepsilon_{0i} < 0 
\end{cases}
\]

\[
D_{1i} = \begin{cases} 
1 & g(X_i|Z_i = 1) + \varepsilon_{1i} > 0 \\
0 & g(X_i|Z_i = 1) + \varepsilon_{1i} < 0 
\end{cases}
\]
Estimating predicted compliance under the framework above is straightforward. As above, predicted compliance is

\[
E(D_1|X) - E(D_0|X)
\]

\[
Pr(g(X|Z = 1) + \varepsilon_1 > 0) - Pr(f(X|Z = 0) + \varepsilon_0 > 0)
\]

where functions \(g(X)\) and \(f(X)\) are then estimated using only the treatment and control groups, respectively. After estimation, these functions can be used to generate a predicted compliance for the entire sample:

\[
\hat{C}_i = \hat{D}_{1i} - \hat{D}_{0i} = \hat{g}(X_i) - \hat{f}(X_i)
\]

### 3.6 Empirical Applications

In order to demonstrate the feasibility of this methodology and estimate the potential gains in compliance rates that might be achieved using real experimental data, I made use of the publicly-available Oregon Health Insurance Experiment (Finkelstein et al., 2012) data ("OHIE"). This study represents a promising test case for my proposed methodology, given its large size (approximately 75,000 total participants, nearly 30,000 of which were randomly assigned to the treatment group), and its relatively low compliance rate (~0.25).

**Background**

In 2008, Oregon provided uninsured low-income households otherwise ineligible for Medicaid with the opportunity to obtain health insurance through the state’s Medicaid expansion program. Since anticipated demand for this insurance far exceeded supply (funds allowed for roughly 10,000 additional "slots"), the state
conducted a lottery. Lottery winners (the "treatment group") were given the opportunity to apply for insurance under this expansion program; those who lost (the "control group") were not). If lottery winners applied and met the program’s criteria, their household would receive health insurance coverage.

This framework meant that compliance with the randomization would be imperfect – many of those who won the lottery did not ultimately obtain health insurance coverage (the "treatment"), either because they failed to apply, or because they failed to meet the program’s criteria. Similarly, many of those who did not win the lottery were nonetheless able to obtain health insurance by other means. Thus, using the treatment-effects parlance described in detail above, households in the control group that obtained coverage were "always-takers," and those in the treatment group who did not obtain coverage were "never-takers." The experiment was conducted over the course of eight lottery drawings occurring between March and September of 2008. Conditional on the number of individuals covered by the application, treatment assignment within each drawing was random (Finkelstein et al., 2012).

The OHIE dataset used in this paper contains limited baseline characteristics, all of which comprise the vector $X$ throughout these empirical applications. Demographic information is limited to the applicant’s birth year, gender, the number of individuals covered by the application (1, 2, 3 or more – a rough proxy for household size), as well as whether the applicant provided a phone number or P.O. Box address, lives in an US Census Bureau Metropolitan Statistical Area (MSA), or requested materials in a language other than English. Also included are data from administrative records indicating the amount of SNAP and TANF benefits (if any) that the applicant received in 2007, the year prior to application. In addition, the
dataset provides variables indicating whether applicants signed up on the first or last day of the application window.

Since the experiment was rolled out over the course of several lottery drawings, the OHIE data can be used to evaluate a hypothetical application of the methodology outlined in this paper: what if the observed takeup patterns exhibited by participants in the early drawings were used to select the participants to be randomized in the later drawings?

**Application #1: Experiment Size is Capped**

For this empirical application, suppose that after conducting the first five lottery drawings (comprised of almost 38,000 individuals), it was determined that there were only sufficient resources to incorporate an additional 5,000 individuals into the experiment. Given concerns regarding the statistical power of the experiment, the goal is to select 5,000 participants from the remaining pool of roughly 36,000 applicants such that the compliance rate is maximized. This will allow the experiment to more precisely estimate the impact of insurance expansion on health care utilization – specifically, whether the patient visited a provider within 12 months of the drawing.

Following the framework laid out in Section 3.5, I use the data from the first five drawings to estimate a compliance prediction function via OLS. First, within the treatment group, I estimate the probability that an individual takes up treatment

$$D_{1i} = X_{1i}\varphi_1 + \epsilon_{1i}$$

then estimate the same equation within the control group

$$D_{0i} = X_{0i}\varphi_1 + \epsilon_{0i}$$
and using the two vectors of coefficients estimated above, define an individual’s compliance score to be

$$\hat{C}_i = X_i(\hat{\phi}_1 - \hat{\phi}_0).$$

After generating compliance scores for all individuals in the remaining pool of applications, I select the 5,000 individuals with the highest compliance scores for inclusion in the experiment.

**Table 3.1: Application #1 first stage estimates**

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 5,000</td>
</tr>
<tr>
<td>Lottery</td>
<td>0.380***</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
</tr>
<tr>
<td>Observations</td>
<td>5,000</td>
</tr>
</tbody>
</table>

*Note: * $^*p<0.1; ^{**}p<0.05; ^{***}p<0.01$

Tables 3.1 and 3.2 present the first stage and LATE estimates under three samples drawn from the remaining applicants: column (1) was generated using the 5,000 most likely compliers, following the procedure outlined above; column (2) represents estimates generated using a random sample of 5,000 individuals; and column (3) contains the estimates that would have been generated had the entire remaining population of applicants been included in the experiment.\(^2\) The results in Table 3.1 demonstrate that by selecting 5,000 of the most likely compliers, the first stage estimate is increased by nearly 70% relative to the scenario in which

\(^2\)Note that the number of observations in the LATE table includes only those for whom a 12-month survey response was obtained.
5,000 participants were chosen randomly. This in turn results in much greater precision when estimating the LATE (Table 3.2), as the standard error is reduced by nearly one half,\(^3\) while the point estimate closely approximates that of the full population.

**Table 3.2: Application #1 LATE estimates**

<table>
<thead>
<tr>
<th></th>
<th>Visited Provider in Past 12 Months</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 5,000</td>
<td>Random 5,000</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Insurance</td>
<td>0.171***</td>
<td>0.136</td>
<td>0.215***</td>
</tr>
<tr>
<td></td>
<td>(0.052)</td>
<td>(0.097)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Observations</td>
<td>1,712</td>
<td>1,603</td>
<td>11,519</td>
</tr>
</tbody>
</table>

*Note:* \(p<0.1; \, **p<0.05; \, ***p<0.01\)

**Application #2: Incorporate into Power Calculation**

In this empirical application, suppose that after conducting the first five lottery drawings, rather than facing a hard cap on the number of subsequent experiment participants, the researcher wishes to incorporate compliance prediction into the power calculation used to determine the number of participants necessary to carry out an experiment. In this case, the data from the first five drawings can be used to estimate the number of individuals needed to obtain sufficient power when they are drawn from the pool of applicants who are most likely to comply. In order to do so, the researcher must construct a compliance prediction function as described

\(^3\)Note that the empirical relationship between the first stage estimate and the estimated standard error is consistent with the formula derived in Appendix B.1.
in Application #1 above, and generate compliance scores for all individuals in the first five drawings. Then, generate a series of first stage estimates, using individuals ranked in the top 1, 2, 3... 100 compliance score percentiles.

Since the statistical power of an experiment can be represented as a function of its first stage estimate and sample size (see Section 3.3), this series of estimates can be used to plot a power curve as a function of the number of individuals with the highest compliance scores (see Figure 3.1). This curve shows that an estimated statistical power of slightly greater than 0.8 could be achieved with 2,000 participants drawn from those within the remaining pool of applicants.

\[\text{Estimated power of slightly greater than 0.8 could be achieved with 2,000 participants drawn from those within the remaining pool of applicants.}\]

\[\text{4The power estimates reflected in Figure 3.1 assume an effect size equal to the observed LATE within the first five drawings.}\]
with the highest compliance scores. Column (1) of Table 3.3 presents the LATE estimates generated using this population, which can be compared to the LATE estimates generated using the entire remaining pool of applicants in column (2). This comparison shows that the point estimate from the small subsample of 2,000 individuals is very close to that of the entire applicant pool, and sufficiently precise to make it statistically significant ($p < 0.05$).

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Visited Provider in Past 12 Months</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 2,000</td>
<td>All</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td>0.224**</td>
<td>0.215***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.090)</td>
<td>(0.033)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>691</td>
<td>11,519</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* $^*p<0.1; ~^**p<0.05; ~^***p<0.01$

### 3.7 Limitations

#### External Validity of Treatment Effects

As discussed in Section 3.4, the LATE estimate recovered from a selected high-compliance sample of individuals may or may not generalize well to that of the broader population of compliers. However, to the extent that the distribution of treatment effects is likely to be similar to that of previously conducted experiments (or quasi-experiment), the relationship between predicted compliance and treatment effects in these studies can be used to assess the generalizability of the LATE
found in the selected sample. For example, using the framework discussed above, generate predicted compliance for each participant, select a subsample $S$ of those most likely to comply, and using 2SLS compute $\hat{LATE}^S$ within this subsample and $\hat{LATE}^{-S}$ within the remaining participants. A simple $t$-test can then be used to test for significant differences between the two estimates under the null hypothesis that they are the same:

$$t = \frac{\hat{LATE}^S - \hat{LATE}^{-S}}{\sqrt{\hat{\sigma}_S^2 + \hat{\sigma}_{-S}^2}}$$

where $\hat{\sigma}_S$ and $\hat{\sigma}_{-S}$ are the estimated standard errors of $\hat{LATE}^S$ and $\hat{LATE}^{-S}$. However in practice, an RCT may not have sufficient power to detect meaningful differences between the two estimates, leading to under-rejection.
Alternatively, the relationship between predicted compliance and treatment effects can be explored graphically. Consider grouping individuals into bins according to their predicted compliance quantile, estimating the LATE within each bin, and plotting these estimates. Figure 3.2 presents an example of this exercise using individuals in the first five drawings of the OHIE. In this case, the plot shows a near-zero correlation between treatment effects and predicted compliance. This is consistent with the fact that the LATE within the high-compliance group closely mirrored that of the broader population in the empirical applications carried out in Section 3.6 above.

External Validity of the Compliance Prediction Function

Another limitation of this methodology concerns the generalizability of the estimated compliance prediction function from one sample to another. This is due to the fact that an individual’s compliance with treatment assignment is not a universal fixed trait – rather, it depends on the parameters of the experiment. For instance, a complier in the OHIE will not necessarily be a complier in an experiment that randomly assigns access to a job training program. It is therefore important that the experiment in which the researcher wishes to increase compliance rates bears substantial similarity to the experiment upon which the compliance function is estimated.

For instance, in the applications presented in Section 3.6, it was plausible that the individuals who were compliers in the first five OHIE drawings would be similar to compliers within the remaining pool of applicants, as they were drawn from the same population, and the parameters of the experiment remained unchanged. For these reasons, this compliance prediction methodology seems particularly well-suited improving the efficiency of trials that are rolled out in
stages (like the OHIE), in which the nature of the experiment remains largely unchanged in each stage.

**Overfitting the Compliance Prediction Function**

However, even if the experiments and population from which participants are drawn are identical, there is still another issue which can complicate the estimation of a useful compliance prediction function: overfitting. When generating a prediction function from a data sample, it will reflect both the true relationship between the independent and dependent variables in the population, as well as spurious correlation which exists between them only in the sample at hand. This degrades the quality of the prediction function when applied to another sample. Since the purpose of the compliance prediction function is to predict the likelihood of compliance among potential participants in a similar subsequent experiment, this poses a problem.

Moreover, as the number of independent variables grow relative to the size of the sample, so does the severity of the overfitting problem. So while obtaining a large number of relevant baseline characteristics provides the researcher with an opportunity to build a powerful predictor of compliance, overfitting must be taken into consideration when estimating such a predictor. In such cases, the functional form of the predictor should not be constructed using OLS (or other unpenalized regression methods) as described above, but through the use of machine learning algorithms, which have been designed with out-of-sample prediction performance in mind.
3.8 Conclusion

Researchers frequently face constraints on the number of participants they can enroll in an experiment. These constraints are compounded by imperfect compliance, often resulting in underpowered studies, or preventing studies from taking place at all. In this paper, I propose a method that allows researchers to use data from existing experiments or quasi-experiments to improve the compliance rate of their studies by identifying and enrolling participants with above-average compliance rates, thereby increasing the statistical power of their studies. Using data from the OHIE, I demonstrate the feasibility of these methods, showing that substantial compliance gains can be achieved even when there are minimal baseline characteristics with which to identify likely compliers.

That said, the methods proposed in this paper should not be universally adopted into RCT design. For instance, they are ill-suited to experimental designs in which the ITT is the chief parameter of interest, or when drawing from a representative sample of the population is of primary importance. Moreover, recent research has demonstrated the value of including non-compliers in an experiment as a means of assessing the global external validity of treatment effects (Kowalski, 2016a,b). Despite these limitations, in appropriate settings the ability to identify and enroll likely compliers into an RCT provides researchers with a useful tool that may provide much-needed statistical power to underpowered studies, or allow for otherwise infeasible studies to be conducted.
References


Covinsky, Kenneth E, Pierluissi, Edgar, and Johnston, C Bree. Hospitalization-associated disability: "She was probably able to ambulate, but I’m not sure". *JAMA*, 306(16):1782–1793, 2011.


Kowalski, Amanda E. Doing more when you’re running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments. NBER working paper, National Bureau of Economic Research, 2016a.

Kowalski, Amanda E. How to examine external validity within an experiment. Unpublished manuscript, 2016b.


Pines, Jesse M, Mutter, Ryan L, and Zocchi, Mark S. Variation in emergency
department admission rates across the united states. Medical Care Research and

Pope, Devin and Simonsohn, Uri. Round numbers as goals: Evidence from

Pope, J Hector, Aufderheide, Tom P, Ruthazer, Robin, Woolard, Robert H, Feldman,
James A, Beshansky, Joni R, Griffith, John L, and Selker, Harry P. Missed
diagnoses of acute cardiac ischemia in the emergency department. New England

Rubin, Donald B. Assignment to treatment group on the basis of a covariate.

Rui, P, Kang, K, and Albert, M. National hospital ambulatory medical care survey:
2010 emergency department summary tables. National hospital ambulatory
medical care survey, Centers for Disease Control and Prevention, 2013.

Sabbatini, Amber K, Wright, Brad, Hall, M Kennedy, and Basu, Anirban. The cost
of observation care for commercially insured patients visiting the emergency

Schuur, Jeremiah D and Venkatesh, Arjun K. The growing role of emergency
departments in hospital admissions. New England Journal of Medicine, 367(5):

Shah, Anuj K and Oppenheimer, Daniel M. Heuristics made easy: An effort-


Stein, Paul D, Huang, Hsiu-ling, Afzal, Adnan, and Noor, Hussam A. Incidence of
acute pulmonary embolism in a general hospital: Relation to age, sex, and race.

Tanabe, Paula, Gimbel, Rick, Yarnold, Paul R, and Adams, James G. The emer-
gency severity index (version 3) 5-level triage system scores predict ed resource

Thaler, Richard H and Sunstein, Cass R. Nudge: Improving Decisions About Health,

Torio, CM and Moore, BJ. National inpatient hospital costs: The most expensive
conditions by payer, 2013. Statistical brief 204, Agency for Healthcare Research
and Quality, 2016.


WHO. The international classification of diseases, 2016.

Appendix A

Appendix to Chapter 1

A.1 Data

Determining the Month-Age of Patients

The Truven Commercial Claims and Encounters Enrollment Detail Table records individual-level information for each month that the individual is enrolled in a health plan. This information includes the individual’s year of birth and integer age as of the beginning of each calendar month. Given these two facts, it is straightforward to determine the individual’s month of birth so long as during the observed period (2005 - 2013), she is enrolled in her month of birth and in the month following it, or is enrolled for twelve consecutive months. This is true of roughly 95% of all enrollees. However, the remaining 5% of individuals are not enrolled in the months necessary for this to be observed. For each enrollee whose month of birth cannot be determined with certainty, I randomly assign a month of birth drawn from the enrollee’s potential months of birth (i.e. the calendar months that cannot be ruled out as the enrollee’s month of birth). This imputation
could lead to classical measurement error of the discontinuities at age 40, thereby attenuating my estimates. However, as noted in Section 1.3, my results are robust to the inclusion/exclusion of individuals with imputed months of birth.

**Imputing IHD Testing in Truven Inpatient Claims**

Due to the manner in which inpatient and outpatient claims are filed with insurers, the troponin test, which I use as a proxy for IHD testing, can only be readily identified in outpatient ED claims. But at minimum, the troponin test should be ordered for all patients in the ED who are diagnosed with IHD. I therefore make the assumption that all ED patients who are admitted to the hospital with an IHD diagnosis received this test, and assume that all other admitted patients did not. To the extent that this fails to account for ED patients who are admitted to the hospital without a diagnosis of IHD but nonetheless received a troponin test, this will understate the overall testing rate. To the extent that there exists a discontinuous jump in admissions at the age-40 threshold, this assumption might attenuate my estimated discontinuity in testing. However, empirically I find that this assumption has no meaningful impact on the magnitude of the discontinuity in testing.
A.2 Robustness Checks

RD Bandwidth

Tables A.1 through A.3 present local linear estimates of the discontinuities in physician behavior using the optimal bandwidth (15 months) described in Section 1.4, as well as using bandwidths half and twice this size, as suggested by Lee and Lemieux (2010). The estimated discontinuity in the heart attack testing rate is very similar in each specification, as would be expected given the highly linear relationship between age and testing rates shown in Figure 1.1. On the other hand, the magnitude of the discontinuities in the IHD diagnosis and admission rate estimates are diminishing in the size of the bandwidth (although still statistically significant). This is consistent with the patterns displayed in Figures 1.2 and 1.3, which show that the slopes of these rates grow flatter near the threshold.

Table A.1: Heart attack testing rates, by bandwidth

<table>
<thead>
<tr>
<th></th>
<th>Tested for Heart Attack x100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 Months</td>
</tr>
<tr>
<td><strong>Turned40</strong></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>0.712***</td>
</tr>
<tr>
<td></td>
<td>(0.191)</td>
</tr>
<tr>
<td><strong>Intercept</strong></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>9.448***</td>
</tr>
<tr>
<td></td>
<td>(0.150)</td>
</tr>
<tr>
<td><strong>Relative Change</strong></td>
<td>7.5%</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>312,452</td>
</tr>
</tbody>
</table>

Note: See Table 1.1.
Table A.2: *IHD diagnosis rates, by bandwidth*

<table>
<thead>
<tr>
<th></th>
<th>Diagnosed with IHD x100</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 Months (1)</td>
<td>15 Months (2)</td>
<td>30 Months (3)</td>
<td></td>
</tr>
<tr>
<td><strong>Turned40</strong></td>
<td>0.175***</td>
<td>0.131***</td>
<td>0.092***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.024)</td>
<td>(0.025)</td>
<td></td>
</tr>
<tr>
<td><strong>Intercept</strong></td>
<td>0.634***</td>
<td>0.679***</td>
<td>0.701***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.015)</td>
<td>(0.017)</td>
<td></td>
</tr>
<tr>
<td><strong>Relative Change</strong></td>
<td>27.7%</td>
<td>19.3%</td>
<td>13.1%</td>
<td></td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>663,437</td>
<td>1,325,771</td>
<td>2,742,363</td>
<td></td>
</tr>
</tbody>
</table>

Note: See Table 1.2.

Table A.3: *IHD admission rates, by bandwidth*

<table>
<thead>
<tr>
<th></th>
<th>Admitted with IHD x100</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 Months (1)</td>
<td>15 Months (2)</td>
<td>30 Months (3)</td>
<td></td>
</tr>
<tr>
<td><strong>Turned40</strong></td>
<td>0.091***</td>
<td>0.068***</td>
<td>0.048**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.023)</td>
<td>(0.020)</td>
<td></td>
</tr>
<tr>
<td><strong>Intercept</strong></td>
<td>0.349***</td>
<td>0.382***</td>
<td>0.394***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.015)</td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td><strong>Relative Change</strong></td>
<td>26.1%</td>
<td>17.9%</td>
<td>12.3%</td>
<td></td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>663,437</td>
<td>1,325,771</td>
<td>2,742,363</td>
<td></td>
</tr>
</tbody>
</table>

Note: See Table 1.3.
Duration of Heart Attack Follow-up Period

In the Tables A.4 (a) and (b) below, I demonstrate the robustness of the discontinuity in female missed diagnoses found in Table 1.6. Using follow-up periods of 30 and 180 days, the estimates are consistent with my primary specification of 90 days, and as expected, grow in the number of days observed post-visit.

Table A.4: Subsequent heart attack diagnosis rates

(a) 30 Days

<table>
<thead>
<tr>
<th>Dependent variable: Missed IHD, 30 Days (per 100,000)</th>
<th>All</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Intercept</td>
<td>65.913***</td>
<td>45.151***</td>
<td>91.713***</td>
</tr>
<tr>
<td></td>
<td>(2.243)</td>
<td>(5.327)</td>
<td>(10.155)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>-9.1%</td>
<td>-50.5%</td>
<td>16.9%</td>
</tr>
<tr>
<td>Observations</td>
<td>1,315,741</td>
<td>732,085</td>
<td>583,656</td>
</tr>
</tbody>
</table>

(b) 180 Days

<table>
<thead>
<tr>
<th>Dependent variable: Missed IHD, 180 Days (per 100,000)</th>
<th>All</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Intercept</td>
<td>117.748***</td>
<td>86.661***</td>
<td>156.342***</td>
</tr>
<tr>
<td></td>
<td>(5.152)</td>
<td>(7.491)</td>
<td>(16.038)</td>
</tr>
<tr>
<td>Relative Change</td>
<td>-1.2%</td>
<td>-31.2%</td>
<td>20%</td>
</tr>
<tr>
<td>Observations</td>
<td>1,315,741</td>
<td>732,085</td>
<td>583,656</td>
</tr>
</tbody>
</table>

Note: See Table 1.6.
A.3 Cost of Testing for IHD

Throughout this paper, I use the troponin blood test as a proxy for IHD testing. I do so not only because it is the first-line test for ruling out a heart attack, but also because it is used almost exclusively for this purpose, making it a uniquely precise proxy for IHD testing. However, it is not the only procedure that a physician might conduct when exploring an IHD diagnosis. To quantify the effect of the age-40 heuristic on all relevant IHD testing costs, I must consider its effects on other tests as well.

Ideally, I would do so by measuring the discontinuity at the threshold for each procedure separately, but the nature of these procedures make it difficult to do so. First, by far the two most common are the ECG and chest X-ray, but both are also frequently used in pursuing diagnoses unrelated to IHD. Second, the remaining procedures, which include stress testing and imaging, are more costly and thus much less frequently used than the troponin, ECG, or X-ray tests. As a result, it is difficult to precisely measure the effect of the age-40 heuristic on these procedures individually. However, since they are typically only used in diagnosing IHD in conjunction with a troponin test, the rate at which these other procedures are performed within the population that is administered the troponin test serves as a reasonable approximation of the frequency with which these tests are used to diagnose IHD.¹

Table A.5 below presents how I arrive at the IHD testing cost estimate used in the cost-benefit analysis performed in Section 1.5. Among the 40-year-old female outpatients in my analysis sample who are given the troponin test, column

¹To the extent to which the marginal patient affected by the age-40 heuristic is less likely to receive these additional procedures than the average patient receiving a troponin test, this estimate likely overstates the aggregate cost of IHD testing for marginal patients.
(a) states the proportion that are also given additional screening procedures. Column (b) displays the 2017 national CMS fee-for-service reimbursement rates (physician plus outpatient hospital service payments, or lab payments in the case of the troponin test), which I use as an estimate of the cost of performing these procedures. Column (c) is the product of columns (a) and (b). Thus the sum of column (c) is my estimate of the average aggregate cost of IHD testing per 40-year-old female patient.

Table A.5: *Average aggregate IHD testing cost*

<table>
<thead>
<tr>
<th>Procedure Name</th>
<th>Testing Rate (a)</th>
<th>CMS Payment (b)</th>
<th>Weighted Cost (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Troponin</td>
<td>1.000</td>
<td>13.50</td>
<td>13.50</td>
</tr>
<tr>
<td>Chest X-Ray</td>
<td>0.792</td>
<td>79.48</td>
<td>62.95</td>
</tr>
<tr>
<td>ECG</td>
<td>0.925</td>
<td>65.64</td>
<td>60.72</td>
</tr>
<tr>
<td>Stress ECG</td>
<td>0.048</td>
<td>269.35</td>
<td>12.93</td>
</tr>
<tr>
<td>Echo</td>
<td>0.030</td>
<td>505.88</td>
<td>15.18</td>
</tr>
<tr>
<td>Stress Echo</td>
<td>0.016</td>
<td>513.42</td>
<td>8.21</td>
</tr>
<tr>
<td>Chest CT</td>
<td>0.133</td>
<td>349.57</td>
<td>46.49</td>
</tr>
<tr>
<td>MPI</td>
<td>0.025</td>
<td>1223.10</td>
<td>30.58</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>$250.56</strong></td>
</tr>
</tbody>
</table>
Appendix B

Appendix to Chapter 3

B.1 Reformulated LATE Standard Error

The formula for the variance of the bivariate instrumental variables coefficient\(^1\) can be presented as:

\[
\text{Var}(\beta_{IV}) = \frac{\text{Var}(u|Z)}{N \times \text{Var}(D) \times \rho_{D,Z}^2}
\]

where \(u\) is the error term, \(Z\) is the instrument, and \(D\) is the treatment, as defined in this paper, and \(\rho_{D,Z}^2\) is the square of the correlation coefficient between \(D\) and \(Z\) (Wooldridge, 2006). The correlation coefficient is defined as:

\[
\rho_{D,Z} = \frac{\text{Cov}(D, Z)}{\sqrt{\text{Var}(D) \times \text{Var}(Z)}}
\]

which implies that

\(^1\)For ease of exposition, the formulas I display here assume that errors are homoskedastic, which simplifies the notation while clearly conveying the intuition. Allowing for heteroskedastic errors does not change the conclusions ultimately reached.
\[
Var(\beta_{IV}) = \frac{Var(u|Z)}{N \times Var(D) \times \frac{Cov(D,Z)^2}{Var(D)Var(Z)}} = \frac{Var(u|Z)}{N \times \frac{Cov(D,Z)^2}{Var(Z)}}.
\]

Recall that the first-stage coefficient \( \beta_{FS} \), resulting from the regression of \( D \) on \( Z \), is by definition:

\[
Var(\beta_{FS}) = \frac{Cov(D, Z)}{Var(Z)}
\]

which in turn means that

\[
Var(\beta_{IV}) = \frac{Var(u|Z)}{N \times \beta_{FS}^2 \times Var(Z)}
\]

and the asymptotic standard error of the coefficient is

\[
\sigma_{IV} = \sqrt{Var(\beta_{IV})} = \frac{\sqrt{Var(u|Z)}}{\sqrt{N} \times \sqrt{Var(Z)} \times \beta_{FS}}.
\]

By the analogy principle, the estimated standard error is therefore

\[
\hat{\sigma}_{LATE} = \frac{\sqrt{\sum_{i=1}^{N} \hat{u}_i^2 / \sum_{i=1}^{N} (Z_i - \bar{Z})^2}}{\sqrt{N}\hat{\beta}_{FS}}.
\]

To further simplify notation, substitute \( \hat{\omega} = \sqrt{\sum_{i=1}^{N} \hat{u}_i^2 / \sum_{i=1}^{N} (Z_i - \bar{Z})^2} \) into the above expression, yielding

\[
\hat{\sigma}_{LATE} = \frac{\hat{\omega}}{\sqrt{N}\hat{\beta}_{FS}}.
\]