# Essays in Health Economics

# Essays in Health Economics

A dissertation presented

by

Steven White

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

August 2019

Dissertation Advisor: David Cutler Steven White

**Essays in Health Economics**

## Abstract

This dissertation presents three empirical studies that are broadly concerned with the valuation of human health and the evaluation of policies designed to improve it. The first study evaluates how a supply-side policy intended to restrict drug diversion, prescription monitoring, affects opioid prescribing and pain management in hospitals and homes. The results indicate that prescription monitoring reduces opioid use in outpatient settings not but in hospitals, and appears to have only modest effects on pain management, with suggestive evidence indicating that it enables more effective targeting of opioid therapy. The second study reevaluates the labor market evidence on compensating differentials for fatal injury risk, showing that the standard estimator for the sample mean value of a statistical life (VSL) is biased when the compensating differentials vary across the wage distribution and that correcting for this bias is quantitatively significant. The last study revisits an old question with a new identification strategy, using the financing mechanism for state medical boards as an instrument for local physician supply in order to evaluate the impact of supply shocks on local health care markets. The results indicate that a larger physician supply leads to changes in the style of medical practice but the welfare implications of this are unclear.

# Contents

# List of Tables

# List of Figures

## Acknowledgments

I am extremely grateful to my advisors, David Cutler, Mandy Pallais, and Danny Shoag. This dissertation benefited enormously from their support, corrections, and suggestions. I also benefited from comments on preliminary versions of this work presented in at the PF-Labor lunch at Harvard, especially comments from Larry Katz. I gratefully acknowledge funding from Harvard University throughout my graduate studies. I would also like to thank Tiantian, and my parents for believing in me for all these years.

# Introduction

Americans spend 18% of their incomes on healthcare but a majority of the population routinely expresses concerns about the quality and value of their care or their access to it. Health care is incredibly valuable because it can save your life, but not all healthcare provides significant enough benefits to be worth its price. The dividing line between worthwhile and excessively costly care remains murky and while many policies have been proposed to increase the quality or cost-effectiveness of care, their actual effects in the market are often unclear. This dissertation presents evidence from three empirical projects broadly unified by their focus on healthcare and its valuation.

The first chapter focuses on the opioid crisis in America, which is taking almost 50,000 lives each year through fatal drug overdoses. While most overdoses are caused by illicit drugs, most users start by abusing prescription opioids before progressing to illicit substance use. The chapter investigates the impacts of prescription monitoring programs, tools that can be used to restrict access to prescription opioids and thus help reduce diversion and abuse. By restricting access, however, these programs run the risk of harming legitimate pain patients since monitoring and the potential for sanctions could cause even well-intentioned doctors to be excessively cautious when prescribing opioids.

I assess this hypothesis by looking at shipments of opioids, tracked by the DEA's ARCOS system, to hospitals and pharmacies. Prescription monitoring is only intended to impact outpatient prescribing, where diversion and abuse are feasible, and consistent with that intention I find modest reductions in shipments to pharmacies but no effect on hospitals. I also evaluate the effect of prescription monitoring on the prevalence of severe pain based on two surveys

and find small but insignificant improvements in pain management. In summary, prescription monitoring does reduce consumption of opioids, but not in ways that unintentionally harm pain patients, and it might be helping legitimate patients access opioid therapy.

The second chapter looks at the value of interventions that save lives. If a technology can reduce the risk of dying by 1 percentage point but costs $100,000 is it worth the cost? The standard economic method for answering this question is to look at the choices people make when facing similar tradeoffs in the marketplace. The most common example is that some jobs require doing dangerous labor but pay higher wages as compensation. The amount of extra compensation per unit of risk, or value of a statistical life (VSL), is thus informative about how much life-saving technologies are worth to those workers, but I show that the standard statistic used to estimate the population VSL is biased except under implausible conditions. Specifically, I show that when the compensation workers require varies with their underlying earnings potential, and the elasticity of the VSL and earnings potential is not exactly one, then the standard statistic is biased.

In order to get a sense of the direction of the bias and whether it is quantitatively significant I estimate a quantile regression model which allows for some heterogeneity in estimated VSLs across the wage distribution. I then employ recently developed distributional methods to recover the distribution of VSLs in the population and provide a new estimate of the mean VSL. My estimate is significantly larger than an estimate obtained using the standard method on the same data and I show that further corrections for excluding high income workers from the sample raises the estimated mean VSL even further.

The final chapter looks at how changes in the supply of healthcare inputs can influence access to care. While having more hospital beds, scanning machines, and healthcare professionals seems like it would naturally improve access to care, there are reasons that some economists are skeptical. Physicians make recommendations about care so they can induce extra demand for their services by recommending more intensive treatment, especially when demand would otherwise be slack. In addition, productivity in the healthcare sector appears to vary enormously across regions, potentially in response to supply pressures. Areas with few doctors

might have to develop processes to use their resources efficiently while areas that have a richer supply of resources can provide all necessary care even with a poor allocation of resources.

I use two methods to assess how the supply of doctors in an area influences the physicians and patients there. First, I use new methods to probe the robustness of ordinary least squares estimates against selection bias, finding that the negative association between the number of doctors working in an area and lower physician incomes is unlikely to be fully explained by selection bias. Second, I use an instrument for physician supply to get two-stage least squares estimates of the effects of supply on various outcomes. While some of the estimates are noisy, I find relatively robust evidence that supply appears to influence how time is allocated between seeing patients and administrative work, with a larger supply leading to a greater focus on administration and each physician spending less time with patients.

# Chapter 1

# Does prescription monitoring harm patients?

## 1.1 Introduction

The United States is suffering from an epidemic of fatal drug overdoses, fed by an underlying epidemic of abuse of and dependence on opioids. While many deaths from overdoses result from consumption of non-prescription opioids including illicit fentanyls and heroin, most opioid use disorders appear to start with abuse of prescription opioids. At least 70% of those seeking inpatient treatment for heroin, for example, report first experimenting with prescription opioids supplied by doctors before progressing to heroin use and among recreational users the vast majority of pills are obtained from prescriptions directly from a doctor (27%) or a prescription for a friend or relative (52%) (Cicero et al. 2014, Jones et al. 2014). The epidemics have generated enormous interest in supply-side tools to limit opioid prescriptions to situations where there is a manifest need.

Many commentators have hypothesized that the opioid epidemic is intertwined with a related epidemic of chronic pain. While the origins of the pain epidemic are unknown, it appears that rising rates of obesity and disability are causal factors and the stagnation of middle income wage growth could also be a factor (Case and Deaton 2015, Case and Deaton 2017). The pain

epidemic hypothesis has led to widespread concern among both providers and patients that efforts to combat the drug epidemic by restricting opioid availability could have the unintended side effect of curtailing access to effective treatment for legitimate pain. Not much is known about these potential spillovers because data on volumes of prescriptions alone are not informative about mechanisms (National Academies of Science, Engineering and Medicine 2017). For example, researchers have shown that the tightening of rules regulating the prescription of hydrocodone combination products (e.g. Vicodin) in October 2014 led to decreases in prescribing (Chumpitazi et al. 2017, Oehler et al. 2016) but it remains unclear if patients were harmed. Similarly, the CDC's guidelines opioid prescribing guidelines, issued in 2016, appear to have curbed the initiation of high-dose opioid therapy but anecdotes suggest they have also unintentionally force taper and discontinuation of opioid therapy even for chronic pain patients who had been on stable doses long term (Dowell et al. 2019). Ongoing debate about how and whether to revise the guidelines has been hampered by a lack of systematic evidence on their overall effects for patients.

The basic problem is that supply-side policies can have different effects on different populations. If prescribers reduce the availability of pills only to abusers or those diverting them to secondary markets then there is unlikely to harm patients If prescribers reduce access uniformly then a rising tide of pain is more likely. Without knowing which categories patients fall into, or, better yet, having data on pain-related outcomes, it is difficult to assess overall effects of these policies.

Prescription Drug Monitoring Programs (PDMPs), state run databases of prescriptions for controlled substances, are one of the supply-side interventions that have generated the most interest among policymakers. The deployment of these monitoring programs in 49 of the 50 states and the District of Columbia, mostly over the past two decades, provides a natural experiment well suited for evaluating whether patients are harmed as a side effect of curtailed supply. Since these programs were developed at different times in different states they enable analysis in a difference-in-differences framework where states implementing the programs at different times serve as control groups for one another. Another feature of the implementation

of PDMPs is also useful in shedding light on how supply-side policies work. PDMPs were typically developed in two stages, first achieving capacity to collect data on prescribing patterns, as a deterrent against illegitimate prescribing, and a variable number of years later developing the capacity to provide information to prescribes in order to inform treatment decisions. As I discuss more below, the phased introduction of different elements of the PDMPs enabled me to distinguish different mechanisms of supply-side influence.

This chapter also contributes to the growing literature evaluating the effectiveness of PDMPs. This literature is divided with some research showing PDMPs have no effect on prescription volume, others showing minimal effects, and a few showing large effects (Ponnapalli et al. 2018). Some of the disagreement in the literature may be due to variation in the populations and time periods studied, but, as reviewed in more detail below, there is also substantial disagreement about the fundamental issue of what constitutes a PDMP and when each state's program qualifies as operational. I draw on the latest legal and historical research on the development of PDMPs and use a clear, consistent coding for what counts as an operational PDMP and the latest data on the dates these features were implemented. As future research builds on this approach it may help to resolve some of the discrepancies in the literature.

The rest of this chapter is organized as follows. The next section briefly reviews the context of of the opioid epidemic, first reviewing the historical evolution of the opioid crisis and then specifically reviewing the development of PDMPs and their most important features. The following section discusses the potential mechanisms by which PDMPs can influence prescribing of controlled substances and population health and my identification strategy for disentangling the different mechanisms. The following sections introduce the data used for the empirical analysis, discuss the econometric details of the empirical strategy, and present the results of my data analysis. The final section concludes.

## 1.2 Background

### 1.2.1 Opioids Background

Prescription opioids were a contributing factor in 17,029 deaths in the U.S. in 2017 and have been killing at similar rates for over a decade (Scholl et al. 2019). Just as concerning as the direct death toll from prescription opioids is their evident function as a gateway to other forms of opioid abuse including heroin and "fake pills" containing illicitly manufactured fentanyl (Cicero et al. 2014). These other opioids caused an additional 30,571 deaths in 2017 above and beyond those linked to prescription drugs directly (Scholl et al. 2019).

The opioid epidemic has additional health costs due to widespread addiction and abuse. SAMSHA estimates there were 1.753 million Americans 12 and older suffering from a pain relief substance use disorder in 2016 and hundreds of thousands also suffer from heroin use disorder (Han et al. 2015). Even with effective evidence based treatment these addictions result in substantial morbidity for years or decades, perhaps exceeding the direct cost from overdose deaths. As a result Eric D. Hargan, Acting Secretary of Health and Human Services, declared the "opioid crisis" a public health emergency in the fall of 2017 in recognition of its importance and the urgency need for a policy response (Hargan 2017). This subsection provides a brief background on the chemistry and history of opioids.

Opioids are chemicals, either extracted from the opium poppy plant (morphine and codeine) or synthesized in labs (oxycodone, fentanyl, etc.), that are similar to natural signaling molecules produced by humans called endorphins. Opioids have similar effects to endorphins since they bind to the same receptors dispersed throughout the body, leading to a variety of effects, mostly notably drowsiness, analgesia, euphoria, and slower breathing (Rosenblum et al. 2008). The opium poppy has been under human cultivation as its euphoric effects have long been recognized, but the modern medical and chemical understanding of opioids developed in the 19th century after German scientists extracted the active ingredients from raw opium in 1803. By 1820, extraction was taking place on an industrial scale and by the mid-19th century artificial synthesis of both the natural chemical as well as derivatives (e.g. heroin) was possible on an

industrial scale (Meldrum 2003). Opioid use was rare in the U.S. until the American Civil War prompted widespread use to treat the painful injuries suffered on the battlefield. This inaugurated what some historians call the first American opioid epidemic in the later third of the 19th century (Meldrum 2003, Booth 1996). The epidemic eventually prompted a federal response Harrison Narcotic Tax Act and related legislation in 1914 which banned the importation and distribution of opium and its derivatives (and cocaine) except by doctors and pharmacists "in the course of [their] professional practice only" (Day 1919). The act appears to have reshaped social views toward the use of opioids, even for severe pain, and these changes alon with vigorous enforcement influenced "physicians and patients alike to avoid opiates" for the following 50 years (Jones et al. 2018).

Attitudes began to change in the medical community in the 1970s as palliative care doctors used opioids to comfort patients on the cusp of death and later to manage chronic pain in cancer patients, not all of whom were expected to die (Tompkins et al. 2017). In 1986 the World Health Organization (WHO) adopted the position that cancer pain was inadequately treated and that opioids should be more widely available as an opioid for cancer patients. It claimed that "[p]sychological dependence is not an issue when strong opioids are taken to relieve cancer pain" and that "doses of morphine and other strong opioids can be increased indefinitely" (Meldrum 2003).

As views about malignant (cancer) pain evolved some doctors began to experiment with treating acute severe pain and chronic non-malignant pain with opioids. Two key papers published during this period led to a substantial reevaluation of the perceived risks of long term opioid therapy, although neither paper provides strong evidence on the question. Porter and Jick (1980), a short research letter in the *New England Journal of Medicine*, searched an early computerized record of inpatients and found that only 4 of 11,882 patients treated with opioids while hospitalized became addicted in some unspecified time frame and with unknown attrition. Portenoy and Foley (1986) looked at a convenience sample of 38 chronic pain patients taking doses equivalent to around 20mg of oxycodone (or less) per day, about half of whom had been using opioids for several years. They found 24 of the 38 got adequate relief although

none had substantial improvement in their employment due to treatment and about 5% (2/38) showed signs of dependence and drug abuse which was dismissed as a small risk.

Misreading of these papers led to a growing belief among doctors that "drug-seeking behavior synonymous with drug addiction does not occur in patients after pain relief with opioids" (McQuay 1999). As attitudes evolved professional organizations began to push for recognition of changing standards of treatment. For example, in 1995 the American Pain Society (APS) began a campaign to consider pain "the fifth vital sign" (Jones et al. 2018). In effect this meant that there should be more evaluation, monitoring, and treatment of pain, with opioids prescribed for use as needed. This campaign influenced the VA to adopt new pain management standards in 1999 and the Joint Commission on Accreditation of Healthcare Organizations (today known simply the Joint Commission) made similar revisions in 2000 (Baker 2017). Nevertheless, many doctors in the mid-2000s continued to feel that pain remained widely undertreated. Apfelbaum et al. (2003) is representative, reporting that while 90% of inpatients were satisfied with their pain medication "pain continues to be undermanaged" because "many patients continue to experience intense pain after surgery." Regulators and law enforcement began to take notice of these changing norms, dropping their nearly century old hostility toward opioids. The Federation of State Medical Boards issued a statement in 1998 that they would avoid regulatory scrutiny of heavy prescribers and the DEA adopted what it called a "balanced" policy toward outpatient opioid use (Tompkins et al. 2017). While historically doctors had only been under legal compulsion to avoid excessive prescribing of opioids, patients began to successful bring tort suits for inadequate treatment of pain in the 21st century. In 2001, for example, a California court ordered a judgment of $1.5 million against a physician found liable for undertreating a dying elderly man's pain (Meldrum 2003).

Due to these changes in knowledge, attitudes, and policy opioid prescribing increased dramatically through the 1990s and early 2000s as documented in a growing literature. Pletcher et al. (2008) found that opioids were prescribed at 23% of pain-related visits to emergency departments in 1993 and this rose to 37% by 2005 with no evidence that growth was slowing down. Olsen et al. (2006) found opioid prescribing by primary care physicians rose by 50% from 1992-

1993 to 1998-1999 while Kenan et al. (2012) found growth on both the intensive and extension margins for prescriptions between 2000 and 2009. The number of opioid prescriptions rose by 35.2% while the weight (in mg) rose by over 100%, implying dramatic growth in the number (or size) of pills in each prescription.

This sea change in prescribing based more on norms than on evidence inevitably resulted in pushback from some stakeholders. Journalists documented reports of widespread abuse and diversion of opioids with particular attention focused on OxyContin, an extended release version of oxycodone nicknamed "Hillbilly heroin." The original FDA-approved label on OxyContin, used from 1996 to 2001, said that "[d]elayed absorption as provided by OxyContin tablets, is believed to reduce the abuse liability of the drug" and that addiction to opioids is "very rare" when they are legitimately used to manage pain (Esch 2017). By 2001 regulators saw a need to act. The FDA required a new label for OxyContin which emphasized that OxyContin had an "abuse liability similar to morphine" and showed some vigilance in pushing Purdue to ensure promotional materials directed readers to a boxed warning on advertisements and pill bottles (GAO 2003). When reports of widespread abuse of OxyContin in Maine came to the attention of Jay McClosky, the U.S. attorney for that district, he sent a letter to Purdue and Maine physicians warning about reports of abuse (Pacheco 2002). This eventually developed into a lawsuit for "misbranding" the pharmaceuticals that ended when Purdue Frederick Company Inc, a holding company for Purdue Pharma, pled guilty and agreed to pay $634 million in fines. Three company executives also pled guilty individually and paid fines of $7.5 million to $19 million. Similar concerns prompted action by the U.S. congress which appropriate funding to expand prescription monitoring programs starting in 2002 (see section 1.2.2 below). The Senate Committee on Health, Education, Labor and Pensions began hearings on the issue in 2003 and requested a government accountability office report (GAO 2003).

A series of papers began to appear starting in 2005 that documented the quantitative extent of the growing epidemic in terms of diversion, addiction, diversion, and overdose deaths. Inciardi et al. (2007) warned of widespread and growing diversion while Cicero et al. (2005) documented rates of abuse of OxyContin, noting importantly that abuse of less potent hy-

drocodone combination products was potentially more prevalent than OxyContin abuse. A series of papers by Leonard J. Paulozzi of the CDC and coauthors noted high, rising, and variably mortality related to prescription opioids (Paulozzi et al. 2006 and Paulozzi and Ryan 2006 are representative). The anecdotal and statistical data prompted action on the part of most states in the middle and later years of the 00s. The majority of states developed modern prescription monitoring programs during this period and the Joint Commission revised its standards on pain assessment and management (Baker 2017). Knowledge of the widespread abuse of OxyContin and other opioids led the FDA to put pressure on manufacturers to develop abuse deterrent versions. Manufacturers were happy to oblige since generally the abuse deterrent mechanisms are patentable and thus insulated from competition. The fist formulation to hit the market was a new OxyContin in late 2010 followed by Opana (2011), Embeda (2012) and a flood of others in the following years (Rauck 2019).

As overdose death rates continued to climb the CDC began to view opioid abuse as a public health emergency and began work on guidelines for prescribers treating chronic non-malignant pain. This project built off a report the Agency for Healthcare Research and Quality commission a few years earlier (Chou et al. 2014) and emphasized the basic problem that at the time there was "no study [that] evaluated effects of long-term opioid therapy versus no opioid therapy" available to guide physicians. They also noted that literature on the rates of abuse had estimates ranging from 0.6% to 37.1% depending on the definition of what counted as abuse. As the CDC prepared its guidelines the Joint Commission began work on revising its standards related to treatment of pain and a fully revised set were issued in 2018 to emphasize non-opioid modalities for pain management, setting "reasonable expectations" about pain as opposed to the older focus on complete elimination, and to emphasize enabling provider access to PDMPs.

Since the blowback against opioids was, like the initial enthusiastic wave, based more on anecdotes and intuitions than hard evidence, there has been a predictable cry that "the pendulum has no swung too far" (Rothstein 2017) the past few years as doctors, patients and other advocates have expressed fears about undertreatment of pain in patients with legitimate needs. Doctors Stefan G. Kertesz and Adam J. Gordon, for instance, have cautioned about "inhumane

treatment" of chronic pain patients as practices change, especially hard limits on how much individual physicians feel comfortable prescribing and seven day limits on prescriptions for acute injuries (2017). Kertesz argues that there has been excessive focus on prescribing that has hit "diminishing returns" and will likely "creat[te] significant risks for patients" (2017). Other physicians have warned about the misinterpretation of the CDC's guidelines and legal requirements. For instance, one doctor reportedly told his patient that he had to reduce their opioid dosage "due to state law," apparently interpreting the CDC guidelines as a binding standard for what constitutes acceptable practice (Freyer 2016). The CDC lent credence to these argument in a recent article by the lead authors of the guidelines there clarified their proper use (Dowell et al. 2019). Rubin (2019) documents the work of Dr. Thomas Kline, a North Carolina physician who focuses on "pain refugees" who are in withdrawal and severe pain after their former doctors refused to continue opioid therapy they had started years earlier. Kline believes suffering from undertreatment of pain due to the opioid crackdown is widespread and is leading many people to take their own lives in desperation. He has cataloged 40 cases he claims illustrate the problem and estimates there are many more he is unaware of. Advocacy groups, such as the U.S. Pain Foundation have also mobilized, criticizing a "climate of fear" about opioid prescribing that has recently developed. Their work has helped to prompt to a series of articles from many media outlets with anecdotes about patients suffering (Taylor 2018, Huber 2018, Ehley 2018, Stone 2018, Firth 2019). While many of these criticisms are fairly non-specific with none offering statistical evidence and few giving concrete examples of patients hurt by supply-side restrictions, the sheer number of stories suggests that this is a serious concern worth looking into.

### 1.2.2 PDMP background

PDMPs have historically evolved in a series of phases from reactive tools of interest only to law enforcement into flexible decision-support tools used by healthcare providers. There is no consensus on the key stages of this transition but I conceptualize the evolution as taking place over three phases. The earliest prescription monitoring programs instituted special documen-

tation requirements for physicians prescribing certain controlled substances particularly likely to be abused. Typically this would involve acquiring special multi-copy prescription pads, often with special serial numbers for tracking, from a state authority. The typical multi-copy pad would produce three copies: one for the physician's personal records, one for the patient to take to a pharmacist, and a third that would be forwarded to a state agency within thirty days of the prescription either being written. Technically New York started the first such program in 1918 but provider opposition to the paperwork burden led the state to halt the program after just three years. (PDMP TTAC 2018). As a result, California's program, started in 1939, can be taken as the starting point for the first phase of PDMP development. California's records were kept by a newly created department, the Bureau of Narcotic Enforcement, operating within its attorney general's office. Seven states developed similar programs over the next fifty years in this first phase of PDMP development, primarily following California's lead in using multi-copy pads and housing authority for their PDMPs within a law enforcement agency (Fishman 2004).

These programs were designed with two main goals in mind. First, by theoretically monitoring all prescriptions for high-risk substances they could deter unscrupulous patterns of prescribing in the first place. Second, in cases where a physician was brought to trial for violating controlled substance regulations the paper records could be brought together to paint a picture of unusual prescribing volume. In practice the burden of compiling a large amount of paper by hand limited this use to serious cases such as those involving "pills mills" handing out prescriptions to abusers indiscriminately. For example, PDMP data assisted in prosecutions of "sleep" clinics flooding the streets with Quaaludes and barbiturates in the 1970s and "weight loss" clinics selling stimulants in the 1980s (PDMP TTAC 2018).

Oklahoma recognized that PDMPs had more potential if the records could be stored is a more useful format and inaugurated the second generation of PDMPs by developing an electronic record keeping system that came online in 1990. This effort spurred the American Society for Automation in Pharmacy (ASAP) to develop standards for the formatting of electronic data on controlled substances which were formally issued in 1994 and five other states

(Massachusetts, Utah, Nevada Indiana, Kentucky) followed Oklahoma's lead in the following decade, but most states remained skeptical of the need for these programs, particularly when they would have to be financed by local taxes. Notably, the eight states with preexisting PDMPs continued to use paper records as the conversion to electronic systems was deemed too expensive (Sacco et al. 2018).

The lower cost of electronic data storage compared to paper records spurred a few notable further innovations in this second phase. First, states began to track longer lists of controlled substances, eventually including all substances on either the state or federal list of scheduled, or controlled, drugs. The drug schedule system has existed on the federal level since the passage of the Controlled Substances Act (CSA) in 1970 and most states have their own sets of schedules as well. The scheduling system uses numbers to classify drugs based on their relative medical value and risk of abuse, with schedule I indicating drugs with "no currently accepted medical use" that also high potential for abuse such as heroin and LSD. Schedule I drugs are effectively illegal to manufacture or distribute. Schedules II through V represent progressively less dangerous drugs that all have an accepted medical use. The federal Schedule II includes most pure opioid preparations while schedule III includes some opioid combination products and less potent opioids mainly used for purposes other than pain relief like buprenorphine. Schedule IV includes drugs the DEA considers to have "low potential for abuse and low risk of dependence" in relative terms including all benzodiazepines like Xanax. Tramadol, an opioid that has unique chemical properties but is commonly use to treat chronic pain, was moved up from schedule V to schedule IV in 2014. All of the early multi-copy prescription pad programs only covered schedule II drugs due to the paperwork burdens but electronic programs all covered schedules II-IV and some also included schedule V.

The actual functions of PDMPs, however, largely remained reactive during this second phase. The electronic data was housed solely on the computers of the agency in charge of drug monitoring and was thus largely siloed off from clinical practice. A few states experimented with ways to make the data more useful instead of simply waiting for requests during prosecutions. Massachusetts, for instance, allowed physicians to request reports on the prescribing

history of specific patients they were concerned might be liable to abuse or divert prescriptions. After a physician made a request the PDMP would gather data on that patient and mail it on a floppy disc to the physician within a few days. Other states, including Nevada, experimented with faxing records to expedite delivery, but with a latency of days or hours these data retrieval processes were too slow to inform prescribers who had to make decisions while seeing patients. Recognizing this problem, Nevada tried mailing unsolicited reports to all physicians prescribing to a patient that was flagged for high risk of abuse and but most states ultimately decided it was important to find a way to give prescribers and pharmacists direct access to data on their patients, which lead to the third phase of PDMP development.

The opioid abuse epidemic, discussed in section 1.2.1, provided the impetus state and federal legislators needed to finance this next phase of PDMP development (Sacco et al. 2018). Congress authorized the a grant program, Harold Rogers Program Drug Monitoring Program, amidst concerns about growing OxyContin abuse in Appalachia in 2003. The program was named after its champion, a congressperson from eastern Kentucky, and consistent with the early focus of PDMPs as tools for law enforcement, the grant programs was administrated by the Bureau of Justice Assistance within the Attorney General's office. The $81.58 million in grants allocated to the states between 2003 and 2013 financed an explosion of development of new PDMPs and the modernization of existing PDMPs. A second grant program, the National All-Schedule Prescription Electronic Reporting Act (NASPER) of 2005 better reflects the focus of PDMP development during this period, although the program ultimately only allocated $2 millions of funding due to a lack of congressional appropriations to finance it. The NASPER program was administrated within the Center for Substance Abuse Treatment within the Department of Health and Human Services, reflecting the growing interest among physicians and other healthcare professionals in using PDMPs to identify and prevent substance abuse.

During this period states began to develop online portals that allowed registered physicians to request and receive reports on patients in real time. A handful of states developed this capability in 2003 or earlier but the federal grant programs led to a flood of systems coming

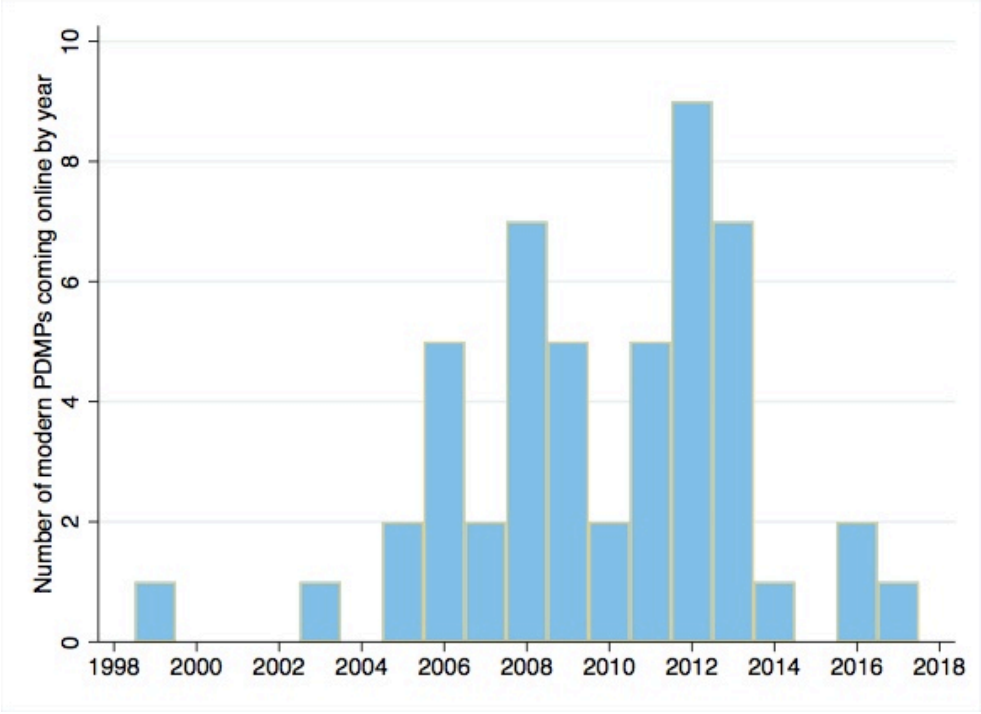


Figure 1.1: The rapid deployment of modern PDMPs

online after 2006 with 42 systems becoming operational between 2006 and 2013 (see Figure 1.1). By 2017 all but one state (Missouri) had a modern PDMP that was accessible to end users.

In the later parts of this period states began to enhance their PDMPs with new functionalities and variation in PDMP capabilities began to grow once again. For example, between 2012 and 2017, 17 states enhanced their PDMPs with mandates that either the prescriber or pharmacist dispensing the medication check the PDMP under certain circumstances. In most cases this mandate was added to an existing PDMP but in a few cases the mandate went into force as soon as the PDMP became operational (New York, West Virginia, and Pennsylvania). A few states have also started to integrate more information beyond just prescription records into their PDMP databases so as to further inform prescribers. Wisconsin and Utah, for instance, started integration of conviction records for drug violation in 2016 and some states are adding information on the form of payment for prescription drugs since cash payment is known to be associated with abuse (PDMP TTAC 2018).

## 1.3 PDMP mechanisms and identification

### 1.3.1 PDMP mechanisms

With that history in mind I now turn to conceptualizing the potential effects of PDMPs through the lens of economics. We have seen that PDMPs are designed to provide two distinct but interrelated functions: (1) accumulate information for the purposes of law enforcement and (2) provide otherwise unavailable information for healthcare providers. These two functions can modify prescribing and impact health through three channels:

1. deter illegitimate prescribing and associated diversion to secondary markets
2. curtail legitimate prescribing due to fears of potential sanctions
3. resolve information asymmetries between a prescriber and patient

In the following subsections I discuss each of these channels in turn.

#### Deterence channel

The deterrence channel is the simplest and oldest channel by which prescription monitoring can influence behavior. The Controlled Substances Act of 1970 criminalizes prescribing any narcotic except by DEA-registered prescribers "for a legitimate medical purpose" when the prescriber is "acting in the usual course of his professional practice" (21 CFR 1306.11, National Academies of Science Engineering, and Medicine 2017). The Supreme Court clarified the meaning of this vague standard in United States v. Moore (Powell 1975) listing common signs that prescribing was outside the bounds of normal practice:

1. prescribing for pain treatment without a physical exam
2. prescribing "as much and as frequently as the patient demanded"
3. charging per pill prescribed instead of for specific medical services rendered

The DEA has occasionally found evidence of even more flagrant violation of legitimate practice in a handful of cases, with doctors selling prescriptions directly in exchange for cash or sexual

favors. One U.S. attorney has gone so far as to say these doctors are simply "drug dealers in white coats" (Kanno-Youngs 2018).

Prescription monitoring puts these kinds of physicians on notice that the government is monitoring their behavior and that unusual volumes or patterns of prescribing are likely to draw attention. By raising the likelihood of sanctions, prescription monitoring raises the marginal cost of illegitimate prescribing and should deter some in equilibrium.

It remains unclear how common illegitimately prescribing is, with or without prescription monitoring, due to the difficulty gather data on such clandestine activity. By some measures it is a widespread problem with some research indicating the DEA may open as many as 1,000 investigations of indiscriminate prescribing per year (Libby 2005, Nedelman 2017). Yet the DEA only files for administration action against a handful of the prescribers investigated. ranging from a low of 88 in 2011 to a high of 479 in 2016 over the past ten years (Nedelman 2017). In addition to the uncertainty about the scale of illegitimate prescribing there is also considerable uncertainty about the significant of the marginal deterrence PDMPs provide against this type of behavior so it is hard to get a quantitative sense of whether this channel is likely to be important.

An indirect method to bound the potential scale of illegitimate prescribing is to measure its downstream effects such as diverted drug sales and use. The National Academies of Science, Engineering and Medicine (2017) examined the National Survey on Drug Use and Health (NSDUH) surveys to estimate recreational consumption of prescription opioids, finding that is likely accounts for at least 17% of opioids dispensed, by weight, but potentially much more because drug abuse is likely underreported and the NSDUH does not survey the homeless or institutionalized, including those in residual drug treatment or prison, who plausibly contribute significantly to overall drug abuse. Of the fraction diverted it also remains unclear how much was supplied from doctors who knowingly prescribed illegitimately and how much was obtained by other means including theft and deceiving careless or gullible prescribers.

**Chilling effect channel**

The second channel by which PDMPs can influence prescribing and health is closely related to the deterrence channel but has different implications for patient health and welfare so I examine is separately. This channel involves the discouragement of legitimate prescribing due to fears, warranted or not, of potential sanctions. This kind of unintended side effect is called a chilling effect in the legal literature and I adopt that term for this mechanism.

Research into chilling effects in other domains, such as libel and censorship, suggest it is most likely when surveillance or monitoring is routine and the standards separating criminal for lawful behavior are ambiguous (Johnson 1996). For example, surveillance and censorship of publications to screen out material that is "politically destabilizing" appear to chill political speech quite generally as authors cannot tell what prosecutors and courts will consider "destabilizing." Fishman (2004), for example, claims that "fear that their prescribing patterns of these heavily regulated drugs will be intensely monitored by legal authorities" prevents some physicians from specializing in pain management while another doctor, wishing to remain anonymous, told reports that he believed that for every doctor investigated by the DEA "there are a hundred doctors scared to prescribe proper pain medication for fear of going to prison" (Owen 2003).

It is easy to understand why doctors would find the standards for legitimate prescribing ambiguous. While it is obvious that a prescription sold for cash is not a "legitimate prescription" written in the "usual course" of practice, there are many common practices that remain in a gray area. Indeed, the DEA's manual for practitioners is 60 pages long with 5 pages dedicated to clarifying what counts as a legitimate prescription written under a DEA registration (DEA 2000). Even then, it remains unclear to what extent the DEA guidance is binding or even describes usual professional practice. For example, the DEA believes that prescriptions for painkillers given to patients after appointments shorter than 25 minutes are suspect, but many doctors routinely see patients in prescheduled 15 minute windows. In addition, the DEA believes a prescription for an opioid without explicit evaluation of pain outside the scope of standard practice, it appears that at least 30% of prescription are written as refills at the end of

appointments focused on other medical problems (Tisamarie et al. 2018). Congress attempted to clarify some of this ambiguity years ago in the Pain Relief Promotion Act, by adding the language "alleviating pain or discomfort in the usual course of professional practice is a legitimate medical purpose" to the CSA but the act never became law due to concerns that its vagueness could effectively legitimize so-called pill mills.

While all supply side drug policies, including the mere existence of the Controlled Substances Act could have a chilling effect, PDMPs seem likely to have particularly potent chilling effects because they involve pervasive monitoring. Indeed, one U.S. attorney recently spoke to the press about how his office "has put [doctors] on notice" because it has access to state PDMPs without a subpoena and they serve as a vital new tool in cracking down on opioid prescribing.

In light of these ambiguities it is clear why some physicians would rather be conservative, withholding medicine they believe might be beneficial for patients in order to avoid DEA scrutiny (Rankin 2018). But just as with the deterrent effect discussed above it is difficult to estimate the potential scale of a chilling effect. Survey evidence suggests it could be substantial as 38% of physicians in one survey noted that they would consider cutting back on controlled substance prescribing if a PDMP was implemented in their state (Rutkow et al. 2015). Furthermore chilling effects are not strictly limited to preventing prescriptions since pharmacies can also be sanctioned for filling illegitimate prescriptions. This mechanism may be quantitatively less important as only 13.87% of pharmacists agreed that they "would be discouraged from dispensing controlled substances if a PDMP is implemented" but is still a potential cause for concern (Fass and Hardigan 2011). That said, there are also reasons to think PDMPs will have only modest chilling effects. Most importantly, physicians are likely aware that insurance companies and data vendors like IMS Health also keep records on prescriptions that can be linked to the original prescriber. PDMPs thus probably add little marginal information beyond what exists in these disparate databases waiting to be compiled. Second, in a recent review on chilling effects of drug control policies Davis et al (2018) expressed skepticism about this channel, noting the pervasive lack of hard evidence for chilling effects in the literature, despite

widespread concern, though they went on to note that the lack of high quality research methods may be the cause and concluded that "research into these questions is urgently needed."

**Asymmetric information channel**

The third and most inherently economic effect of PDMPs is resolving an information asymmetry between providers and patients. Patients know what drugs they have prescriptions for and how many pills they have, but prescribers without access to a PDMP generally only know what they personally prescribed to a patient and whatever information the patient is willing to share. Modern PDMPs which provide real-time access to patient's prescription history can largely eliminate this asymmetry and lead to better informed patient care.[1] PDMPs are specifically helpful in identifying one type of behavior commonly associated with opioid abuse and diversion: doctor shopping. Doctor shopping involves deceiving multiple doctors in a short timeframe in order to acquire multiple controlled substance prescriptions for an excessive number of pills which are then abused for recreational use or diverted to illicit drug markets. The complaint could be genuine or faked as long as the consultations involve trying to deceive the physicians about care and prescriptions available from other prescribers. PDMPs have obvious utility in identifying doctor shoppers since doctors can see a complete record of all controlled substances prescribed to the patient in the state. They can also help identify other patients at high risk of abuse such as patients seeking an early refill, who have overlapping prescriptions, or who get other non-opioid drugs that increase the risk of overdose such as benzodiazepines from another prescriber. It is unclear how common the later situations are and, unlike with doctor shopping, less clear how doctors should response.

The prevalence of doctor shopping remains unclear, largely because of disagreement about the precise quantitative definition of doctor shopping. The simplest operational definitions are based on the number of different doctors where one obtains on opioid prescription or the number of different pharmacies where a patient fills opioid prescriptions. For example, Buchmueller and Carey (2018) use a cutoff of using five or more pharmacies and report that 0.59%

[1]Many PDMPs have the limitation that they only cover prescription within a particular state, so if patients obtained prescriptions from providers in different states this would not show up in most PDMPs at this time.

of opioid users in their Medicare sample doctor shop. Other researchers have used softer cutoffs like four or more pharmacies or four or more doctors and naturally find a greater prevalence of shopping and rates, of course, also vary with the population studied (Sansone and Sansone 2012). Some researchers have attempted to generate multidimensional measures of doctor shopping that integrate information about the number of physicians used, number of pharmacies used, as well as form of payment, age, and the types of prescriptions obtained. Carlson and McDonald (2013) use machine learning-based clustering approach applied to a convenience sample from IMS Health representing the vast majority of prescriptions in the U.S. to define two tiers of doctor shopping. Their top tier, clear doctor shoppers, account for 1.9% of all opioid prescriptions, 4% of opioids by weight, and around 5-6% of opioids by milligrams of morphine-equivalent (MME), a measure discussed more below that combines weight and potency into a summary measure. Simeone (2017) uses a similar but more conservative approach and estimates that between 2 and 3% of opioids were diverted by doctor shoppers in the 2008-2012 period and that doctor shopping waned over this time frame, declining an estimated 26%.

In contrast to the chilling and deterrence effects, the asymmetric information channel relies on doctors registering with the system and actively querying it for data. One major concern in the healthcare literature is that only a small fraction of prescribers are registered to use a PDMP so it is implausible to expect a big effect through this channel (Deyo et al. 2018). This is a serious concern and deserves a brief investigation.

A number of studies have found that only a minority of physicians use modern PDMPs in the years shortly after they become operational, but the registration and use rates vary substantially across specialties. Irvine et al (2014) looked at registration rates across specialties and practice locations and found that emergency medicine, primary care, pain medicine, and addiction medicine specialists were the heaviest users of PDMPs and that clinicians in safety net clinics and emergency departments had particularly high rates of use and registration. These patterns make sense as doctor shopping across emergency departments is easy since no patients can be turned away and the other specialties prescribes large volumes of opioids. Rutkow et al. (2015) look specifically at primary care doctors since they prescribe just slightly under half

of all opioid prescriptions and found that most primary care physicians were aware of their state's PDMP with about 53% registered. Those who register query the PDMP on average for about 1 in 23 patients or approximately half of the patients prescribed an opioid. Hafajee et al. (2015) have the most comprehensive data on registration across states and report a median registration rate of 35% for physicians in general, but this plausibly represents about a large fraction of opioids prescribed given the higher rates of registration among high-prescribing specialties noted above.

Table 1.1: Opioid prescribing rates by specialty

| | **Physicians** | | | **Opioid Prescriptions** | |
|---|---|---|---|---|---|
| **Specialty** | **number** | **share (%)** | **prescribe opioids (%)** | **per prescriber** | **share (%)** |
| Pain medicine | 3,783 | 0.51 | 83.1 | 2454.7 | 4.41 |
| Physical, rehab | 8,218 | 1.11 | 81.8 | 1028 | 3.95 |
| Rheumatology | 4,949 | 0.67 | 85.7 | 775.4 | 1.88 |
| Ortho. Surgery | 24,385 | 3.29 | 89.3 | 750.6 | 9.34 |
| Sports | 2,593 | 0.35 | 89.8 | 552.9 | 0.74 |
| General practice | 203,576 | 27.43 | 84.2 | 492.1 | 48.21 |
| Neruosurgery | 5,540 | 0.75 | 83.2 | 470.4 | 1.24 |
| Anesthesiology | 32,585 | 4.39 | 37.4 | 460.6 | 3.21 |
| Emergency | 33,375 | 4.50 | 90.3 | 429.7 | 7.40 |
| Neurology | 14,092 | 1.90 | 75.7 | 321.6 | 1.96 |
| Geriatrics | 4,518 | 0.61 | 77.1 | 275.2 | 0.55 |
| Plastic surgery | 7,914 | 1.07 | 88.4 | 237.9 | 0.95 |
| Otolaryngology | 9,588 | 1.29 | 88.5 | 226.9 | 1.10 |
| Gen. surgery | 29,965 | 4.04 | 81 | 224.9 | 3.12 |
| Urology | 10,099 | 1.36 | 88.1 | 219.9 | 1.12 |
| General, other | 6,593 | 0.89 | 83.9 | 218.3 | 0.69 |
| Heme-Oncology | 19,156 | 2.58 | 84.5 | 196.7 | 1.82 |
| Ob-Gyn | 32,794 | 4.42 | 85.8 | 166.1 | 2.67 |
| Cardiovascular | 23,274 | 3.14 | 77.2 | 51 | 0.52 |
| Pediatrics | 50,584 | 6.81 | 67.5 | 24.5 | 0.48 |
| Other | 214,716 | 28.93 | 59.3 | 1072.7 | 4.63 |

**Notes:** Author's calculations based on data from Currie and Schnell (2018) and AAMC (2018).

What fraction of doctor shopping could this share of registered physicians using the PDMP detect? The data in Table 1.1 can provide a rough calculation since the answer depends on the shares of opioids dispensed by the different specialities. If we suppose that the registration rate is 0% among physicians that do not prescribe opioids (24%), 50% higher on average in the

specialties identified by Irvine et al. (2014) and that the overall registration rate is 35% as noted by Hafajee et al. (2015), then it would be approximately 57% for active pain medicine, general practice, emergency medicine, who dispense 60.02% of opioids and make up 27.57% of physicians, 38% for other active physicians who prescribe the remaining 39.98%, for a total coverage of 49.4%. If registration within a specialty is biased toward prescribers who write more prescriptions as seems plausible then coverage would be even higher. Based on these results we should clearly not expect PDMP use to eliminate all doctor shopping and other forms or abuse it but it seems plausible to think that if doctor shopping accounts for 6% of opioids dispensed (by MME) then a PDMP effect on the order of 3% from reducing asymmetric information alone is plausible. If doctor shopping arounds for more than 6% of opioids dispensed, as it would if we include McDonald and Carlson's (2013) second tier then doctor shopping could potentially reduce prescribing by more.

In addition to preventing doctor shopping the asymmetric information channel can have other effects on prescribing. One widely acknowledged problem for physicians is that they worry about to prescribing pills that might be abused or diverted. Without a way to reassure themselves that a patient can be trusted they find that they withhold prescriptions from patients that might benefit from them. Schnell (2017), for example, developed a structural model of opioid prescribing and estimates that physicians would write 22% more opioid prescriptions to legitimate patients if they were not concerned about diversion to secondary markets. By providing an independent source of information that can verify that patients do not doctor shop, refill prescriptions early, or demonstrate otherwise aberrant behavior, PDMPs can help expand access for legitimate pain patients. While it is hard to assess the likely scale of this channel *ex ante*, there is some evidence that suggests it could quantitatively important. Weiner et al. (2013) conducted an experiment where 38 emergency room physicians at two academic medical centers were instructed to formulate a treatment plan when assessing a patient presenting with a complain of acute pain and then to review PDMP data and make adjustments if it seemed appropriate. Physicians revised plans in 9.5% of cases, adding an opioid prescription in 6.5% and eliminating a prescription in 3.0%, suggesting that not only is more liberal prescribing possible

but that it might predominate over restricted prescribing to doctor shoppers. This study might lack external validity since the participating hospitals were located in the downtown area of a city where drug diversion is a widely acknowledged problem, but it illustrates that by reducing asymmetric information PDMPs can help both by preventing high-risk prescriptions and by enabling more access to treatment for patients in need who otherwise appear too risky.

### 1.3.2 Identification

In this subsection I build on the discussion above to outline on my general identification strategy for separating out the three channels for PDMP effects. A key element of this strategy is careful attention to classifying and coding PDMPs in my data so I begin with that as a preliminary discussion of the challenges associated with that task and how attempts to resolve it motivated a key part of that strategy.

A major problem for the literature examining the effects of PDMPs is inconsistency in classifying whether a state has an operational PDMP. There are two publicly available datasets on PDMP legislation from the Prescription Drug Abuse Policy Systems (PDAPS) and the National Alliance of Model State Drug Laws (NAMSDL) and most studies utilize one of both of these studies combined with individual research to verify the dates (Horowitz et al. 2018). This has led to widespread inconsistencies across studies as illustrate in Table 1.2 which compares that dates use in four studies for a subset of states.[2]

In only 1 of the 18 states is there full agreement on the year of PDMP implementation and on average the largest subset that agrees about the year is 2.52, meaning on average there are 1.48 data sets with a different year. In addition, even in cases where there is agreement about the year of implementation there is often disagreement about the exact quarter or month within the year. For example, all of the studies agree that North Dakota's PDMP was first operational in 2007 but two studies report that it was operational in the first half of the year while the other two report that it become operational only later.

[2]One of the studies focused on a short time period where only the 18 states below made modifications so they only report dates for these states.

Table 1.2: PDMP start dates used in four studies

| state | Study #1 | Study #2 | Study #3 | Study #4 |
|---|---|---|---|---|
| Alaska | 1/1/12 | 2008h2 | 1/1/12 | 7/31/11 |
| Arizona | 12/1/08 | 2007h2 | 12/1/08 | 9/30/08 |
| Arkansas | 3/1/13 | 2011h2 | 3/1/13 | 2/28/13 |
| Delaware | 8/21/12 | 2012h1 | 8/1/12 | 1/1/11 |
| Florida | 10/14/11 | 2009h2 | 10/1/11 | 8/31/11 |
| Georgia | 5/8/13 | 2011h2 | 7/1/13 | 6/30/13 |
| Kansas | 4/1/11 | 2008h2 | 4/1/11 | 1/31/11 |
| Maryland | 1/1/14 | 2011h2 | 1/1/14 | 8/19/13 |
| Minnesota | 4/15/10 | 2007h2 | 4/1/10 | 1/3/10 |
| Montana | 10/15/12 | 2011h2 | 10/1/12 | 3/11/12 |
| Nebraska | 3/2/09 | 2012h2 | pre-2000 | 4/13/11 |
| New Hampshire | 10/16/14 | 2011h2 | pre-2000 | 9/1/14 |
| New Jersey | 1/4/12 | 2009h2 | 1/1/12 | 8/31/11 |
| North Dakota | 4/1/07 | 2007h2 | 1/1/07 | 8/31/07 |
| Oregon | 9/1/11 | 2009h2 | 9/1/11 | 5/31/11 |
| South Dakota | 3/1/12 | 2010h2 | 3/1/12 | 12/4/11 |
| Washington | 1/4/12 | 2007h2 | 1/1/12 | 10/6/11 |
| Wisconsin | 6/1/13 | 2010h2 | 5/1/13 | 3/31/13 |

Some of the disagreement is easy to understand. Three of the studies report that Florida's PDMP became operational in late 2011 but study #2 codes it as starting operation in late 2009. The discrepancy appears to be due to study #2 using a lower standard for functionality. Florida's PDMP only became accessible to doctors in 2011 but the states appropriated funds for its development in 2009 and it appears to have started collecting data on prescriptions in that year. Disagreements in coding for Nebraska are also illustrative. Nebraska had a multi-copy program since the 1990s but its program was only modernized around 2010. Study #3 thus codes Nebraska has having a PDMP prior to 2000 because of the multi-copy program while the other sources give different dates for the modernization ranging from 2009 to 2012. The basis for the disagreement about the date of modernization in that case is unclear but it likely has to do with the standards for what features are required to count as fully functional.

A team of legal scholars led by Jill Horowitz were disturbed by this problem, noting that the disagreements are understandable since "even experienced legal researchers have difficult creating consistent and reliable measure of PDMP law enactment and operation dates." There

are a large number of judgment calls that go into determining what count as an "operational" PDMP since programs vary in a large number of specific details that may or may not be considered important in the context of any particular study. In light of this the team proposed creating a dataset based on a public and detailed protocol for resolving all ambiguities and that explicitly details the features that qualify a PDMP as operational. Indeed, even this proved difficult as the team disagreed about what standard of functionality should be used so it ended up with two sets of dates capturing different criteria. The protocol and results were published in Horowitz et al. (2018) and the reader is referred to the original source for more information. I briefly review the most important results from their research that are applicable to this chapter.

One important aspect of PDMP development noted above is that most PDMPs were developed in stages, with different features coming online at different times. Horowitz et al. focus on two main levels of overall implementation. First is a stage where the PDMP is capable of data collection, maintaining a database of all applicable controlled substances prescribed in the state. To qualify as being operational for data collection the program must have systematically tracked all prescriptions for some set of drugs. In effect this is a loose criterion meant to capture the idea that prescription are being monitored with the data being pooled into a central database for later use. By this standard even the most basic early multi-copy prescription pad programs count as PDMPs with data collection. There are two main ambiguities that arise when coding the specific date for PDMP enactment with this standard. The first is that a few states required pharmacies to keep long term records of all controlled substances dispensed but did not require these to be forwarded to a central database. Horowitz et al. do not count these as PDMPs since the data will never collated a centralized data that could practically be used to support enforcement of controlled substance regulations against prescribers, although they acknowledge that they might have been useful for enforcement against pharmacies and pharmacists. The second ambiguity is that in some states legislation specifies that a PDMP will be created while separately authorizing data collection to begin on a particular date. In these cases it is possible for the system be functional but lack legal authority to begin operation until sometime in the future or, conversely, for data collection to be legally authorized but the pro-

gram incapable of actually collecting the information. In these cases the team records the date when the both data collection was authorized and the program had the capability to collect data.

The second stage of PDMP development that they collected dates for is what they call the establishment of a modern system but which I will call having user access. This functionally often came online years or in a few cases decades after the database function was developed and was difficult to code because, in practice, there is substantial variation in how end-user access is implemented. They set three criteria to ensure consistent coding. First, the system must be accessible to the entire population of end users. Many states started trial periods to test their software before making it available to all doctors. In some cases these trial periods lasted for substantial periods fo time. Virginia, for instance, started a large trial in 2003 and failed to conclude the trial and expand access to all doctors until 2006, so Horowitz et al. use the later date for when Virginia's PDMP became a modern system. The second criteria is that users must be able to query the system using the internet in, approximately, real time. This rules out some of the late phase 2 systems discussed earlier that allowed for queries by fax or that would mail discs of information. In practice this type of data retrieval was simply too slow to be a useful decision support tool and was rarely used as a result, so it would be misleading to classify these are user accessible. Finally, Horowitz et al. do not put specific requirements on which schedules must be reported but data collection must take place at least weekly in order for the system to count as up to date for end users. In practice this requirement is, to my knowledge, never binding and most systems are updated for more often, usually daily.

The distinction between the two functionalities forms the basis for a key aspect of my identification strategy. Data collection enables the deterrence and chilling effect channels but cannot resolve asymmetric information since the information was not made available to the end user. In contrast, the user access function enables asymmetric information and might have a modest chilling effect, but has no relevance to the deterrence channel. The staggered introduction of these elements of the programs provides variation that can help disentangle the different channels by which PDMPs can influence behavior. For instance, if the volume of opioids pre-

scribed to outpatients drops when data collection begins but not when user access begins then the PDMP must be having its effect through a chilling effect or deterrent effect but it could not be by reducing asymmetric information.

Unfortunately, the chilling effect could theoretically operate both when data collection begins or when user access becomes available so this strategy cannot completely distinguish the channels. Intuitively it would seem like data collection would have a larger chilling effect since the hypothesized mechanism is that knowledge of close government surveillance discourages physicians from prescribing, and data collection enables the surveillance, but user access could have a chilling effect because the origin of the chilling effect is in the ambiguity of the standards for legitimate prescribing and beliefs about this standard may be reshaped after the PDMP is available to end users. Physicians might feel they cannot prescribe "legitimately" without consulting the system, which is time consuming, or that they can more easily be held liable for "inappropriate prescribing" if they prescribe to a patient that the system shows other physicians not longer prescribe to.

In light of these concerns I supplement my data analysis by studying how PDMPs effect two more outcome variables. First, I gathered data on the prevalence of severe pain in the general population. A chilling effect must, by definition, cause harm to patients because it limits the availability of effective pain treatment. As a result it should cause a detectable increase in the prevalence of severe pain. On the other hand, the deterrence channel should have no impact on pain while the asymmetric information channel could either no effect pain patients or, by enabling more aggressive treatment of legitimate patients, even lower the severe pain prevalence. Second, I investigated the effects on patients in a context where two of the channels are inapplicable: inpatients. Hospital physicians know exactly what medications their patients are taking since the medications are directly provided, often by IV, so the asymmetric information channel is not operative. The deterrence channel is also inoperative at hospitals since hospitals only provide small quantiles of narcotics for immediate consumption and thus cannot reasonably function as pill mills. Indeed, while many states now have mandates that require PDMP use under certain circumstances, these regulations generally include a blanket exception for

Table 1.3: Summary of Identification Strategy

| | Mechanism | | |
|---|---|---|---|
| **Test** | **Deterence** | **Chilling** | **Asymmetric information** |
| prescriptions | $\downarrow$ | $\downarrow$ | $\uparrow$ or $\downarrow$ |
| pain prevalance | negligible | $\uparrow$ | negligible or $\downarrow$ |
| inpatient opioids | n/a | $\downarrow$ | negligible |
| inpatient pain | n/a | $\uparrow$ | negligible |

opioids provided in the course of inpatient care as the PDMP is not meant to be useful in those situations and querying it would only waste time. As the deterrent and asymmetric information channels are not plausibly operative at hospitals, they provide a direct test of the chilling effect hypothesis–any decline in prescribing at hospitals can only be attributed to the chilling effect. Unfortunately the data on opioid use at hospitals is quiet noisy so I also look at a parallel set of data to investigate how PDMPs affect the quality of pain management in hospitals. This is a less direct test but the data are much less noisy so it provides a much more powerful test of whether PDMPs harm inpatients. Table 1.3 summarizes the four tests and the expected effect for each channel.

## 1.4 Data

This section discusses the sources of the data used in section 1.6 for the empirical analysis. Most of the discussion involves the validity of the data and discussion of any cleaning or editing involved to prepare it for analysis. In addition I present summary statistics for each source and review key trends or patterns that are salient or otherwise noteworthy.

### 1.4.1 ARCOS

The federal Controlled Substances Act requires manufacturers and distributors of schedule II drugs and a few narcotics on other schedules, to report all transactions to the Attorney General, who has delegated authority for collecting and analyzing these reports to the DEA. The transaction reports are collected electronically and compiled in the ARCOS (Automation of Re-

ports and Consolidated Ordering System) system and are made available to the public after Freedom of Information Act requests. These data specifies how much of each type of opioid, in grams, was shipped to each type of retail distributor or consumer (e.g. pharmacy, hospital, etc.) for each state and year. The only opioid that is not monitored by this system is the tramadol, which is has a unique mechanism of action that led to it being placed on schedule V and exempted from tracking. This is unfortunate because despite its scheduling there is some evidence that tramadol can be abused and tramadol use was risen dramatically in recent years, perhaps reflecting substitution as concerns about higher scheduled drugs has increased (Jeffrey et al. 2018).

Because opioids vary dramatically in potency it is natural to reweight the masses of each opioid by its potency so that they can be compared. The standard in the literature is the convert amounts into milligrams of morphine-equivalent (MME). For example, oxycodone is thought to be about 50% more potent than morphine based on bioavailability and binding affinity so 1g of oxycodone should be considered equivalent to 1.5mg of morphine. Although there is some disagreement about the exact conversion factors to use, the CDC's estimates (Table 1.4) are an emerging standard so I adopt them for this study. I also exclude three opioids that, while potentially addictive and tracked by ARCOS, are not primarily used to manage pain, namely opium (used to treat diarrhea) and methadone and buprenorphine (for managing opioid addictions). It is particularly important to exclude the later two drugs since they can confound analysis of the effects of PDMPs. For example, if PDMPs have a deterrent effect we would expect to see smaller volumes of most opioids prescribed but potentially larger volumes of buprenorphine and methadone used as addicts might turn to providers to help manager their addictions.

One minor limitation of the ARCOS data is that it tracks distribution to pharmacies and hospitals, not to actual patients, which is theoretically the actual metric of interest, leading to potential discrepancies between distribution measured by ARCOS and the actual distribution of interest. The magnitude of these differences, however, are likely to be minor since pharmacies and hospitals generally only order drugs that they expect to and eventually do distribute to patients. A potentially bigger problem is that even if the amount distributed to pharmacies

Table 1.4: MME per mg of each opioid used in this study

| **Opioid** | **MME per mg** |
|---|---|
| Fentanyl | 100 |
| Levorphanol | 11 |
| Hydromorphone | 4 |
| Noroxymorphone | 4 |
| Oxymorphone | 3 |
| Oxycodone | 1.5 |
| Hydrocodone | 1 |
| Morphine | 1 |
| Tepentadol | 0.4 |
| Pethidine | 0.3 |
| Dihydrocodeine | 0.25 |
| Alfentanil | 0.25 |
| Codeine | 0.15 |
| Propoxyphene | 0.1 |
| Buprenorphine | not included |
| Opium tincture | not included |
| Methadone | not included |

or hospitals exactly matched the amount distributed to patients, patients still sometimes never use the drugs. This is unlikely to be a major problem in hospitals where healthcare providers often inject the drugs or watch patients take the pills, but could be a problem for the pharmacy measures. Some studies document that many opioids distributed to post-surgical patients are never consumed, at least by the post-surgical patient (Howard et al. 2018). Still, if some fraction of the opioids distributed to pharmacies are never sold or sold but never consumed, this fraction likely is similar across states and over time so it is unlikely to be a major threat to the validity of the empirical strategy discussed below.

While the ARCOS data tracks distribution to final six retail providers of drugs (pharmacies, hospitals, doctors, teaching institutions, mid-level practitioners, and narcotic treatment programs), this study focuses only on the first two as the the vast bulk of opioids included in this study are distributed to pharmacies and hospitals (98-99% each year). Practitioners and teaching hospitals make up a maximum of 1.1% of total distribution in 2010 and in most years far less and while narcotic treatment programs receive substantial fractions of the buprenorphine and methadone, as one would expected, but only trivial fractions of the other drugs.

The ARCOS data is, in principle, a census of all opioids distributed legally within the United States but one would expect that inevitably double-counting and misreporting will lead to some biases and other data quality issues. Careful inspection of the data reveals at least two outliers which I manually corrected. South Carolina's hospitals received 7,871.7g of fentanyl in 2013 but in other years averaged 616.9 grams. The 7,871.7g might be the result of a typo or an entry of the amount of milligrams in place of grams at one hospitals. I impute a value of 610.5g based on the neighboring years. South Dakota similarly has a surge in hospital fentanyl distribution in 2011, from an average of 314.2 grams in prior years to 14,451.9g in that year. I imputed a value based on the neighboring years of 480g.

It is probably not an accident that the two obvious outliers involve fentanyl data at hospitals. The year to year variation in fentanyl distribution to hospitals is much larger than for other opioids. For example, the state-level standard deviation of the percent change in fentanyl distributed to hospitals (after correcting the outliers above) is 0.3 and 0.13 for pharmacies with the analogous standard deviations for oxycodone are 0.11 and 0.12 and for hydrocodone 0.13 and 0.08. One year, 2010, in particular stands out for having large fluctuations compared to earlier years. Due to concerns about the fentanyl data quality I investigated whether there was anything particular to fentanyl that could explain why the hospital quantities appear for unstable and potential inaccurate.

There are two main differences concerning fentanyl that could serve as explanations. First, fentanyl is a synthetic opioid designed to more easily pass into the brain compared to other opioids. As a result fentanyl is about 100 times as potent as morphine so doses are generally reported in in mcg (micrograms). It is easy to imagine data entry errors where someone typed in the mcg amount but listed the units as mg, overestimating the distribution by a factor of 1000. If a hospital in a small state did this repeatedly over the course of a year then that could conceivably explain why the total quantities in South Carolina and South Dakota for the years mentioned earlier are about ten times larger than in other years. A hospital that normally receives 1% of the fentanyl would report receiving 1000% of the fentanyl for a typical year and the total amount distributed would be about ten times the state average based on that single

hospital's mistake. On the other hand, if these types of data entry errors are only occasionally made for small orders then the errors might just show up as noise, with all quantities in all years exaggerated but by varying amounts. This seems plausible but does not help explain why 2010 in particular shows exaggerated quantities in most states. Perhaps there was a change in the data entry procedures but I could not find any evidence about it.

Another issue for all opioids is that the quantities can be reported in two forms, as a salt or as an anhydrous base. Most opioids are electrically charged so are usually distributed packaged together with another molecule as a salt, like how the sodium we consume is packaged as a salt together with chloride. The most common counter ion for most opioids is, like for sodium, chloride or hydrochloric acid, which are both small so 90% of the weight of the salt (oxycodone HCl) is oxycodone so there is only a small bias from incorrectly reporting that the salt is pure oxycodone or vice versa Fentanyl, in contrast, is usually packaged with the much larger counter ion, citrate, and this fentanyl citrate salt is only 63.56% fentanyl by weight. ARCOS allows entry of either the weight of the pure opioid or the weight of the salt which it will automatically convert to the equivalent weight of the pure substance and since the conversion factors are larger for fentanyl are much larger any errors in entry of this characteristic will lead to particularly large biases. This does not explain why there is particular changes to fentanyl quantities at hospitals in 2010 but, as noted earlier, changes in reporting in 2010 might have caused this noise to be magnified. Due to concerns about the fentanyl hospital quantity quality I do robustness checks excluding fentanyl on all regressions involving hospital opioid quantities. The results are not sensitive to the inclusion or exclusion of fentanyl.

Summary statistics for the cleaned data are presented below and a few patterns are worth noting. First, the mass of MME used at pharmacies generally dwarves the mass used in hospitals–at the mean pharmacies received more than ten times as much as hospitals. Second, the amounts shipped to pharmacies and hospitals vary dramatically across states, although the variation is much larger in the pharmacy category, from a high of nearly 1.5 grams per person (Tennessee in 2013) to less than 0.2 grams per person (D.C. in 2017). Third, oxycodone makes up

Table 1.5: ARCOS summary statistics

| | N | mean | StD | min | max | p25 | p75 |
|---|---|---|---|---|---|---|---|
| MME per capita (pharmacy) | 612 | 606.2 | 191.9 | 176.6 | 1,380 | 470.9 | 711.0 |
| MME per capita (hospital) | 612 | 56.05 | 22.27 | 19.07 | 163.9 | 41.23 | 64.29 |
| MME per day (hospital) | 612 | 92.97 | 42.25 | 33.51 | 270.4 | 65.94 | 108.2 |
| oxycodone MME per capita (pharmacy) | 612 | 262.4 | 116.4 | 54.00 | 892.8 | 182.7 | 325.8 |
| hydrocodone MME per capita (pharmacy) | 612 | 108.3 | 67.70 | 7.485 | 376.0 | 61.48 | 139.6 |
| other MME per capita (pharmacy) | 612 | 235.5 | 62.95 | 61.96 | 452.5 | 192.7 | 275.3 |
| MME per capita (hospital), no fentanyl | 612 | 64.06 | 34.15 | 18.65 | 246.1 | 41.80 | 74.23 |

nearly half of the opioid MME shipped to pharmacies, consistent with intuition since it thought to be widely used both for outpatient management of chronic pain and for recreational use.

It has been widely noted that opioid consumption has been on the rise in the U.S., especially for outpatients and the ARCOS data bear out these trends. Figure 1.2 shows the opioid distribution per capita to pharmacies and hospitals per year between 2006 and 2017. The pharmacy amounts rise from around 450 MME in 2006 to just over 675 MME in 2011 then steeply decline to just slightly over 450 MME in 2017. Hospital amounts show a modest increase from 2006 to 2011 but declined in tandem with pharmacies after 2011, with particularly steep declines in 2016 and 2017. The later drop-off may have less to do with changing practice patterns and more to do with manufacturing problems involving injectable opioids. These trends help provide face validity for the ARCOS data.

Other checks on the validity of the data also look good. The distribution of types of opioids should vary across pharmacies and hospitals, with morphine and fentanyl more common in hospitals, often used for IV administration and anesthesia respectively, and oxycodone and hydrocodone, primarily packaged as pills, more common for pharmacies (see Table 1.6). The ARCOS data bears this out as pharmacies distribute 42% oxycodone, 19% hydrocodone, only 10% morphine and 22% fentanyl, with other drugs accounting for the remainder. It would have been surprising if oxycodone were not the top drug at pharmacies since it is often prescribed to outpatients in large doses (up to 160mg per tablet) while hydrocodone is usually limited to 5mg doses. Hospitals have a very different pattern with just 27% oxycodone, 11% hydrocodone,

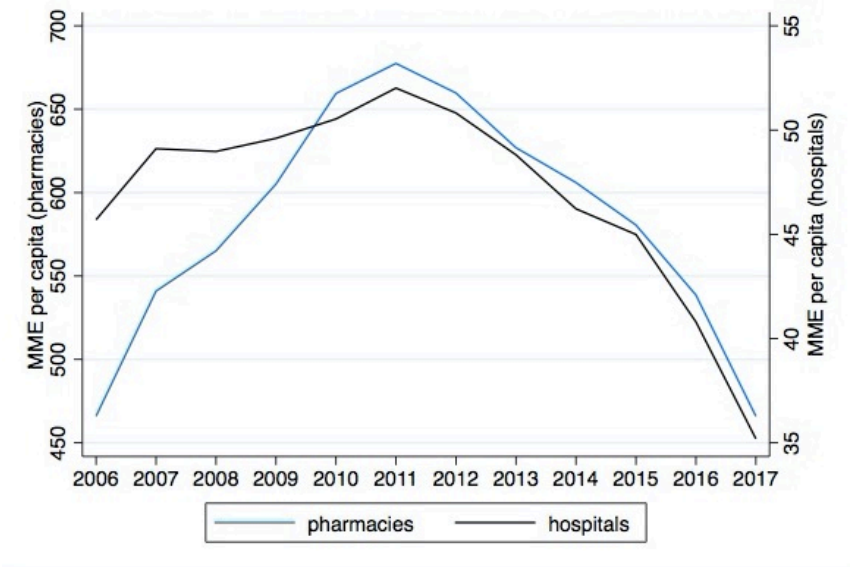


Figure 1.2: Trends in dispensing in ARCOS

Table 1.6: Opioid MME shares in Hospitals and Pharmacies

| | MME share of opioids (%) | |
|---|---|---|
| | Pharmacies | Hospitals |
| Oxycodone | 42.19 | 27.2 |
| Hydrocodone | 18.78 | 10.86 |
| Morphine | 10.05 | 17.98 |
| Fentanyl | 21.92 | 32.31 |
| Other | 7.06 | 11.65 |

18% morphine (nearly double the pharmacy fraction), and 32% fentanyl with the remainder composed of other drugs.

One major potential threat to the validity of my empirical strategy is worth discussing. The premise of investigating pharmacies and hospitals separately is that it helps to distinguish the channels by which PDMPs can exert their effects. This approach depends on the accurate coding of establishments under business code A (for pharmacies) and business code B (hospitals) in the ARCOS. This seems easy enough for most cases, but may be more difficult when hospitals contain in-house pharmacies that fill prescriptions for patients who are being discharged as well as patients seen at onsite outpatient clinics, emergency departments, and urgent care clinics. In discussions with the DEA it became clear that while it is preferred that

these hospital-based pharmacies register separately as pharmacies (class A) there is no official criteria that mandates this at the federal level since states manage this registration process. This leads to the concern that some states may simply grant hospitals with pharmacies a single hospital (class B) license and file all opioid used under that code, leading to the possibility that measured opioid use at hospitals declines after the implementation of a PDMP when in fact this is actually activity in their pharmacies that was misclassified. I investigated this issue in detail and found that is unlikely to be a problem for several reasons.

First, and most importantly, calibration calculations suggest that the per capita and per day MME used at hospitals in ARCOS are approximately what we would expect based solely on inpatient use for medical and surgical patients. My basic approach for the calibration is to estimate the amount of MME used per day by each of three groups of patients at hospitals with widely different needs and then estimate the number of days in the hospital (per capita) for each to scale up to total use. The three main categories are surgical admissions, medical admissions, and emergency department outpatients, who sometimes get medicine dispensed for pain while in the emergency department. Herzig et al. (2014) use billing records for non-surgical inpatients at 286 hospitals to document typical opioid use and variation in opioid use. They find that hospitals range from using opioids in a low of 5% of medical inpatients to a high of 72% with a mean of 51%. This variation narrows, but remains enormous, after adjusting for patient characteristics, consistent with the wide state to state variation in opioid distribution per capita to hospitals. Of the 51% of patients given opioids during their stay, the mean daily dose was 68 MME. Since surgeries are often painful and because fentanyl is sometimes used for anesthesia the opioid consumption for surgical patients is naturally much higher. Koepke et al. (2018) note that "almost all patients in the USA receive opioids during a surgical encounter" and Pizzi et al. (2012) report a median dose of 60 MME per day with a mean probably around double that, so I suppose that surgical patients average 120 MME per day. These data can be combined with MEPS estimates of admissions rates for medical and surgical services and AHA data on hospital utilization to estimate the opioid use per inpatient day.

Table 1.7: Relationship of Hospital MME and types of hospital volume

| **Dependent:** | $MME_{hospital}$ **per capita** | |
|---|---|---|
| inpatient days (per capita) | 0.0593** | 0.0414* |
| | (0.0214) | (0.0186) |
| ED visits (per capita) | -0.0012 | -0.0043 |
| | (0.0455) | (0.0278) |
| outpatient visits (per capita) | 0.00182 | 0.000932 |
| | (0.00441) | (0.00241) |
| Constant | 64.61** | 38.53** |
| | (15.92) | (12.32) |
| N | 51 | 561 |
| $R^2$ | 0.157 | 0.831 |
| Fixed effects? | n/a | state, year |
| Robust standard errors in parentheses<br>** p<0.01, * p<0.05, + p<0.1 | | |

$$MME \text{ per day} = 68(0.51)(0.45) + 120(0.55) = 81.6$$

The observed amount in ARCOS is just 79.44 MME per inpatient day leaving none of the opioid use unaccounted for.

A second, indirect, piece of evidence that the ARCOS hospital measures represent use in hospitals as opposed to distribution to outpatients is that in cross-state regressions the number of days spent in a hospital, per capita, is strongly correlated with hospital MME use per capita while number of outpatient visits and emergency department visits are not (Table 1.6). This suggests that even if the ARCOS measures include some hospital-based pharmacies by mistake these pharmacies are not used by outpatients and are instead primarily used to fill prescriptions for inpatients at discharge.

A third, related, piece of evidence is that the amount of MME used at hospitals does not predict drug overdose deaths, in contrast to pharmacy MME quantities which are a strong predictor both in cross sectional and fixed effects regressions. Figure 1.3 shows results from regressing ln(drug over deaths per capita) on ln(MME pharmacy per capita) and ln(MME hospital per capita) for each year in the sample. In most years a 1% increase in MME pharmacy per capita is associ-

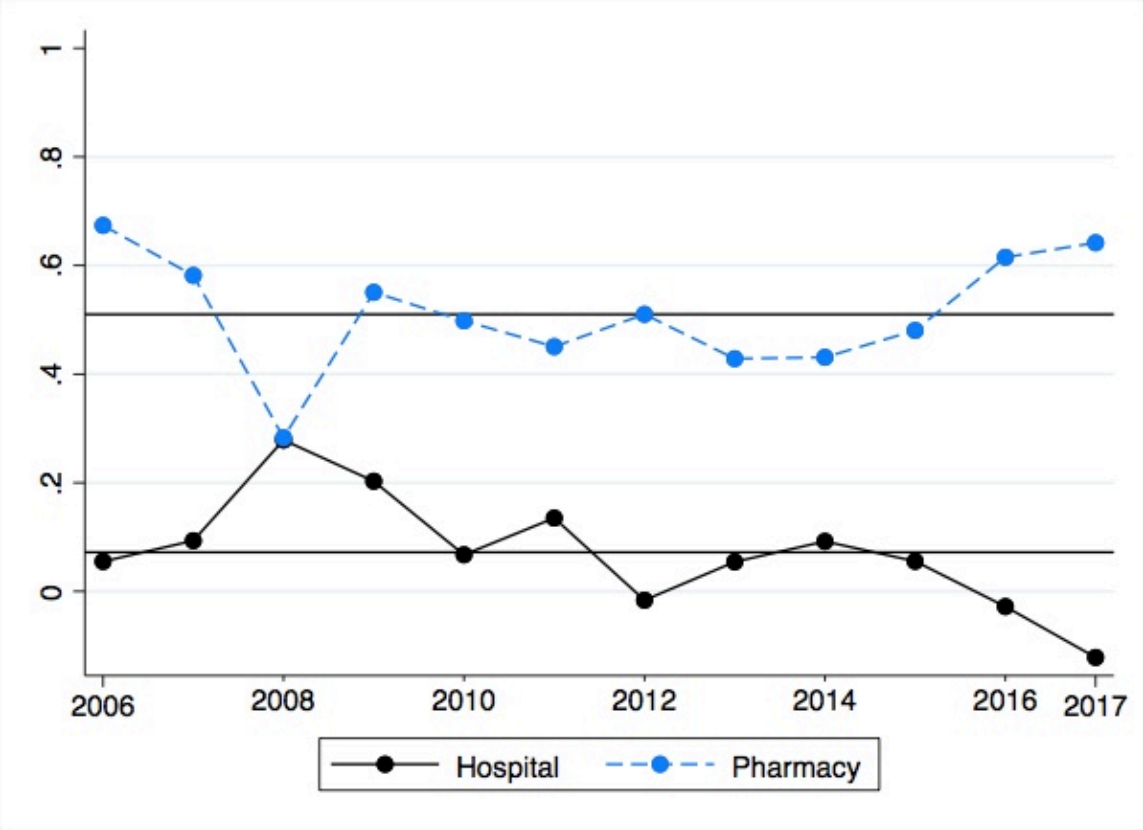


Figure 1.3: MME at pharmacies predicts overdoses but MME at hospitals does not

ated with around a 0.5% increase in drug overdose deaths while a 1% increase in MME hospital per capita is associated with a modest, insignificant increase or decrease.

Fixed effect regressions reported in Table 1.8 show a similar pattern. This fits with an interpretation where the ARCOS hospital measures MME used in the hospital and not opioids dispensed to opioids that could potentially be diverted.

In summary the ARCOS hospital data appears to accurately reflect opioid use in hospitals and not opioids distributed by hospital-based pharmacies.

### 1.4.2 HCAPHS

The Hospital Consumer Assessment of Healthcare Providers (HCAPHS) is one of two data sources I use for assessing uncontrolled pain in the population. HCAPHS is a quality reporting system for inpatient hospital care managed by the Centers for Medicare and Medicare services

Table 1.8: Relationship of opioid distribution and overdose deaths

| **Dependenat variable:** | **ln(prescription opioid deaths)** | | | | **ln(pure prescription deaths)** | |
|---|---|---|---|---|---|---|
| ln($MME_{hospital}$ per capita) | -0.035 | -0.114 | -0.0392 | -0.131 | 0.0505 | -0.164 |
| | -0.0558 | -0.0823 | (0.0562) | (0.0832) | (0.0900) | (0.119) |
| ln($MME_{pharmacy}$ per capita) | 0.6151*** | -0.0891 | 0.609*** | 0.0646 | 0.781** | 0.622** |
| | -0.199 | (0.213) | (0.198) | (0.206) | (0.386) | (0.255) |
| percent white | | | -1.680 | -1.802 | -3.728 | -1.102 |
| | | | (1.067) | (1.679) | (2.228) | (2.030) |
| rural | | | 0.124 | 0.728*** | 0.519 | 0.957** |
| | | | (0.215) | (0.264) | (0.347) | (0.391) |
| percent children | | | -0.348 | 0.338 | -1.337 | -5.597 |
| | | | (3.577) | (4.532) | (4.770) | (6.161) |
| percent seniors | | | 0.0447 | -2.979 | -9.542* | -12.03** |
| | | | (4.282) | (4.332) | (5.506) | (5.110) |
| UE rate | | | 0.264 | -1.480 | -1.171 | -6.138*** |
| | | | (1.193) | (1.716) | (2.301) | (2.188) |
| poverty rate | | | 0.599 | 4.182** | 1.626 | 7.386** |
| | | | (2.408) | (2.047) | (3.610) | (3.150) |
| Constant | -1.13 | 3.519 | 0.178 | 3.593 | 0.479 | 0.911 |
| | (1.24) | (1.328) | (1.744) | (2.731) | (4.239) | (3.250) |
| N | 357 | 612 | 357 | 612 | 357 | 606 |
| $R^2$ | 0.881 | 0.844 | 0.883 | 0.857 | 0.848 | 0.814 |
| Fixed effects? | all include state and year | | | | | |
| Years | 06-12 | all | 06-12 | all | 06-12 | all |

Standard errors clustered by state in parentheses
*** p<0.01, ** p<0.05, * p<0.1

**Note:** See text for description of how opioid overdose deaths are classified.

(CMS). Its two primary functions are to gather data about the patient experience in hospitals and, on the basis of this data, incentivize better care. In practice the incentives come in two forms, adjusting Medicare reimbursement rates and providing hospital ratings on the Hospital Compare website that can steer customers toward better rated hospitals.

The surveys have been in widespread use since 2006 after a period of development and testing by the Agency for Healthcare Research and Quality and the National Quality Forum. The survey gathers demographic information along with ratings on seven aspects of care such as treatment from nurses, experience in the hospital, and the discharge experience. Many questions ask respondents how often some unpleasant experience happened and give options for responses. The main question is interest for this study asks "During this hospital stay, how often was your pain well controlled?" with options of "always," "usually," "sometimes," and "never." Medicare bins responses of "sometimes" and "usually" together before releasing the data so in my data there are only three categories for responses.

Although surveys are used to adjust Medicare payments rates the surveys are given to a random sample of (almost) all adult patients with only a few categories for exclusion such as foreign patients and patients discharged to hospice. Some hospitals such as Critical Access Hospitals, small hospitals mostly located in rural areas, are not required to use HCAPHS surveys since they use a different payment system and cannot be penalized. Surveys are administrated in multiple available languages and multiple methods (mail, telephone, mixed methods and interactive voice response) in order to ensure representative samples, although since hospitals collect their own data there is some scope for manipulating how the data is collected in order to get an improved score.

CMS makes some adjustments to the data before its public release on the Hospital Compare website, from which is where I obtained the data. These include adjustments for case mix and demographics since certain types of patients tend to systematically give lower scores, and further adjustments based on mode of survey since mail surveys tend to lead to lower scores than telephone surveys (Lemeneh, Lerhman, and Conway 2016)). CMS believes that for reliable results hospitals need to obtain 300 or more surveys per year and the vast majority of hospitals

far exceed this threshold with a median of 589 responses across hospitals and a much larger average, but a subset of hospitals fail to meet this threshold. In my baseline analysis I focus only on acute care hospitals that have at least 300 responses in a given year, but in robustness checks I include those with fewer responses and Critical Access hospitals. It turns out that results are not sensitive to the sample.

My baseline sample spans the years 2007 to 2016. The sample starts in 2007 because hospitals only had a financial incentive to start reporting in the middle of that year. A few hospitals failed to start reporting until 2008 but since that year the sample is effectively a census since all of the nation's approximately 3,300 acute care hospitals participate in Medicare and thus report HCAPHS data to avoid financial penalties. The sample ends in 2016 because the relevant questions about pain were no longer reported after the fourth quarter of 2016 due to concerns that they were creating bad incentives for healthcare providers. Physicians for Responsible Opioid Prescribing alleged that the old question, by focusing on uncontrolled pain, had the "unintended consequence of encouraging aggressive opioid use in hospitalized patients" (Chen et al. 2016) and Adams et al. (2016) criticized it for helping to feed a misperception that patients should expect to feel no pain.

Table 1.9: HCAPHS summary statistics

| | **N** | **mean** | **StD** | **min** | **max** | **p25** | **p75** |
|---|---|---|---|---|---|---|---|
| pain uncontrolled (%) | 38,315 | 30.03 | 6.058 | 0 | 100 | 27 | 33 |
| pain SN controlled (%) | 38,315 | 7.016 | 3.414 | 0 | 96 | 5 | 8 |
| number of patients surveyed | 37,507 | 801.5 | 939.4 | 50 | 13,956 | 239.7 | 1,010 |
| acute care hospital | 38,315 | 0.756 | 0.430 | 0 | 1 | 1 | 1 |
| critical access hospital | 38,315 | 0.238 | 0.426 | 0 | 1 | 0 | 0 |
| non-profit hospital | 38,315 | 0.602 | 0.489 | 0 | 1 | 0 | 1 |
| for profit hospital | 38,315 | 0.174 | 0.379 | 0 | 1 | 0 | 0 |

I now turn to a brief summary of the key facts and trends for HCAPHS pain scores. The most salient fact, as seen in Table 1.7, is that most hospitals have good scores for managing pain. Over 99% of hospitals have a majority of respondents report that their pain is "always" well controlled and nearly 70% of patients, on average, report that their pain is "always" well controlled. Less than 10% complain that their pain was "never" well controlled with the re-

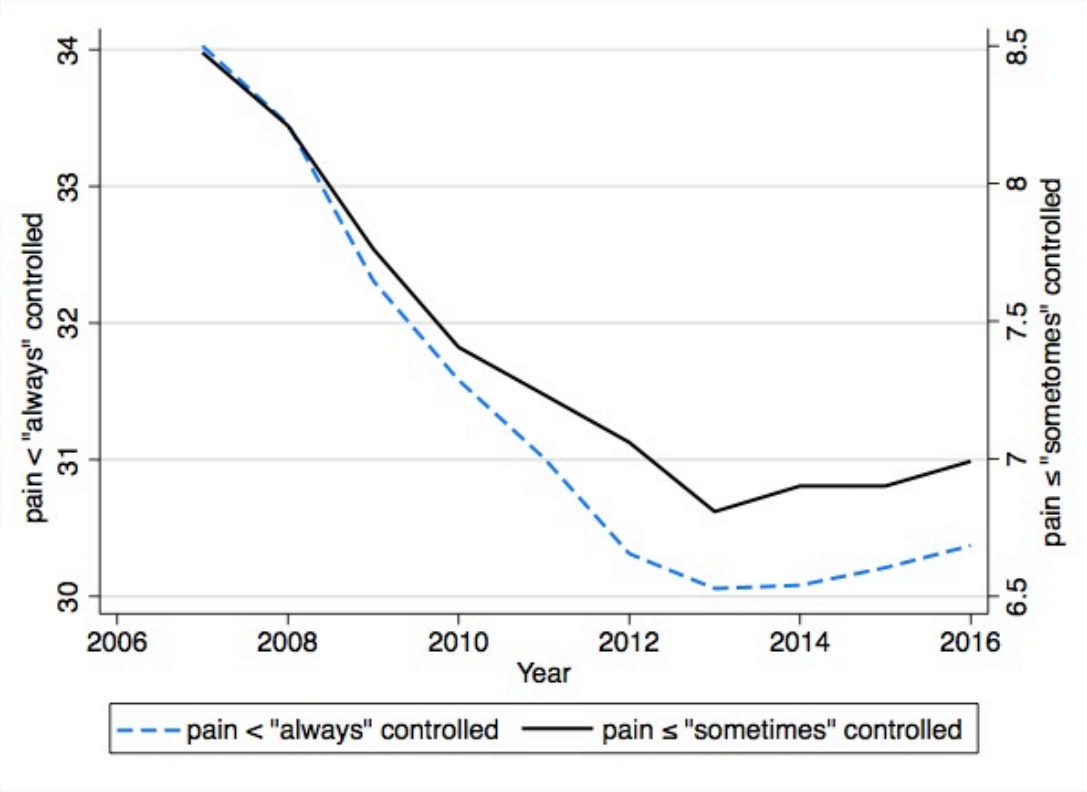


Figure 1.4: Trends in uncontrolled inpatient pain (HCAPHS)

mainder in the sometimes or usually category. Aside from a few outliers these scores do not vary that much across hospitals, with the 25th percentile for pain always well controlled at 27% and the 75th percentile only slightly higher at 33%. Responses of pain being sometimes or never well controlled show even less variation.

A second key fact, illustrated in Figure 1.4, is that scores have improved considerably over time, roughly in parallel, although progress slowed down later in the sample and even reversed in the last few years of data collection. The fraction reporting that pain was always well controlled has risen by 3.1 percentage points and the fraction reporting that pain was never well controlled dropped by 1.3 percentage points in the same time frame or 13.6% from the 2007 baseline. Scores are fairly stable within hospitals as 61% of the variance in scores for uncontrolled pain can be explained by regressions with only indicators for the hospital alone and this rises to 64.3% when year indicators are included. This implies that fixed effects regressions on

a time-year panel of the HCAPHS have high power, especially compared to the noisy ARCOS hospital data.

Finally, there are a number of statistical considerations about specification and interpretation of the HCAPHS data that each deserve a brief discussion. The first issue is that the Hospital Compare website reports data quarterly but each quarterly report includes data from the previous four quarters averaged together. At first glance it might be preferable to use the quarter as the unit of observation since it is more fine grained than by year but without knowing the sample sizes for each quarter there is no way to get a pure scores for each quarter separately. The alternative is to use the quarterly score for each quarter with the previous three quarters included but this creates severe autocorrelation in a dependent variable which leads to biased standard errors. The only feasible solution is to only use the subset of reports that have data for a particular calendar year and ignore the other data-points that mix data across years to avoid serial correlation.

A second issue is how to code uncontrolled pain. One seemingly natural option is to leave the three categories CMS reports and do an ordered probit or logit analysis, but this comes at a cost of substantial computational burden. The alternative strategy which has become standard in the literature using these data is to pick a cutoff for what counts as uncontrolled pain. For example, if the respondent says that their pain was "always" well controlled then it is coded as a 0 (not uncontrolled) and any other responses is a 1 (for uncontrolled). Alternatively, I could code "always" and "sometimes/usually" as a 0 for not uncontrolled and only consider "never" well controlled as a 1. The emerging standard is to consider any response other than "always well controlled" to be uncontrolled (Jena, Goldman and Karaca-Mandic (2016) and Jung et al (2018)) and I adopt that for comparability with other studies. I use the alternative coding in robustness checks.

A third statistical issue concerns weighting. It seems natural to want to weight the responses so that each patient's responses would contribute equally in evaluating if uncontrolled pain is getting better or worse in a state. In effect this would involve giving more weight to larger hospitals but I lack data on the actual volume of patients seen at any particular hospital.

I do have data on a proxy, the number of surveys filled out by patients at the hospital, which at first glance this seems like a good proxy, but on reflection has two major problems. First, hospitals have control over what fractions of patients they attempt to survey. This is clear from how the number of responses can vary dramatically from year to year at a particular hospital when there is no evidence it either expanded or contracted collection significantly. Second, response rates vary from hospital to hospital, in part due to differences in survey methods but likely also due to differences in quality of care that affect the likelihood of responding to a survey. Both of these concerns illustrate the the number of survey responses collected in endogenous to the quality of care, so with these problems in mind I use unweighted response in my baseline results and only conducted weighted regressions for robustness checks.

### 1.4.3 Gallup Daily Tracking

My last outcome variable and primary measure of pain prevalance comes from Gallup, Inc.'s tracking surveys from 2008 to 2016. The organization runs two surveys, the Gallup Sharecare Well-Being Index survey and the U.S. Daily survey, that had pain related questions between 2008 and 2016. These tracking surveys are conducted over the phone, drawing a sample by random digit dialing a mix of landline and mobile phones. The goal is to ensure a random sample of the non-institutionalized adult (18 years and old) American population. Weighing based on demographics to match the CPS further ensures the data are representative. Interviews are conducted with about 500 respondents each day throughout the year except on major holidays to ensure a sample of around 175,000 for each survey.

Until 2012 both surveys asked all respondents "Did you experience the following feelings during A LOT OF THE DAY yesterday: How about _____?" where the blank would be filled in with a variety of potential emotions or feelings such as "enjoyment," "worry", "stress", and, importantly for this study, "physical pain." From 2013 to 2016 only Sharecare survey asked about physical pain so samples were reduced in size but still substantial.

This question is simple to understand and transparently captures information about levels of pain in the country but it has a few undesirable features. The most obvious potential problem

is that it only captures information about the extensive margin of unmanageable or severe pain. If a respondent was in pain the day before but did not feel that they were in "a lot" of pain the question fails to capture anything about that experience. Similarly, if respondents think that they are in a lot of pain every day but that their pain has gotten worse, or better, the question is ill-suited to capture any information about those trends. That said, it seems likely that any changes on the intensive margin in pain will be reflected on the extensive margin as well. For example, if fewer people have access to opioids and as a result more people are in pain this question should detect that because some will go from having minimal amounts of pain that is managed to having "a lot" of pain and the fraction answering yes will rise.

A second potential problem is that the context in which the question is asked has varied over time as questions have been added to and removed from the Gallup Daily and Sharecare surveys. For example, between 2008 and 2013 there were other questions about recurring pain in certain areas ("In the last 12 months, have you had any of the following, or not? How about neck or back condition that caused recurring pain? . . .") It is unclear if these questions were asked before or after the question of interest for this study but if they were asked before then they might have helped prime respondents to think more about their pain, potentially making them more likely to say "yes" that they did have a lot of pain. When these questions were removed that changed the context and could thus change responses. This not a major concern, however, for the validity of this study for two reasons. First, the empirical strategy detailed below includes time fixed effects which control for time-varying factors that affects all states simultaneously. Second, there are no discontinuous jumps in the "yes" or "no" response rates from year to year, making it seems unlikely that the ordering of the questions are other contextual changes mattered significantly.

Summary statistics for the question of interest are presented in Table 1.10 below.

A few trends are worth noting. First, the samples are larger, yielding fairly precise estimates of the proportion in pain. Across all years the average sample is 5,400 respondents per state and still average 3,400 in the later years. Since the sampling is done at random the number of responses is roughly proportional to state population. As a result, the smallest samples

Table 1.10: Gallup summary statistics

| | N | mean | StD | min | max |
|---|---|---|---|---|---|
| Pain yesterday | 459 | 24.3 | 3.08 | 15.0 | 35.5 |
| **Top 5 states** | | **(average)** | **Bottom 5 states** | | **(average)** |
| West Virginia | | 33.6 | D.C. | | 17.3 |
| Kentucky | | 30.6 | North Dakota | | 19.6 |
| Arkansas | | 28.6 | Hawaii | | 20.4 |
| Alabama | | 28.3 | Minnesota | | 20.5 |
| Oklahoma | | 28.3 | South Dakota | | 20.6 |

are in the later years for D.C., hitting a minimum of 395 in 2015, and the smallest non-D.C. sample is for Delaware in 2016 with 465 responses. Another possibly surprising thing to notice is that there was only a modest upward trend in reported pain in most states and for the U.S. as a whole, from 23.45% with a lot of pain in 2008 to 24.65% in 2016. This modest increase is consistent with an aging population but might be surprising in light of a growing body of literature that suggests the U.S. is in the midst of a growing pain epidemic. Figure 1.5 shows that pain on the Gallup survey rose steadily as opioid prescribing was loosening (2008-2010), tightening (2014-2016), and roughly steady (2011-2013).

In light of this surprising trends I thought it was worth doing some tests for the validity of the Gallup data to reassure that it appears to accurately capture information about pain prevalance. The most basic test for validity is to look at which states score high and low on reported pain, as listed in Table 1.10. It is widely thought that many of the poorer, older, most obese states suffer the most from pain so we should expect that these states top the list. Indeed they do with the top states for pain including West Virginia, Kentucky and Tennessee. The states with the lowest average scores for pain are also generally unsurprising and fit with intuition as they tend to be richer and younger. That said, a few states have surprising positions such as Florida (24.9%) and Maine (25.3%), near the average despite having the largest fraction of seniors. and Washington (25.7%) and Oregon (27.2%), near the top despite being relatively healthy, affluent states, but the overall patterns fit with outside evidence and known patterns.

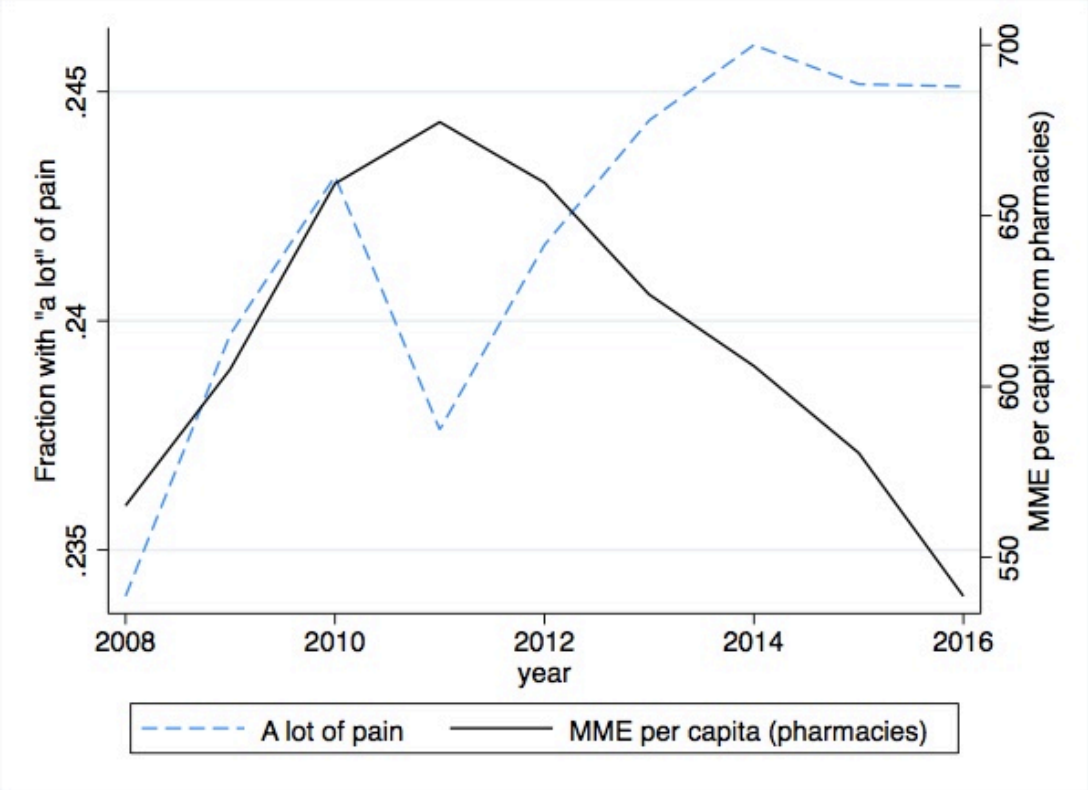


Figure 1.5: Trends in Gallup measure of Pain

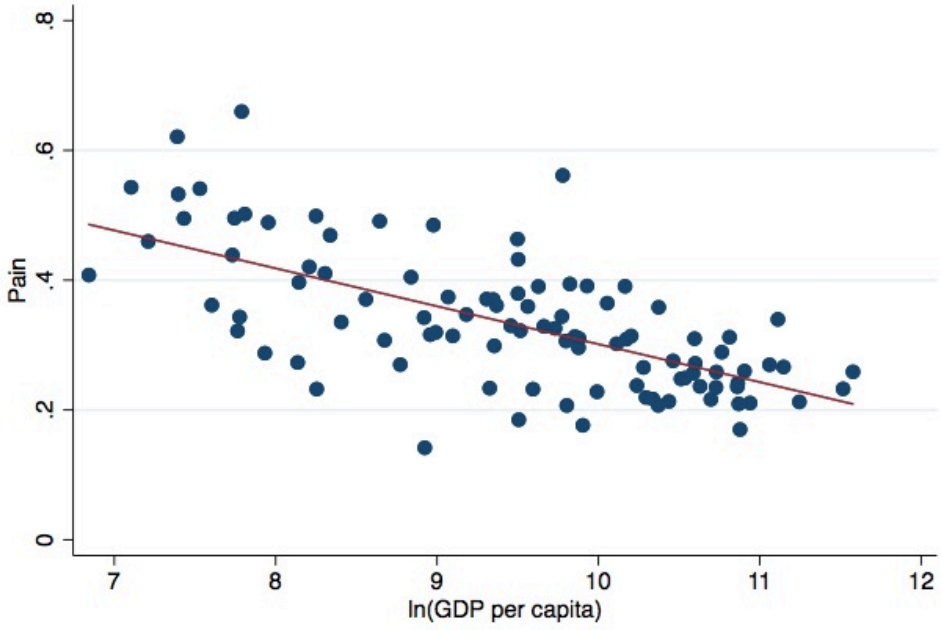


Figure 1.6: Worldwide association of GDP and pain

Another test for face validity is available since Gallup has used the same question in the Gallup World Poll since 2006. In 2018 over 142 countries were included in the World Poll so we can assess to what extent pain rates are higher in poorer countries where health tends to be worse. It has sometimes been suggested that pain surveys are biased because the respondent has to subjectively determine what counts as "severe" or "a lot" of pain and these subjective thresholds adjust based on circumstances, complicating interpretation. Blanchflower and Oswald (2019), for example, report that in a sample from the International Social Survey Program in 2011 that asked about pain, the U.S. stands out as a major outlier with by far the highest reported levels of pain, followed by other rich countries including the U.K., Australia, and Norway. It is puzzling why these countries score as having more pain than the far poorer and generally less healthy South Africa, China, Chile and Turkey.

Figure 1.6 shows the results of a regression of pain prevalence from the Gallup World Poll against log per capita income and it is clear the Gallup data does not suffer from the same problems as the International Social Survey Program. The line of best fit is $pain = -0.0584 \ln(\text{GDP per capita}) + 0.886$ with the coefficient on GDP per capita highly significant

($t = -9.79$ ) at conventional levels. This is reassuring and helps validate that the Gallup question is not subject any large or obvious bias.

### 1.4.4 PDAPS and Horowitz et al. (2018)

One of the key variables in all of the empirical specifications for this study is a dummy variable indicating whether a PDMP exists for that state and year. As discussed earlier, I use the coding standards from Horowitz et al. (2018) developed by a team of legal scholars. The dates for when the PDMP was enacted and began operational are shown in the Table 1.11 where "early" in the first column means the state had begun data collection at some point before 1990. As is evident from the table many states had PDMPs capable of data collection before my analytic data set starts (in 2006) and a few had modern systems with user access but 26 states started data collection and 40 first enabled user access during the 2006-2017 time period I study. By the end of the study period all states had data collection and all but systems except Missouri's enabled user access. The third column in the table shows the datas that states that had paper-based systems of data collection converted their systems to electronic forms of surveillance. There is arguably an important distinction between the two types of systems as electronic systems might have a larger deterrent effect, so I use these dates in robustness checks.

In addition to these dates for the key independent variable I draw dates when several other potentially confounding policies that could have impacted opioid use when implemented from the Prescription Drug Abuse Policy System (PDAPS) database. These policies suffer from some of the same issues as PDMPs in that there is some ambiguity about what counts as an implemented regulation or program, but PDAPS provides a detailed codebook on their criteria. These potentially confounding programs I include as controls are medical marijuana regulations, mandates to use a PDMP, and regulations on pain management clinics. I discuss each briefly in turn. Medical marijuana access might influence opioid dispensing rates to the extent marijuana is a substitute both as a recreational drug and as a treatment of pain. Some studies have found that when medical marijuana is available there is less demand for diverted opioids and thus less doctor shopping and fewer prescriptions (Bachuber et al. 2014, Powell et al 2018).

Table 1.11: Dates of PDMP data collection and user access

| State | Data Collection | User Access | Electronic |
|---|---|---|---|
| Alabama | Nov-05 | Apr-06 | |
| Alaska | Sep-08 | Jan-12 | |
| Arizona | Sep-07 | Dec-08 | |
| Arkansas | Mar-13 | May-13 | |
| California | early | Sep-09 | Jan-05 |
| Colorado | Jun-05 | Feb-08 | |
| Connecticut | Oct-06 | Jul-08 | |
| Delaware | Sep-11 | Aug-12 | |
| D.C. | Feb-14 | Oct-16 | |
| Florida | Dec-10 | Oct-11 | |
| Georgia | Jul-11 | May-13 | |
| Hawaii | early | Feb-12 | Dec-96 |
| Idaho | early | Apr-08 | Apr-00 |
| Illinois | early | Dec-09 | Apr-00 |
| Indiana | early | Jul-07 | Jul-07 |
| Iowa | May-06 | Mar-09 | |
| Kansas | Jul-08 | Apr-11 | |
| Kentucky | Jul-98 | Jul-99 | |
| Louisiana | Jul-06 | Jan-09 | |
| Maine | Jan-04 | Jan-05 | |
| Maryland | Oct-11 | Dec-13 | |
| Massachusetts | Dec-92 | Jan-11 | Feb-13 |
| Michigan | early | Jan-03 | Jan-02 |
| Minnesota | Jan-09 | Apr-10 | |
| Mississippi | Jun-06 | Jul-08 | |
| Montana | Jul-11 | Oct-12 | |
| Nebraska | Aug-11 | Jan-17 | |
| Nevada | Jan-96 | Feb-11 | |
| New Hampshire | Jun-12 | Oct-14 | |
| New Jersey | Aug-09 | Jan-12 | |
| New Mexico | Jul-04 | Aug-05 | |
| New York | early | Jun-13 | Oct-06 |
| North Carolina | Jan-06 | Jul-07 | |
| North Dakota | Dec-06 | Oct-08 | |
| Ohio | May-05 | Oct-06 | |
| Oklahoma | Jun-91 | Jul-06 | |
| Oregon | Jul-09 | Sep-11 | |
| Pennsylvania | early | Aug-16 | Jun-15 |
| Rhode Island | early | Sep-12 | Aug-95 |
| South Carolina | Jun-06 | Feb-08 | |
| South Dakota | Mar-10 | Mar-12 | |
| Tennessee | Jan-03 | Jan-10 | |
| Texas | early | Aug-12 | Sep-99 |
| Utah | Jul-95 | Jan-06 | |
| Vermont | Jun-08 | Jan-09 | |
| Virginia | Sep-03 | Jun-06 | |
| Washington | Aug-11 | Jan-12 | |
| West Virginia | Jun-95 | May-13 | Sep-02 |
| Wisconsin | Jun-10 | Jun-13 | |
| Wyoming | Jul-03 | Jul-13 | |

In addition, marijuana may be an effective treatment for pain and is sometimes combined with opioids in the treatment of cancer .

Some PDMPs have been enhanced with regulations that mandate use under certain circumstances. The most common mandate requires that physicians at narcotic addiction treatment programs use the PDMP when admitting a patient, but since this does not affect treatment for pain PDAPS does not count as a mandate. A handful of states have enhanced these mandates to require PDMP use during treatment or acute or chronic pain or both and some require prescriptions to check the PDMP every time they prescribe an opioid. PDAPS counts all of these as mandates as long as they apply to all outpatient practitioners under some circumstances. These mandates generally when into effect after a modern PDMP became operational so they are unlikely to confound identification of the modern PDMP effect but they do help to explain some of the variance in opioid consumption, especially in the later part of the sample and to that extent help improve the precision of the estimates.

Pain management clinics are, like sleep clinics for downers and weight loss clinics for stimulants, thought to be hubs for the mass diversion of controlled substances. They have a particularly notorious reputation in Florida as they contributed to the state's reputation as the "Oxy Express" until new regulations were issued in 2010 amidst a crackdown (Meinhofer 2015). A few other states, mostly in the south, added new regulations on pain clinics during the sample period in order to limit abuse and this could plausibly directly affect opioid dispensing rates from pharmacies so it is important to include them as a control.

The table below shows summary statistics for all of the variables coding for laws or regulations.

There are 612 observations for each since each of the 51 states (including D.C.) are followed from 2006 to 2017 giving 12 observations each. About 81.2% of the observations have a PDMP enacted while 64.4% have a modern PDMP. Mandates were only enacted in a handful of states and mostly in the later period so only 8.5% of observations have mandates and pain clinic laws are similarly rare. Marijuana laws are somewhat more common but were still only in force in about a quarter of the state-year pairs in the sample.

Table 1.12: PDAPS and Horowitz Data - summary statistics

| | N | mean | StD | min | max | p25 | p75 |
|---|---|---|---|---|---|---|---|
| PDMP - Data Collection | 612 | 0.853 | 0.354 | 0 | 1 | 1 | 1 |
| PDMP - Electronic Data | 612 | 0.825 | 0.380 | 0 | 1 | 1 | 1 |
| PDMP - User Access | 612 | 0.644 | 0.479 | 0 | 1 | 0 | 1 |
| PDMP - Mandate | 612 | 0.114 | 0.319 | 0 | 1 | 0 | 0 |
| Pain Clinic Law | 612 | 0.0866 | 0.281 | 0 | 1 | 0 | 0 |
| Marijuana Law | 612 | 0.289 | 0.454 | 0 | 1 | 0 | 1 |

### 1.4.5 Other data sources

I used a small number of other sources for data on population, demographics, and hospital utilization in order to construct per capita and per hospitalization measures of opioid consumption as well as control variables. These are all well documented and widely used data sources I only briefly discuss them.

For year by year state population estimates I used the Census Bureau projections as of June 2018. These include intercensal estimates for the population of each state as of July 1 for each year from 2006 to 2009 which use the 2000 Census as a basis and intercensal estimates for 2010 to 2017 which use the 2010 Census as a basis.

I gathered state level data on hospital utilization including inpatient days per 1,000 people, hospital admissions per 1,000 people, and outpatient visits per 1,000 people in order to construct estimates of opioids dispensed per day to inpatients. These data are publicly available through the Kaiser Family Foundation website for the relevant years and are based on original data collected by the American Hospital Association through an annual survey of members which are combined with Census population estimates.

Finally, I gathered data on demographic controls from two sources. First, I gathered data on age, race, ethnicity and poverty rates through the IPUMS USA program which provides access to the underlying American Community Survey (ACS) data and collapsed these into state by year sets of controls (Ruggles et al 2019). Second, I gather state level unemployment rates from the Bureau of Labor Statistics which are publicly available on its website.

Table 1.13: Minor data sources - summary statistics

| | N | mean | StD | min | max | p25 | p75 |
|---|---|---|---|---|---|---|---|
| population (millions) | 612 | 6.127 | 6.885 | 0.522 | 39.54 | 1.643 | 6.841 |
| inpatient days (per 1000) | 612 | 641.3 | 205.8 | 336 | 1,674 | 512 | 697.5 |
| rural (share) | 612 | 0.197 | 0.169 | 0 | 0.670 | 0.0580 | 0.276 |
| white (share) | 612 | 0.724 | 0.155 | 0.328 | 0.969 | 0.610 | 0.841 |
| black (share) | 612 | 0.117 | 0.109 | 0.00498 | 0.560 | 0.0381 | 0.156 |
| Asian (share) | 612 | 0.0458 | 0.0727 | 0.00633 | 0.536 | 0.0173 | 0.0477 |
| Hispanic (share) | 612 | 0.107 | 0.0994 | 0.00681 | 0.487 | 0.0396 | 0.123 |
| children (age < 18) | 612 | 0.234 | 0.0212 | 0.167 | 0.314 | 0.222 | 0.245 |
| seniors (age >= 65) | 612 | 0.140 | 0.0205 | 0.0685 | 0.201 | 0.129 | 0.153 |
| UE rate | 612 | 0.0711 | 0.0230 | 0.0234 | 0.149 | 0.0543 | 0.0857 |
| poverty rate | 612 | 0.165 | 0.0326 | 0.0988 | 0.272 | 0.141 | 0.188 |
| admissions | 561 | 111.5 | 24.41 | 69 | 238.3 | 93 | 125 |
| inpatient days | 561 | 641.3 | 204.8 | 336 | 1674 | 512 | 697.5 |
| outpatient visits | 561 | 2426.1 | 894.8 | 941 | 5891 | 1734 | 2959 |

Summary statistics for these data sources are presented in Table 1.13. A few clarifications are worth mentioning. The rural population share is the fraction of the population that lives outside an urban area or urban cluster. Urban clusters can have as few of 2,500 people so the majority of the deep South is rural by this standard although in common parlance might be considered mostly rural. The rural population shares are only significant in the Mountain West and parts of the midwest dominated by corn farms, areas that have opioid consumption far below average. In addition the last three variables which report on hospital utilization are per 1,000 people so, for example, the mean state averages 0.11 admissions per capita although this varies considerably across states.

## 1.5 Empirical Strategy

The basic empirical strategy used in the following section is a standard difference-in-differences framework, viewing the introduction of prescription monitoring as a natural experiment. The key assumption is that, while the states have widely varying pre-PDMP levels of opioid consumption and reported pain, they would have followed a parallel trend in the absence of the new programs. To some extent this assumption can be assessed by looking at pre-trends, as

shown in figures in the following section, but it cannot be tested directly since we do not know what would have happened in the absence of the PDMP.

The basic regression specifications report in tables all take the form

$$Y_{it} = \alpha_i + \beta_1 \text{PDMP-DC}_{it} + \beta_2 \text{PDMP-UA}_{it} + \gamma X_{it} + \delta_t + \varepsilon_{it}$$

where $i$ indexes the state, $t$ indexes the year, $Y$ is some response variable (e.g. MME of opioids sent to pharmacies), $PDMP_{it}$ is a dummy for the existence of a modern PDMP in that state and year, $\delta$ is a year fixed effect, $\alpha_i$ is a state fixed effect, $X_{it}$ is a vector of controls that varies across specifications and $\varepsilon_{it}$ is the error term.

In addition to the specification above I use the following non-parametric event study specification below to assess pre-trends and as a check on the quality of evidence when the basic specifications show significant effects:

$$Y_{it} = \alpha_i + \sum \beta_{1d} \text{PDMP-DC}_{id} + \sum \beta_{2d} \text{PDMP-UA}_{id} + \gamma X_{it} + \delta_t + \varepsilon_{it}$$

Here $d$ is a new index that measures the time since implementation of the modern PDMP. For example, $d = 0$ in the year of implementation, then $d = 1$ in the following year, and $d < 0$ represents the years leading up to implementation. Under the identifying assumptions we should see that $\beta_{nd} < 0$ when $d < 0$ since the PDMP features could not possibly have had effects in the years prior to implementation. If we see significant values for $\beta_{nd}$ or a trend up or down that would suggest that the assumption of parallel trends is likely to be violated since it did not hold in the period prior to implementation.

There are four main substantive choices particular to my data when using these basic specification and in most cases there is not a clearly right approach from among the options. First, should the opioid quantity dependent variables be in raw amounts or per capita amounts, and second, should the dependent variables be modeled in logs or levels when, as with all opioid quantity amounts, the distribution is highly skewed due to some states having extreme high consumption. If the amounts are in levels it seems obvious the per capita amounts are strongly

preferred since it is otherwise difficult to control for variation in the size of the population. If the dependent variables are in logs, however, then log population can be included as a control variable and the coefficient on PDMP has a similar interpretation for the raw log quantity or the log quantity per capita, it is roughly the percentage change in opioid consumption. Both approaches are used in the literature (Pardo 2017, Kilby 2015, and Dowell et al 2016b for examples), but I prefer to use log per capita quantities as my base specification with levels per capita reported as validity checks. The logged dependent variables have a more symmetric, normal distribution so the standard errors have a better behaved (Stock and Watson 2007). In addition, the logged dependent variable is more likely to yield less treatment effect heterogeneity. As discussed earlier, modern PDMPs are designed to reign in doctor shopping, illegitimate prescribing, and other forms of abuse. While the rates of these behaviors vary from state to state the proportion of opioids that are abused may not vary much so it plausible to hypothesize effects on the order of 0.05 log points that are fairly heterogenous. In contrast, the level effects could vary from around 50 MME per capita in a state like Tennessee or Delaware where MME per capita is often over 1,000 to as little as 20 MME in the less heavy using states like Illinois or Minnesota. From one point of view the heterogeneity of expected effects might not be that important since the difference in differences specifications are meant to estimate an average treatment effect, but new research has shown that under conditions of heterogeneous treatment effects the weighted average treatment effect this approaches estimate can be difficult to interpret. The difficulty gets worse the more heterogenous the treatment effects are since, intuitively, the distribution of weights matters more when the underlying observations are more variable.

A third substantive issue is how to code the PDMP-DC$_{it}$ and PDMP-UA$_{it}$ variables during the implementation years. If Florida's PDMP enabled user access in October 2011 does that mean it had a PDMP for the year 2011? If not, what about South Dakota, which enabled user access in March of 2012, so had user access for most of the year? The simplest approaches are to code them all as 1 as long as the PDMP feature was implemented at some point in the year, which is my approach, or to code them all as 0 unless the PDMP feature was operational for the

entire year. A more complex approach would be to require the PDMP to be in place for some minimal fraction of the year, perhaps half. In practice this coding issue is unimportant except in one case discussed in more detail in the results section.

The final substantive issue is what controls to include in the vector $X$. There are a variety of controls that seem like obvious candidates for inclusion since they are likely to influence reported pain, opioid consumption, or both. These include the fraction of the population that are seniors, the unemployment rate, and the fraction that lives in a rural area. All of these are strong predictors of either pain or opioid use, or both, in the cross section, but it turns out that in fixed effects specifications they tend to be less important because they do not vary that much within a state from year to year, so the residual variation is small and unimportant in explaining year to year fluctuations in opioid consumption or pain. For the most part they tend not to be significant when included in my main specifications. Still, to the extent controls help to eat up variation in the residuals they improve the power of the tests so I include all available controls in my regressions.

A final important issue before turning to reporting the results is how to construct standard errors. While the difference-in-differences approach has gained acceptance as yielding unbiased coefficient estimates of treatment effects under the parallel trends assumption, it is now widely accepted that the standard errors must be constructed carefully to avoid excessive rejection rates. Bertrand, Duflo, and Mullainathan (2004) emphasized the observations within a state or within a time period will often have correlated error terms. As a result it is important to do some kind of clustering to avoid having underestimated standard errors. Further research has clarified the conditions under which clustering matters and given practical guidance on how and when to cluster. Cameron and Miller (2015), summarizing prior research, note that a state-year panel like that used in this study is the classic case where errors are correlated within a state. While the state fixed effects eliminate the possibility that any state could have high conditional opioid consumption in all years because it cannot be above (or below) its on average consumption in all years, this does not eliminate the possibility of serial correlation across a few years. If Tennessee had high opioid consumption conditional on covariates and the fixed

effect in 2010 then it is likely to have high conditional opioid consumption again in 2011 because whatever was anomalous about 2010 is also plausibly anomalous about 2011, perhaps lax law enforcement after election of a new government or evolving demographics. Similarly, the regressors of interest themselves, the PDMP dummies, are obviously serial correlated since they are a series of 0s followed by a series of 1s. Under these conditions it is vital to cluster the standard errors by state or, they show, the standard errors can be underestimated severely. For this reason, all standard errors estimated below make some attempt to cluster by state.

Cameron and Miller go on to note the different methods available for clustering. The simplest is to use Liang and Zeger (1986) estimator for the covariance which is consistent as the number of clusters goes to infinity. The practical problem is that the number of clusters is small (51 or less) in state-year panels so Cameron and Miller suggest a finite sample correction which inflates the standard errors slightly. This is generally accepted as reliable in the literature for cases, like mine, where there are more than 20 clusters so these are the basic standard errors I report in regression tables in section 1.6.

Another approach, which can potentially be more efficient, is bootstrapping. In theory, bootstrapped clustered standard errors provide an asymptotic refinement on the Liang and Zeger standard errors, so they will tend to be smaller with sufficiently large samples. When they fail to be smaller they help, correctly, inflate the standard errors for the failure of the distribution to approach its theoretical limit distribution in the finite sample. The basic idea is to creates random resamples of clusters which can be used to estimate a distribution of t-statistics for the coefficients under the null hypothesis of no effect. A p-value is then constructed by looking at where the original regression's coefficient falls within this bootstrapped null distribution. Roodman et al (2018) discuss some practical computation problems that arise when doing the cluster wild bootstrap in Stata and provide software to solve these challenges and perform diagnostics on problems that can arise, such as when one influential cluster leads to an uneven ("lumpy") null distribution. I use the Roodman et al. Stata program to calculate bootstrapped p values for both the PDMP-DC$_{it}$ and PDMP-UA$_{it}$ regressors at the bottom of results tables. These p-values are often smaller than the standard finite-sample adjusted clus-

tered SEs suggesting that the samples are sufficient for reasonably close convergence of the distributions to their theoretical limit distributions.

## 1.6 Results

I now turn to presenting the main results from the empirical investigation, placing emphasis on both the statistical significance of estimates and their economic significance. In many cases I find that the effect of the PDMP effects are statistically insignificant but modern econometrics has reemphasized the importance on considering to what extent the results leave open the possibility of economically significant effects. For example, I find that the effect of PDMP data collection on the distribution of opioids to hospitals is insignificant, but it could be that a lack of power drives this result and that economically significant reductions in use are taking place, so the follow-up question is to ask what is the largest reduction in distribution that we can rule out. In many cases this kind of analysis is informative since we can not only conclude that effects are insignificant but also that economically important effects can be ruled out as well. In some cases however the results are more ambiguous and further research is clearly called for.

### 1.6.1 Effect of modern PDMPs on opioid distribution to pharmacies

I estimated the difference-in-differences specification from section 1.5 with $\ln(\text{MME per capita})$ dispensed to pharmacies as the dependent variable and the results are reported in Table 1.13. The first two columns show the effects of data collection and user access entered independently while the later two columns show the general specification, without weights (third column) and weighted by population (fourth column). The results are consistent across specifications: user access reduces opioid shipments by about 5 log points or 5% and this effect is significant at conventional levels using the state-level clustered standard errors with finite sample adjustments. The bootstrapped standard errors are smaller, suggesting that the data fit the assumptions for clustering and that as a result that bootstrapping only serves as a modest asymptotic refinement. In contrast, data collection does not have a significant effect with small, though

negative, point estimates in all specifications. In terms of the channels for PDMP discussed earlier this evidence is consistent with the asymmetric information channel for PDMP being important and potentially with a chilling effect, but inconsistent with the deterrence channel.

A natural follow up question to ask if to what extent I can rule out economically important effects of data collection. Although the data collect effect is not significant, the point estimates are negative and indicate a 1-2% decrease in opioid MME prescribed. The standard errors on the PDMP-DC coefficients, however, are small, just 0023 in the combined specification so the upper bound on the confidence interval for the estimate effect is 0.0678. Intuitively a 0.0678 log point or about 7% decrease in outpatient opioid usage seems small and unlikely to representa large chilling effect as this coefficient captures the effect of both chilling and deterrence. On the other hand, this reduction might be concentrated into a small subset of users who might be particularly harmed as a result. For example, if the 7% reduction represented a chilling effect on prescribing of hydrocodone combination products for acute injuries it harm a large number of patients, although only for a short time period each. In order to assess this possibility I estimate the effect on pain prevalence in section 1.6.2.

Since the literature on PDMP effects has produced a wide range of estimates I think it is important to try to assess the validity of the assumptions behind the difference in difference analysis so I estimated the event study specification and plotted the coefficients for PDMP-DC in Figure 1.7 and for PDMP-UA in Figure 1.8. Both graphs show no evidence of pre-trends which is reassuring. Consistent with the pre-post regressions in Table 1.14 the PDMP-DC event study shows little evidence of a real effect. Although most of the estimated coefficients are negative the time pattern of effects is more consistent with noise than a genuine effect. Taken at face value the point estimates suggest that there is no deterrent of chilling effect in the year of implementation, then a large effect appears in the second year which slowly fades out over time. If anything we would expect deterrence to grow over time as law enforcement illustrates the effectiveness of the system, and a chilling effect also seems unlikely to fade as the country has become more hostile to opioid prescribing.

Table 1.14: Effect of PDMPs on opioids dispensed to pharmacies (logs)

| **Dependent:** | **ln(MME per capita from pharmacies)** | | | |
|---|---|---|---|---|
| PDMP - DC | -0.0345 | | -0.0227 | -0.0141 |
| | (0.0239) | | (0.0230) | (0.0216) |
| PDMP - UA | | -0.0532** | -0.0493** | -0.0478* |
| | | (0.0182) | (0.0174) | (0.0196) |
| PDMP mandate | -0.0366 | -0.0313 | -0.0316 | -0.0817* |
| | (0.0351) | (0.0357) | (0.0350) | (0.0349) |
| pain clinic law | 0.0303 | 0.0315 | 0.0344+ | 0.0535** |
| | (0.0198) | (0.0193) | (0.0190) | (0.0151) |
| marijuana law | -0.078 | -0.079 | -0.0830 | -0.184 |
| | (0.156) | (0.151) | (0.150) | (0.217) |
| rural (%) | 0.990 | 0.740 | 0.707 | 2.708 |
| | (1.107) | (1.082) | (1.047) | (1.871) |
| white (%) | 2.644 | 2.814 | 2.895 | -1.318 |
| | (2.282) | (2.343) | (2.343) | (2.841) |
| children (%) | -0.427 | -0.373 | -0.315 | -4.329 |
| | (2.221) | (2.267) | (2.225) | (2.929) |
| seniors (%) | 1.362 | 1.516+ | 1.479 | 2.142 |
| | (0.914) | (0.901) | (0.887) | (1.707) |
| UE rate | -0.599 | -0.449 | -0.467 | -1.016 |
| | (0.844) | (0.828) | (0.813) | (1.224) |
| povety rate | -598.1 | -489.1 | -497.3 | -954.6 |
| | (676.4) | (668.0) | (661.1) | (911.7) |
| Constant | 4.935** | 5.029** | 5.045** | 5.283** |
| | (1.103) | (1.082) | (1.078) | (1.235) |
| N | 612 | 612 | 612 | 612 |
| $R^2$ | 0.954 | 0.955 | 0.956 | 0.959 |
| Fixed effects? | all include state and year | | | |
| weight | n/a | n/a | n/a | pop |
| bootstrap p (DC) | 0.188 | | 0.348 | 0.506 |
| boostrap p (UA) | | 0.003 | 0.005 | 0.0180 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

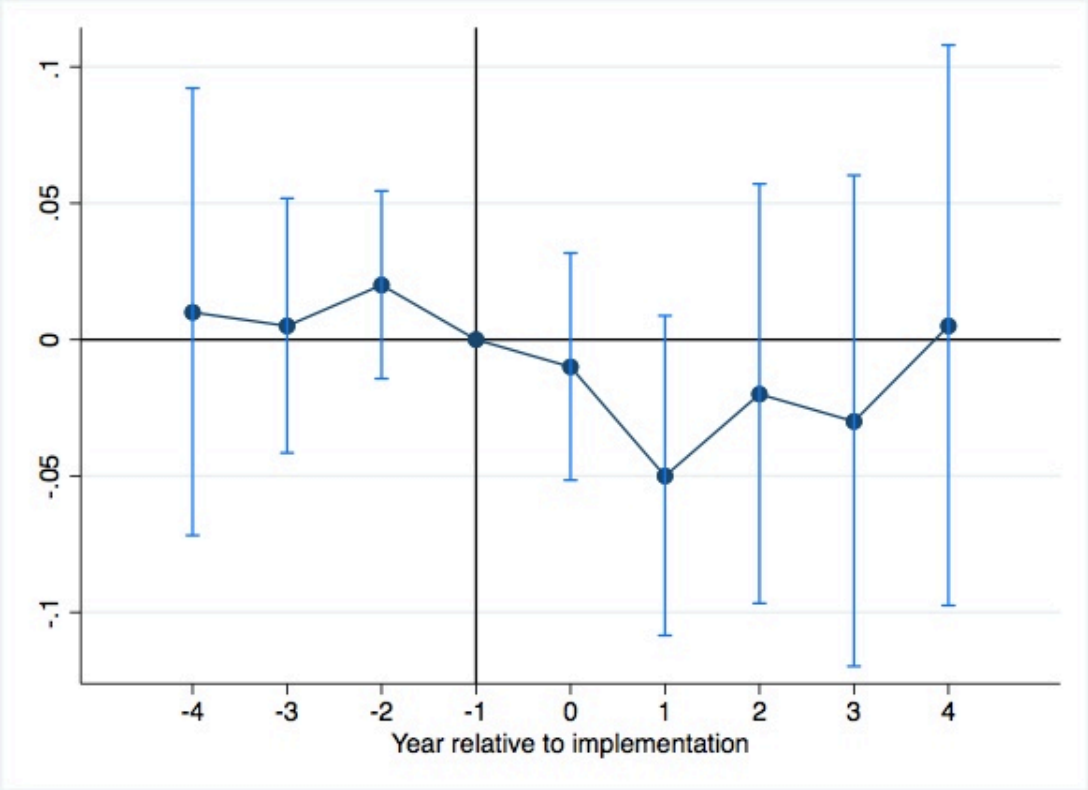


Figure 1.7: Event study for PDMP-DC on ln(MME pharmacy per capita)
**Note:** The dots are coefficients from the event study specification described in the text. Year zero is the year in which the PDMP began collecting data.

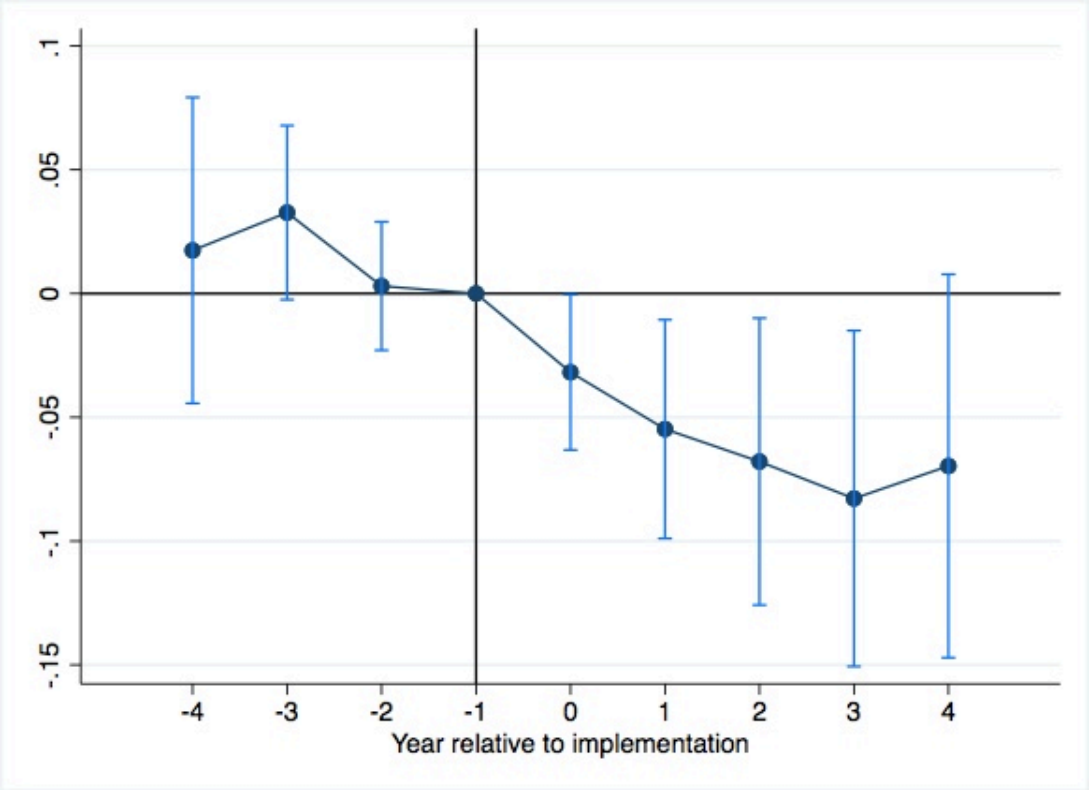


Figure 1.8: Event study for PDMP-UA on ln(MME pharmacy per capita)

The event study diagram for PDMP-UA on ln(MME per capita) also tells a clear story. There is no evidence of pre-trends and then modest effect in the first year of user access which grows over time, presumably because more physicians register to use the system and it becomes easier to detect doctor shopping and other forms of abuse.

I reestimate the effects of PDMPs on the level of opioid dispensing (from pharmacies) per capita and the results are reported in Table 1.15. Consistent with the prior results, in these specifications the user access effect remains highly significant at conventional levels, reducing pharmacy opioid volume by about 39 MME per person per year. The data collection effect remains insignificant and, in the preferred specification, the point estimate is a 10 MME per person per year reduction in MME, or the equivalent of two tablets of hydrocodone per capita.

Another robustness check that can help validate that the asymmetric information channel is operative is to look at the effect on the effects on specific opioids. Most doctor shoppers focus on getting tablets and oxycodone, in particular, is known to be the most popular drug

Table 1.15: Effect of PDMPs on Opioids dispensed to Pharmacies (levels)

| **Dependent:** | **MME per capita (pharmacy)** | | | |
|---|---|---|---|---|
| PDMP - DC | -19.31 | | -10.34 | -2.066 |
| | (14.04) | | (13.79) | (18.09) |
| PDMP - UA | | -39.34** | -37.54** | -44.18* |
| | | (13.24) | (13.03) | (19.42) |
| PDMP mandate | -31.78* | -33.42* | -35.32* | -16.94 |
| | (15.64) | (15.63) | (15.89) | (23.76) |
| pain clinic law | -19.61 | -15.66 | -15.78 | -56.60* |
| | (27.97) | (27.85) | (27.52) | (26.34) |
| marijuana law | 10.85 | 12.68 | 14.01 | 23.77* |
| | (13.65) | (13.15) | (13.15) | (10.36) |
| rural (%) | -206.9* | -208.9* | -210.8* | -381.6* |
| | (93.13) | (88.51) | (88.40) | (144.1) |
| white (%) | 645.6 | 444.8 | 429.7 | 1,950 |
| | (672.4) | (631.6) | (613.7) | (1,501) |
| children (%) | 1,747 | 1,901 | 1,937 | -1,276 |
| | (1,562) | (1,570) | (1,570) | (2,180) |
| seniors (%) | -407.6 | -349.0 | -322.3 | -3,683 |
| | (1,376) | (1,389) | (1,367) | (2,364) |
| UE rate | 1,557* | 1,662* | 1,646* | 2,498 |
| | (774.6) | (751.6) | (749.1) | (1,652) |
| povety rate | -598.1 | -489.1 | -497.3 | -954.6 |
| | (676.4) | (668.0) | (661.1) | (911.7) |
| Constant | -309.8 | -233.2 | -225.7 | -20.51 |
| | (623.4) | (605.9) | (609.2) | (712.5) |
| N | 612 | 612 | 612 | 612 |
| $R^2$ | 0.923 | 0.925 | 0.925 | 0.922 |
| Fixed effects? | all include state and year | | | |
| weight | n/a | n/a | n/a | pop |
| bootstrap p (DC) | 0.161 | | 0.465 | 0.903 |
| boostrap p (UA) | | 0.002 | 0.003 | 0.029 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

of abuse (GAO 2003, Han et al 2015). As a result we should expect to see larger effects of PDMP user access on oxycodone shipments and potentially hydrocodone shipments, since it is distributed almost solely in tablet form, compared to other opioids which come in a mix of tablets, lozenges, patches, and syrups. Table 1.16 shows difference in difference specifications where the dependent variable is the logged quantity of the specified drug per capita shipped to pharmacies.

As expected the data collection effects remain insignificant, although still negative, while the user access effects are significant and negative for the expected drugs. The oxycodone effect, in particular, is somewhat larger than the average effect and the effect on hydrocodone is also significant, although imprecisely estimated. The point estimate for the effect on other drugs indicates a reduction of about 2.5%, a much smaller and insignificant effect. One thing worth noting, although it is beyond the main focus of this research and should be validated in further study, is that mandates for PDMP use appear to have an especially large effect on hydrocodone (almost 14 log points) and other opioid dispensing (almost 8 log points), but have little effect on oxycodone. This is concerning because it suggests that mandates may be restricting access for legitimate patients and doing a poor job reducing prescriptions that are more likely to lead to drug abuse and diversion.

I present a few more robustness checks in Table 1.17, looking at the robustness of the user access effect on subsamples and with the inclusion of time trends. The first column leaves out Florida, a state that had a particularly severe opioid diversion problem and was the main target of DEA raids on "pill mills" in 2010, around the same time its PDMP was developed (from 2009 to 2011). Since I do not have a specific control for DEA enforcement including Florida in the sample could be biasing the results by loading some of the effect of the DEA raids onto the PDMP-DC or PDMP-UA coefficients. The results without Florida help to validate that this is not a major concern, with the coefficients largely unchanged, the data collection effect negative but insignificant, and the user access effect significant only modestly smaller than in the baseline model. The second column restricts the sample to states that changed their PDMP during the sample period (2006 to 2017). This leaves out states that were early adopters of

Table 1.16: Effect of PDMPs on Speciifc Drugs dispensed to Pharmacies

| **Dependent:** | **ln(MME per capita from pharmacies)** | | |
|---|---|---|---|
| **Drug:** | **oxycodone** | **hydrocodone** | **all other** |
| PDMP - DC | -0.0227 | -0.0339 | -0.0158 |
| | (0.0230) | (0.0408) | (0.0231) |
| PDMP - UA | -0.0593** | -0.0416* | -0.0252 |
| | (0.0174) | (0.0270) | (0.0187) |
| PDMP mandate | -0.0286 | -0.139** | -0.0797** |
| | (0.0232) | (0.0340) | (0.0257) |
| pain clinic law | -0.0316 | -0.0194 | 0.00179 |
| | (0.0350) | (0.0231) | (0.0330) |
| marijuana law | 0.0344+ | -0.00989 | 0.0396 |
| | (0.0190) | (0.0365) | (0.0252) |
| rural (%) | -0.383* | -0.547** | -0.337+ |
| | (0.150) | (0.175) | (0.180) |
| white (%) | 0.707 | 1.334 | -0.258 |
| | (1.047) | (2.198) | (0.988) |
| children (%) | 2.895 | 7.069* | 4.459+ |
| | (2.343) | (3.227) | (2.540) |
| seniors (%) | -0.315 | 2.065 | 0.178 |
| | (2.225) | (4.306) | (2.115) |
| UE rate | 1.479 | -0.740 | 0.585 |
| | (0.887) | (1.236) | (0.802) |
| povety rate | -0.467 | 0.129 | -0.543 |
| | (0.813) | (1.228) | (0.852) |
| Constant | 5.045** | 1.596 | 4.502** |
| | (1.078) | (1.556) | (1.190) |
| N | 612 | 612 | 612 |
| $R^2$ | 0.956 | 0.976 | 0.930 |
| Fixed effects? | all include state and year | | |
| weight | n/a | n/a | n/a |
| bootstrap p (DC) | 0.348 | 0.496 | 0.508 |
| boostrap p (UA) | 0.005 | 0.0280 | 0.192 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

Table 1.17: Robustness checks for effect of PDMPs (pharmacies)

| **Dependent:** | **ln(MME per capita from pharmacies)** | | | |
|---|---|---|---|---|
| **Adjustment:** | **no FL** | **subsample** | **coding** | **time trends** |
| PDMP - DC | -0.0177 | -0.0224 | | -0.00557 |
| | (0.0231) | (0.0231) | | (0.0137) |
| PDMP - UA | -0.0456* | -0.0580** | -0.0502** | -0.0343* |
| | (0.0171) | (0.0167) | (0.0174) | (0.0164) |
| DC - electronic | | | -0.0146 | |
| | | | (0.0215) | |
| PDMP mandate | -0.0583** | -0.0366 | -0.0457+ | -0.0350+ |
| | (0.0214) | (0.0272) | (0.0233) | (0.0206) |
| pain clinic law | -0.00554 | -0.0169 | -0.0319 | -0.0297 |
| | (0.0291) | (0.0425) | (0.0354) | (0.0257) |
| marijuana law | 0.0353+ | 0.0272 | 0.0339+ | 0.00863 |
| | (0.0197) | (0.0206) | (0.0193) | (0.0159) |
| rural (%) | -0.346* | -0.345* | -0.380* | -0.0739 |
| | (0.146) | (0.147) | (0.150) | (0.103) |
| white (%) | 0.335 | 1.067 | 0.698 | 1.410 |
| | (0.953) | (1.027) | (1.062) | (1.089) |
| children (%) | 2.832 | 3.911 | 2.836 | -0.798 |
| | (2.304) | (2.371) | (2.340) | (1.852) |
| seniors (%) | -0.444 | 0.959 | -0.372 | -1.247 |
| | (2.129) | (2.116) | (2.224) | (2.204) |
| UE rate | 0.988 | 1.856* | 1.473 | 1.394* |
| | (0.747) | (0.874) | (0.887) | (0.622) |
| povety rate | -0.122 | -0.500 | -0.450 | -0.0974 |
| | (0.763) | (0.840) | (0.817) | (0.691) |
| Constant | 5.309** | 4.315** | 5.065** | 59.56** |
| | (1.081) | (1.096) | (1.070) | (8.691) |
| N | 600 | 504 | 612 | 612 |
| $R^2$ | 0.960 | 0.958 | 0.955 | 0.978 |
| Fixed effects? | | all include state and year | | |
| weight | n/a | n/a | n/a | n/a |
| bootstrap p (DC) | 0.482 | 0.385 | 0.514 | 0.668 |
| boostrap p (UA) | 0.0110 | 0 | 0.00400 | 0.0420 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

electronic user access like Kentucky but under the identifying assumptions should not effect the overall pattern of results since these early adopters serve only as controls in the difference-in-differences specifications. The results do not indicate anything troubling as coefficients show the same overall pattern of significance and magnitude. The third column modifies the coding of the data collection variable so that paper data collection using multi-copy prescription pads does not count. This coding was discussed briefly in section 1.4.4 and only involves five states. It turns out that the recoding is not materially important, with the coefficient on data collection ("DC - electronic") negative but insignificant like in the baseline model. The final column adds state-specific linear time trends, improving the fit of the model slightly and eliminating any bias from varying pre-trends across states, but also adding a lot of degrees of freedom. The state time trends eat into the user access effect slightly, reduce it to about 3.5%, but it remains significant. The data collection effect also drops to nearly zero, reinforcing the overall impress that data collection alone does not impact prescribing.

### 1.6.2 Effect of modern PDMPs on pain prevalence

The results in section 1.6.1 indicated that PDMPs reduce opioid distribution to pharmacies and this appears to be primarily through the asymmetric information channel as the effect only occurs after doctors obtain the ability to query the system, but it is impossible to rule out that this could alternatively be a chilling effect as doctor tighten prescribing to legitimate patients know in light of uncertainty about how the PDMP changes legal norms and sanctions. As such the evidence in section 1.6.1 is theoretically consistent with both helpful and harmful effects for pain patients and it is important to try to assess the impact on pain management directly. Furthermore, the asymmetric information channel, as noted above, might actually improve access for some legitimate pain patients as their doctors can now verify that they do not obtain prescriptions from other doctors, refill prescribers early, or demonstrate other aberrant behavior. Looking directly at pain prevalence can help to assess this theory and its quantitative significance.

With these hypotheses in mind, I focus in this subsection on looking at the effect of prescription monitoring on the fraction of the adult population that reports suffering from "a lot" of pain on the previous day. This metric, from Gallup's daily surveys, is discussed in more detail in section 1.4.4. The results from the basic difference in differences specification (Table 1.18) show mostly null effects and because the share of people in any given state that reports being in a lot of pain varies little from year to year the estimates are quite precise, with standard errors of around 0.2 percentage points in all specifications. Data collection appeared to not impact opioid prescribing so it is reassuring that it appears to have no impact pain prevalence as any other conclusion would suggest misspecification. User access is more interesting as the coefficients are small but negative but insignificant, suggest that user access does not have a chilling effect and likely has no overall effect on pain patients but, if anything, it might be helpful to them. Another interesting, though preliminary, result is that in the specifications not weighted by population (columns 1-3) mandated use of a PDMP appears to lead to 0.6-0.7 percentage point increases in the prevalence of severe pain. These increases are consistent with the results from section 1.6.1 that showed mandated use of a PDMP leads to large drops in the types of opioids least likely to be abused and could be consistent with a chilling effect; however, these estimates are only marginally significant so I leave it as a topic for further study without drawing any firm conclusions.

I ran event study specifications to further explore the effect of user access on pain and the results are plotted in Figure 1.9. Interestingly, the negative effect on pain seems more substantial when viewing the full time course. In the initial year of implementation there is essentially no change in report pain but in all subsequent years there is a modest and collectively significant decrease in reported pain by about 0.5 percentage points. Table 1.19 shows results from robustness checks for the effect of PDMPs on pain. The middle column shows that when the PDMP effect is split into an initial year effect and a subsequent effect the subsequent effect is significant at the 10% level. I take this as suggestive evidence, consistent with the theory that PDMPs can free up physicians to prescribe more liberally to legitimate patients, but perhaps only after developing trust in the program's ability to assist in detecting misuse. Another pos-

Table 1.18: Effect of PDMPs on pain prevalence (Gallup)

| **Dependent:** | **pain yesterday (percentage points)** | | | |
|---|---|---|---|---|
| PDMP - DC | 0.240 | | 0.309 | 0.223 |
| | (0.286) | | (0.257) | (0.166) |
| PDMP - UA | | -0.184 | -0.241 | -0.179 |
| | | (0.228) | (0.220) | (0.149) |
| PDMP mandate | 0.712* | 0.648+ | 0.690* | 0.348 |
| | (0.323) | (0.323) | (0.325) | (0.212) |
| pain clinic law | -0.0725 | -0.0384 | -0.0396 | 0.128 |
| | (0.226) | (0.233) | (0.230) | (0.190) |
| marijuana law | 0.595+ | 0.689+ | 0.646+ | 0.146 |
| | (0.348) | (0.365) | (0.356) | (0.277) |
| rural (%) | -3.404* | -3.486* | -3.441* | -1.476 |
| | (1.362) | (1.372) | (1.350) | (1.358) |
| white (%) | 3.362 | 0.998 | 0.931 | 18.70+ |
| | (8.582) | (8.465) | (8.311) | (9.418) |
| children (%) | -31.88 | -29.37 | -29.56 | 3.335 |
| | (23.36) | (23.13) | (22.96) | (30.63) |
| seniors (%) | 54.52* | 55.44* | 56.11* | 65.20* |
| | (23.56) | (23.63) | (23.30) | (26.95) |
| UE rate | -1.400 | -0.590 | -0.736 | 1.782 |
| | (9.579) | (9.597) | (9.505) | (10.98) |
| povety rate | -18.12 | -18.15 | -17.57 | -3.792 |
| | (13.70) | (13.87) | (13.90) | (7.039) |
| Constant | 25.05* | 26.26* | 26.00* | 2.336 |
| | (11.41) | (11.66) | (11.36) | (12.58) |
| N | 612 | 612 | 612 | 612 |
| $R^2$ | 0.888 | 0.887 | 0.888 | 0.886 |
| Fixed effects? | all include state and year | | | |
| weight | n/a | n/a | n/a | pop |
| bootstrap p (DC) | 0.355 | | 0.542 | 0.581 |
| boostrap p (UA) | | 0.828 | 0.558 | 0.647 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

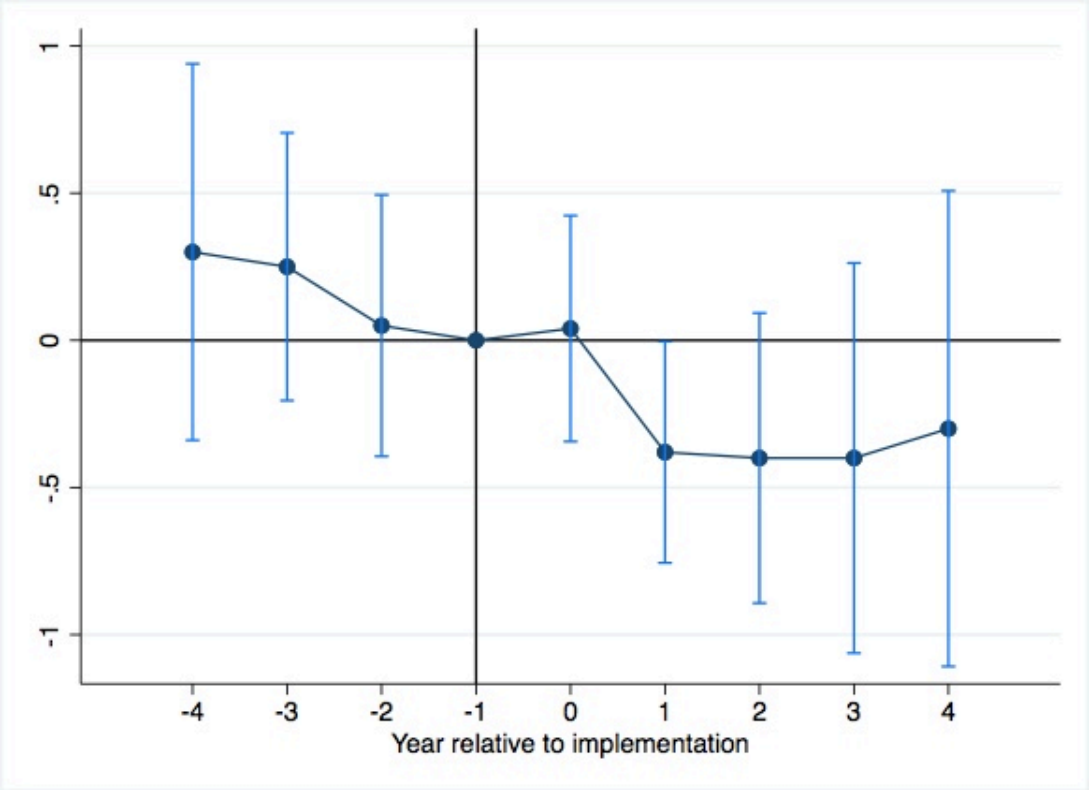


Figure 1.9: Event study for PDMP-UA on pain prevalence

sibility, is that the PDMP user access effect on pain evolves over time, perhaps as other policies have changed and made doctors feel like they cannot prescribe unless they can document reasons to believe patients as complying with the treatment plan. PDMPs, in addition to urine tests and other mechanisms can help to serve this function. Overall, however, the evidence is most consistent with the view that PDMPs do not affect pain patients and without further study that is the most reasonable conclusion.

Although all of the evidence so far is consistent with the view that PDMPs do not appear to harm patients by limiting access to legitimate pain treatment it is important to consider the power of these tests to detect a chilling effect. Can I rule out economically meaningful increase in pain as a result of PDMP implementation? The short answer is probably no because, as discussed more in chapter 2, health is incredibly valuable so even tiny increases in the prevalence of severe pain would be economically important. A few calculations are illustrative. In the preferred specification the net effect of PDMPs on pain prevalence ($\beta_{PDMP-DC} + \beta_{PDMP-UA}$)

Table 1.19: Robustness checks for effect of PDMPs on pain (Gallup)

| **Dependent:** | **pain yesterday (pecentage points)** | | |
|---|---|---|---|
| **Adjustment:** | **subsample** | **implementation** | **time trends** |
| PDMP - DC | 0.383 | 0.254 | 0.322 |
| | (0.279) | (0.258) | (0.345) |
| PDMP - UA | -0.319 | | 0.124 |
| | (0.207) | | (0.219) |
| PDMP - UA (t = 0) | | -0.0292 | |
| | | (0.252) | |
| PDMP - UA (t > 1) | | -0.433+ | |
| | | (0.246) | |
| PDMP mandate | 0.970* | 0.632+ | 0.625 |
| | (0.360) | (0.332) | (0.557) |
| pain clinic law | -0.103 | -0.0220 | -0.200 |
| | (0.226) | (0.235) | (0.419) |
| marijuana law | 0.575 | 0.688+ | 0.618 |
| | (0.350) | (0.343) | (0.470) |
| rural (%) | -4.139** | -3.426* | -5.271+ |
| | (1.352) | (1.333) | (2.661) |
| white (%) | 2.527 | -1.298 | -32.04 |
| | (8.583) | (8.266) | (26.76) |
| children (%) | -38.16 | -26.42 | -32.90 |
| | (23.19) | (22.80) | (36.94) |
| seniors (%) | 73.56** | 60.35* | 23.15 |
| | (21.22) | (24.05) | (97.94) |
| UE rate | 10.98 | -0.0803 | 6.372 |
| | (9.453) | (9.001) | (10.74) |
| povety rate | -24.87 | -17.66 | -18.78 |
| | (15.75) | (13.74) | (16.88) |
| Constant | 24.74* | 26.36* | 419.0 |
| | (11.47) | (11.07) | (312.6) |
| N | 378 | 459 | 459 |
| $R^2$ | 0.875 | 0.877 | 0.893 |
| Fixed effects? | | all include state and year | |
| weight | n/a | n/a | n/a |
| bootstrap p (DC) | 0.207 | 0.331 | 0.282 |
| boostrap p (UA) | 0.333 | 0.0631 | 0.553 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

has a point estimate of 0.044, with a standard error of 0.27, slightly larger than either coefficient alone due to a positive covaraince. This point estimate would indicate that one in every 2,500 adults would suffer from unmanaged pain as a result of the PDMP, which seems small, but does not exclude larger effects as the confidence interval extends to an upper bound of 0.57 or that 1 in 175 adults would suffer from unmanaged pain.

In order to assess the significance of these possible increases in pain we have to assess the economic cost, in dollars, of severe pain, which is not easy. One approach that remains controversial but is common in the public health literature starts by assuming that each year lived in good health (e.g. quality-adjusted life year or QALY) is worth a certain dollar amount, often $100,000. Changes in health can then be valued by assessing how they change the length or quality of life lived in QALYs and converting the QALYs to dollars. The difficult part is in deciding how much any particular health problems, such as pain, reduces quality of life.

Murray and Lopez (1997), the classic reference on this approach, suggest that symptomatic osteoarthritis reduces quality of life by 16% or 0.16 QALYs per year. To round up we might suppose severe pain in general costs 0.2 QALYs. If each QALY is worth $100,000 then suffering from severe pain for a year would have an economic cost of $20,000. At the extreme bound from earlier where 0.57 percent of the adult population suffers severe pain each day due to prescription monitoring that adds up to a cost of $28.5 billion per year. Olafsdottir et al. (2019) use an alternative method to estimate the willingness to pay for pain relief in an elderly sample from the HRS and get numbers in the range of $50-$145 per day. Over a year that adds up to $18,250 to $52,925 which implies final estimates broadly similar to the QALY approach.

The basic problem, in summary, is that although the Gallup data allows fairly precise estimates of the effects of PDMPs on pain, and they appear to be near zero, good health is generally so valuable that even modest increases in pain add up to large social costs. The flip side of this analysis is that to the extent PDMPs could potentially increase access to pain treatment for legitimate patients the corresponding social benefits are extremely large as well. In addition any overdose deaths or addictions prevented would also be enormously valuable. I leave it to

further work to do a general welfare analysis of PDMPs integrating all the potential costs and benefits.

### 1.6.3 Effect of modern PDMPs on opioid distribution to and pain in hospitals

As discussed earlier in section 1.3.2, PDMPs are not designed to influence inpatient care at hospitals. Prescribers in hospitals know exactly what medicines the patients have been given so there is little risk of abuse or diversion. Furthermore, overdoses are rare and can easily be treated since the patient is already under a doctor's supervision (Cauley et al. 2017). Nevertheless, PDMPs could still effect hospital care through a chilling effect on prescribers. This subsection uses data on opioid shipments to hospitals and surveys on pain management in hospitals to assess whether there are unintentional effects of PDMPs on hospitals.

The main results for the effect on opioid use in hospitals are presented in Table 1.20. The point estimates for the effect of data collection are small, ranging from around 2% to 4% reductions and are all insignificant. As noted in the previous section, reductions on this order of magnitude seem unlikely to be materially important for patient welfare, but the standard errors are only small enough to result out decreases greater than 8%. The point estimates for the effect of user access are even more precisely estimated and all very close to zero. It is also worth noting that the other policies, PDMP mandates, pain clinic regulation, and marijuana regulation also have insignificant effect as would be expected since these policies should only influence outpatient care. In summary the results are broadly consistent with the prior subsection and indicate that PDMPs do not have a chilling effect.

For completeness I've included results from some robustness checks for the effects of PDMPs on hospital use of opioids. Table 1.18 shows a variety of modifications for the basic regression specification. The first four columns use the amount of MME per capita as the dependent variable instead of the amount of MME per inpatient day. This has no meaningful effect on the results. The third and fourth column also exclude fentanyl from the opioid measure because, as noted in section 1.4.1, fentanyl quantities for hospitals are particularly noisy, potentially due to data entry errors on DEA tracking reports. Fortunately, the effects are unchanged with the

Table 1.20: Effect of PDMPs on Opioids dispensed to Hospitals (logs)

| **Dependent:** | **ln(MME) per day (hospitals)** | | | |
|---|---|---|---|---|
| PDMP - DC | -0.0426 | | -0.0336 | -0.0243 |
| | (0.0372) | | (0.0367) | (0.0404) |
| PDMP - UA | | -0.00654 | 0.00453 | 0.00410 |
| | | (0.0262) | (0.0251) | (0.0254) |
| PDMP mandate | 0.0214 | 0.0335 | 0.0218 | 0.0277 |
| | (0.0312) | (0.0323) | (0.0315) | (0.0317) |
| pain clinic law | -0.0108 | -0.0105 | -0.0112 | -0.00830 |
| | (0.0227) | (0.0244) | (0.0234) | (0.0226) |
| marijuana law | 0.00807 | -0.000497 | 0.00769 | 0.0328 |
| | (0.0257) | (0.0273) | (0.0263) | (0.0287) |
| rural (%) | -0.102 | -0.0902 | -0.101 | -0.196 |
| | (0.169) | (0.175) | (0.170) | (0.198) |
| white (%) | -2.272 | -2.153 | -2.246 | 0.769 |
| | (1.420) | (1.562) | (1.421) | (1.819) |
| children (%) | 3.561 | 3.313 | 3.538 | 5.196 |
| | (2.279) | (2.383) | (2.298) | (3.598) |
| seniors (%) | -0.939 | -1.114 | -0.949 | -1.905 |
| | (3.288) | (3.483) | (3.315) | (5.346) |
| UE rate | 0.0533 | 0.146 | 0.0426 | 0.472 |
| | (1.108) | (1.134) | (1.102) | (1.082) |
| povety rate | 0.513 | 0.551 | 0.501 | 0.880 |
| | (1.327) | (1.308) | (1.304) | (1.833) |
| Constant | 5.271** | 5.215** | 5.261** | 2.557+ |
| | (1.084) | (1.124) | (1.078) | (1.480) |
| N | 612 | 612 | 612 | 612 |
| $R^2$ | 0.918 | 0.917 | 0.919 | 0.923 |
| Fixed effects? | all include state and year | | | |
| weight | n/a | n/a | pop | pop |
| bootstrap p (DC) | 0.391 | | 0.572 | 0.705 |
| boostrap p (UA) | | 0.832 | 0.859 | 0.866 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

exclusion of fentanyl. The last two columns are the same style of robustness checks used in section 1.6.1 above, recoding data collection to only count electronic surveillance and using the subsample of states that changed their PDMPs during the sample time period. None of these specifications shows materially different results.

Given that the ARCOS data for hospitals is noisy which makes it hard to rule out plausibly meaningful decreases in opioid use I decided to look directly at the effect of PDMPs on pain management in hospitals. These data provide significantly more power for testing the chilling effect hypothesis and, because they represent the ultimate determinant of welfare, of interest on their own. Table 1.22 shows the results for the baseline difference-in-differences specification.

The overall picture is that PDMPs appear to have no chilling effect in hospitals, ruling out increases in uncontrolled pain of more than 0.25 percentage points due to PDMP data collection and more than 0.12 percentage points due to PDMP user access. Interestingly, in the specification where observations are weighted based on the number of surveys obtained by the hospital PDMP user access appears to cause an extremely modest reduction in uncontrolled pain that is significant at the 10% level. Most likely this is a false positive, inevitable when conducting a large number of tests, but could also be consistent with hospitalists feeling freer to use opioids aggressively after being able to verify that their patients do not have a history of misuse or doctor shopping according to the PDMP. The event study plot of coefficients for PDMP-UA do not help to shed light on this issue as the pattern, shown in Figure 1.10, is plausibly consistent with a pre-trend or noise or with potentially modest real effects. I leave this open as a question potentially worth further investigation.

## 1.7 Conclusion

Opioids are dangerous drugs that carry a substantial potential for abuse and diversion. In the U.S. they are the underlying cause of tens of thousands of deaths each year. Yet they are the most effective treatments available for severe, acute pain and potentially also for some forms of chronic pain. As a result any policy designed to curb the availability of opioids to

Table 1.21: Robustness checks for effect of PDMPs (hospitals)

| **Dependent:** | **MME per capita** | | | | **MME per day** | |
|---|---|---|---|---|---|---|
| **level or log?** | **level** | **log** | **level** | **log** | **log** | **log** |
| **Other adjustments:** | **n/a** | | **no fentanyl** | | **coding** | **subsample** |
| PDMP - DC | -4.125 | -0.0337 | -2.338 | -0.0393 | | 0.0349 |
| | (2.560) | (0.0345) | (2.860) | (0.0426) | | (0.0353) |
| PDMP - UA | -0.252 | -0.00756 | 3.336 | 0.0158 | 0.00582 | -0.0119 |
| | (1.920) | (0.0241) | (2.659) | (0.0283) | (0.0252) | (0.0352) |
| DC - electronic | | | | | -0.0385 | |
| | | | | | (0.0335) | |
| PDMP mandate | 1.909 | 0.0208 | 2.012 | 0.0146 | 0.0287 | 0.0568 |
| | (1.955) | (0.0301) | (2.847) | (0.0384) | (0.0317) | (0.0901) |
| pain clinic law | 0.113 | 0.00710 | -1.398 | -0.00368 | -0.0128 | 0.0348 |
| | (1.661) | (0.0240) | (2.373) | (0.0319) | (0.0237) | (0.0411) |
| marijuana law | -0.0709 | 0.00171 | 1.714 | -0.000899 | 0.00902 | 0.0698+ |
| | (1.773) | (0.0282) | (2.666) | (0.0295) | (0.0260) | (0.0403) |
| rural (%) | -6.462 | -0.146 | 1.393 | -0.116 | -0.0973 | -0.230 |
| | (8.651) | (0.124) | (12.96) | (0.174) | (0.170) | (0.199) |
| white (%) | -265.2* | -2.379* | -60.67 | -2.122 | -2.322 | -2.031 |
| | (100.3) | (1.184) | (81.71) | (1.704) | (1.437) | (1.695) |
| children (%) | -9.941 | 1.522 | 242.8 | 3.713 | 3.399 | 3.849 |
| | (145.2) | (2.186) | (235.4) | (3.040) | (2.323) | (4.164) |
| seniors (%) | 155.8 | 0.830 | -90.35 | -0.855 | -1.109 | 2.745 |
| | (228.4) | (3.008) | (277.2) | (3.988) | (3.294) | (3.152) |
| UE rate | -36.18 | -0.00769 | 71.36 | 0.594 | -0.0259 | -1.601 |
| | (90.46) | (1.051) | (119.8) | (1.157) | (1.089) | (1.521) |
| povety rate | 15.97 | 0.717 | 82.35 | 0.630 | 0.547 | -0.825 |
| | (93.82) | (1.280) | (146.4) | (1.596) | (1.297) | (1.085) |
| Constant | 239.5** | 5.207** | 40.97 | 4.665** | 5.360** | 4.981* |
| | (70.65) | (1.077) | (75.27) | (1.256) | (1.090) | (2.053) |
| N | 612 | 612 | 612 | 612 | 612 | 504 |
| $R^2$ | 0.852 | 0.901 | 0.869 | 0.927 | 0.918 | 0.946 |
| Fixed effects? | all include state and year | | | | | |
| weight | n/a | n/a | n/a | n/a | n/a | n/a |
| bootstrap p (DC) | 0.176 | 0.427 | 0.458 | 0.545 | 0.378 | 0.336 |
| boostrap p (UA) | 0.916 | 0.744 | 0.208 | 0.598 | 0.832 | 0.779 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

Table 1.22: Effect of PDMPs on uncontrolled pain for inpatients (HCAPHS)

| **Dependent:** | **pain uncontrolled (percentage points)** | | | |
|---|---|---|---|---|
| PDMP - DC | -0.279 | | -0.208 | -0.137 |
| | (0.248) | | (0.234) | (0.251) |
| PDMP - UA | | -0.304 | -0.278 | -0.355* |
| | | (0.208) | (0.203) | (0.159) |
| PDMP mandate | 0.189 | 0.227 | 0.194 | 0.377+ |
| | (0.219) | (0.217) | (0.216) | (0.213) |
| pain clinic law | -0.175 | -0.143 | -0.131 | -0.332 |
| | (0.239) | (0.264) | (0.259) | (0.235) |
| marijuana law | 0.218 | 0.242 | 0.250 | 0.514* |
| | (0.249) | (0.239) | (0.243) | (0.256) |
| rural (%) | -4.090* | -3.896* | -4.044* | -5.098** |
| | (1.530) | (1.553) | (1.613) | (1.803) |
| white (%) | 4.621 | 3.145 | 1.040 | 11.01 |
| | (15.24) | (14.97) | (14.83) | (14.74) |
| children (%) | 40.68 | 36.91 | 39.96 | 60.23* |
| | (25.88) | (25.39) | (25.15) | (24.95) |
| seniors (%) | 47.52+ | 43.02+ | 46.15+ | 48.43* |
| | (25.20) | (25.09) | (24.51) | (21.80) |
| UE rate | 4.941 | 7.209 | 6.771 | 5.112 |
| | (7.290) | (8.056) | (7.972) | (8.000) |
| povety rate | -24.00* | -26.02* | -24.90* | -27.93* |
| | (10.30) | (10.44) | (10.03) | (10.99) |
| Constant | 18.26+ | 20.76* | 21.09* | 11.31 |
| | (10.52) | (10.01) | (10.04) | (10.30) |
| N | 27,195 | 27,195 | 27,195 | 25,652 |
| $R^2$ | 0.665 | 0.665 | 0.665 | 0.703 |
| Fixed effects? | all include hospital and year | | | |
| weight | n/a | n/a | n/a | surveys |
| bootstrap p (DC) | 0.344 | | 0.469 | 0.678 |
| boostrap p (UA) | | 0.138 | 0.163 | 0.0290 |

Standard errors clustered by hospital in parentheses
** p<0.01, * p<0.05, + p<0.1

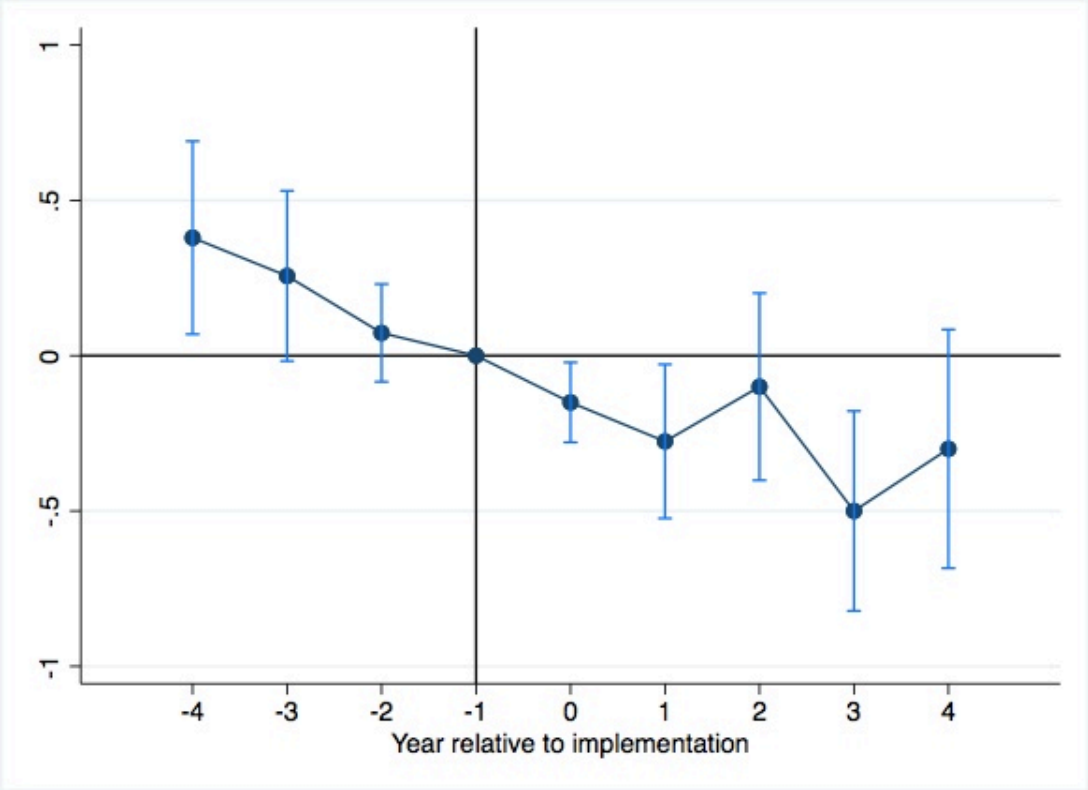


Figure 1.10: Event study for PDMP-UA on uncontrolled inpatient pain

reducing diversion and abuse of these drugs necessarily runs the risk of also curtailing access for legitimate pain patients. This chapter studied one widely advocated supply side method of opioid regulation, prescription monitoring, investing whether it appears to harm patients and, more generally, the mechanisms by which it can alter behavior.

I investigated the mechanisms by which PDMPs exert their effects, using the phasing in of two distinct PDMP features, data collection and end user access, in different states at different times as a natural experiment to distinguish the different channels by which prescription monitoring can have an effect. I supplemented these inquiries by looking directly at the effect of PDMPs on health outcomes, looking at the prevalence of severe pain in the general population and at the quality of pain management for inpatients in acute care hospitals. The preponderance of the evidence indicated that what I called the asymmetric information channel predominates as data collection appears to have no effect on prescribing, either for outpatients as intended or inpatients as a chilling effect would suggest, but giving physicians the ability to query the database has modest but significant effects on prescribing. Although, in theory, the asymmetric information channel could lead to improved pain management, I found no clear evidence that PDMPs impact the prevalence of chronic pain, for the better or worse. In summary the evidence indicates that prescription monitoring does not harm patients.

# Chapter 2

# Are we underestimating the value of (statistical) life?

## 2.1 Introduction

Values of a statistical life (VSLs) are one of the primary inputs into cost benefit analyses of environmental and consumer safety regulations so precise and unbiased estimates of these statistics are particularly important. This chapter investigates biases that afflict the standard method revealed preference method for estimating VSLs when the wage-elasticity of the VSL is large. Theory and some empirical evidence strongly suggest that safety is a luxury good and thus that the value of mortality risk reductions (i.e. the VSL) rises rapidly with income (Murphy and Topel 2006, Kaplow 2005). This appears to be the economic rationale for rapid improvements in safety across workplaces as countries develop (Hammit et al. 2000, Costa and Kahn 2004) and the secular rise in health spending in the late 20th century as countries became more affluent (Hall and Jones 2007). If safety is indeed a luxury good, then there will be enormous heterogeneity in VSLs across the population, just as there is substantial variation in willingness to pay for yachts, Teslas, and other luxury goods. I identify three reasons that this heterogeneity can lead the standard hedonic wage regression approach to yield downward biased estimates of the population mean VSL. First, when there is heterogeneity in compensating differentials

(or more generally, treatment effects) in a linear model then OLS yields a conditional-variance weighted average of the underlying compensating differentials across groups. I show that in a CPS sample of workers the conditional-variance of job-related mortality risk is much higher among less skilled workers with lower household income, so the OLS wage premium (compensating differential) estimate will disproportionally depend on the compensating differentials these workers demand. Second, the standard method of translating an estimate of the compensating differential for fatal injury risk into an estimate of a VSL implicitly assumes that the wage premium and wage have no correlation, but if safety is a luxury good then these variables will have a strong positive correlation and calibration suggests that methods that account for such a correlation will be around 50% larger than estimates using the standard method. Finally, I show that omitting the top-coded observations in datasets, while necessary for econometric purposes, can seriously bias population mean VSL estimates downward since the excluded workers have much larger VSLs.

A large literature has developed estimating the value of a statistical life (VSL) using hedonic wage models, starting with (Thaler and Rosen 1976). Much of this literature is summarized in in the widely cited meta-analyses by Aldy and Viscusi (2003) and Mrozek and Taylor (2002). The typical study focuses on estimating the average market compensation, usually in percent, for a small increment in on-the-job risk of death for some subpopulation. The assumption implicit in these models is that percentage point increase in pay per unit of risk does not vary across the population, at least in any significant way, and thus that the average return estimated is informative about the return a random individual would require to assume that risk.

This chapter fits into a growing literature exploring the potential importance of heterogeneity in compensating differentials. Aldy and Viscusi (2007), for instance, empirically investigate variation in VSLs across age groups, motivated by federal agency interest in potentially adjusting the VSL used in cost-benefit analysis of regulations when seniors are impacted. They also investigate cohort effects, noting that younger, richer, cohorts will likely have a higher VSL at any given point in their lifecycle. Their general conclusion is that there is some variation in VSLs across cohorts and across the lifecycle, with a U-shaped pattern over the lifecycle driven

by liquidity constraints early in life and declining conditional life expectancy later in life. Aldy and Smythe (2016) use a standard life-cycle consumption model to investigate the variation in VSL expected across genders and races due to differences in life expectancy and earnings. They find that there is significant heterogeneity, but it driven predominantly by variation in earnings and that differences in conditional life expectancy are relatively unimportant.

Another strand of research closely related to my focus in this chapter has emphasized the potential importance of a highly income-elastic VSL in explaining and assessing the world-wide rise in spending on health care and biomedical research. Hall and Jones (2007) calibrate a model of health care production and argue that the rapid increase in spending on health care as a fraction of GDP is the optimal response to steady GDP growth which pushes the value of staying alive (VSL) higher at a more than proportional rate. In their model even conservative choices for calibrating the intertemporal elasticity of substitution, the parameter that controls the income-elasticity of the VSL, result in a steady rise in health spending as a fraction of GDP. Jones (2016) follows up on the insights in Hall and Jones (2007) to argue that that the increasing fraction of R&D dedicated to biomedical research is a natural side-effect of rising incomes and income inequality since this results is a higher market value for outputs like new pharmaceuticals that can prolong lives. In related work, Murphy and Topel (2006) build a similar model of the VSL to value reductions in mortality due to medical progress and find enormous values for the median consumer. Their model implies significant variation with income since the income elasticity, although it varies with age and income, is generally above 1.5, but they do not explore the implication of this for variation or what it implies about the distribution of VSLs in the cross section, instead focusing on how growing incomes would lead to higher VSLs over time. In a way this chapter functions as a cross-section counterpart to this literature, asking what a high wage-elasticity implies about the distribution of VSLs at a given point in time and how to manage the potentially large dispersion econometrically.

Finally, this chapter builds on recent work building a quantile regression framework for hedonic wage models. Evans and Schaur (2010) explore variation in compensating differentials by age and conditional wage quantile using the Health and Retirement Study (HRS). They

find some variation in the returns to risk across age groups and enormous variation across conditional wage quantiles. This is suggestive evidence that the VSLs for these older workers are probably highly wage elastic but they cannot get a specific estimate for the wage elasticity since they lack a method for converting the conditional quantiles of their marginal into a marginal distribution of VSLs, a well-known problem in the quantile literature on which substantial progress has been made in recent years (Fortin et al 2011).

The rest of the chapter is organized as follows. Section 2.2 lays out the basic theory of the VSL as used in this chapter. Section 2.3 explains the problems heterogeneity in VSLs creates for estimating the population mean VSL, focusing on the first two technical problems outlined above. This section also clarifies the conditions under which this bias will be large, namely when the VSL is has a wage-elasticity at least somewhat above 1. Section 2.4 reviews the mixed theoretical and empirical evidence on the wage-elasticity of the VSL to assess the extent to which we should expect the biases identified previously to matter. The mixed evidence motivates the work in the following section which gives a detailed description of a quantile hedonic wage model, the data used to estimate the model, and the results of the estimation. Section 2.6 uses the results from the quantile model to calculate a sample-mean VSL and to extrapolate values for censored observations and then uses this information to estimate the worker population mean VSL. Comparison with the results of a baseline OLS hedonic wage model estimated using the same data allows for an assessment of the sizes of the three biases noted earlier in the chapter and their relative importance. Section 2.7 briefly concludes.

## 2.2 Theory of the VSL

The theory of the VSL was developed in several classic papers (Thaler and Rosen 1976, Arthur 1979, Rosen 1988, Shepard and Zeckhauer 1984) and subsequently refined so that there is largely a consensus on the theoretical issues involved in construction and interpretation of this metric. In this section I review the main ideas in the theory that are relevant to this chapter. The reader is referred to Viscusi (2013) for a more detailed review.

Consider a consumer at an arbitrary time period $t$ of her life with expected utility $E[U_t(p,a)] = \max u(c) + p\beta U_{t+1}(a-c)$ where $c$ represents consumption, $a$ represents assets, $p$ is the probability of surviving into the next period, and $\beta$ is the discount factor. Implicitly the utility of being dead is normalized to zero and independent of assets. This is a standard assumption that is not likely to be problematic in the contexts discussed below. The VSL reflects the trade-off a consumer would make between assets and the probability of survival along a constant expected utility locus. Thus we can define a VSL by totally differentiating $E[U_t(p,a)] = \bar{U}$ where $\bar{U}$ is a constant:

$$\begin{aligned} \beta U_{t+1} + p\beta(VSL)\frac{\partial U_{t+1}}{\partial a} &= 0 \\ VSL(p,a) &= \frac{\beta U_{t+1}}{p\beta\frac{\partial U_{t+1}}{\partial a}} \end{aligned}$$

We can simplify this expression by noting that, by the envelope theorem, $u^{\cdot\prime}(c) = p\beta\frac{\partial U_{t+1}}{\partial a}$, so we have

$$VSL(p,a) = \frac{\beta U_{t+1}}{u'(c)}$$

Intuitively, the VSL represents the utility value of remaining alive in utils scaled by $u.(c)$, a conversation factor that tells us how many dollars each util is worth.

Theory has emphasized that for infinitesimal changes in $p$ this quantity can be interpreted either as the willingness to pay (WTP) for improving the odds of survival or as the willingness to accept (WTA) compensation for a higher risk of death. In practice, when changes are not infinitesimal, these quantities may differ slightly, but recent research has emphasized that in the context of compensating differentials in the labor market the two quantities appear to not differ by much in practice (Knieser et al. 2012).[1]

[1]For non-rational consumers the two are likely to be different as many studies using the contingent valuation method have found. People often reject the premise of accepting compensation to take on greater risk so it can be hard to get estimates of a WTA using this methodology and the values stated are for a non-representative subset of the population.

Economists have long recognized that the definition of VSL implies heterogeneity because it is a function of consumer characteristics–in this simple model the baseline risk of death, $p$, and level of assets, $a$, in addition to any variation in underlying preferences that could influence marginal utility ($u'(c)$) like risk aversion. Recently, Camerson (2010) made this point by emphasizing that a VSL is, in effect, a rescaled demand for safety (in the form of higher p) and it is widely acknowledged that individual demand is sensitive to an enormous number of possible influences including the two probably most often cited: individual preferences and income. Based on this brief background on the concept of a VSL and why we would expect it to vary within the population, the next section explores the problems heterogeneity creates in attempting to estimate the mean VSL for a population of workers.

## 2.3 Problems heterogeneity creates for the standard model

Heterogeneity in the VSL is interesting in its own right and could be useful in helping to individuate costs and benefits in analysis of regulations as advocated by Sunstein (2013) and others. Less attention has been paid to the importance of accounting for such heterogeneity when focusing on traditional objects of interest such as the mean VSL for any population being studied. For concreteness we will assume throughout the rest of the chapter that the population of interest is all adult workers and their mean VSL is of interest because it is one of the major parameters of interest in cost-benefit analysis of federal regulations as well as assessments of the value of new health technologies that would save lives at some known cost.

In this section I show that when there is unmodeled heterogeneity in the VSL by wage, which intuitively we expect to be substantial, the standard hedonic wage model suffers from three potential biases. I investigate the conditions under which these biases will be large and provide rough calibrations of the plausible sizes of the biases where possible. Two of the difficulties discussed in the following subsections depend on the main parameter estimated using the hedonic wage regressions varying with wages or other characteristics, so before discussing each of the difficulties I start by briefly reviewing the standard hedonic wage model.

### 2.3.1 Hedonic Wage Models

The hedonic wage model has become the dominant framework for revealed-preference estimates of the VSL, serving as the guiding framework for almost all of the studies cited in recent meta-analysis and by federal regulators (Aldy and Viscusi, 2003; Moran, 2016). The empirical specification for these models is generally $\ln(wage)_i = \delta risk_i + X_i\beta + \varepsilon_i$ where $risk_i$ is a measure of the risk of fatality on the job (often scaled in units of 1/10,000 or 1/100,000) and $X_i$ is a vector of controls typically including age or experience, levels of education, union status, gender, race, and sometimes dummies for industry and occupation categories in cases where these are not perfectly collinear with the measure of job fatality risk. In theory the measure of risk should be whatever the worker has in mind as the subjectivity probability of death in different jobs, but this is obviously impossible to measure accurately, so, in practice, risk is measured by calculating the empirical average risk of fatality over a period of years for each industry or industry-occupation combination.

The coefficient of interest is $\delta$ which is interpreted as the compensating differential, in log point (approximately percent) increase in the wage, for a unit of risk The estimate is only unbiased under the standard selection-on-observables assumption which is, of course, a strong assumption that is occasionally questioned (Hwang et al 1992). In particular, since theory tells us that the VSL rises with income we should expect to see sorting on potential earnings ability, with higher ability workers choosing to "buy" more safety and thus sorting into safer jobs. To the extent earnings ability is not controlled for by observables there will be selection bias; however, this is standard practice in the literature because it is needed in order to make empirical headway without the ability to experimentally vary the on-the-job risk of death. Lavetti (2016) provides a detailed discussion of potential sources of bias in estimating hedonic wage equations in the context of mortality risk as well as the assumptions necessary to eliminate these biases and Heckmen et al. (2004) provide a more general discussion of the problems faced in estimating compensating differentials in general.

The coefficients for the model are typically estimated using ordinary least squares (OLS) and then $\hat{\delta}$ is interpreted as a market average wage premium per hour that compensates for

a one-unit increase in risk. Typically the units of risk are 1/10,000 fatalities per year of work and workers are assumed to work around 2,000 hours per year, so the compensation can be rescaled into a VSL using the equation $\hat{VSL} = \hat{\delta}(2,000\bar{wage})10,000$. If $\delta$ does not vary across the population then this estimate is unbiased, but if it does vary then we have to consider what kind of average $\hat{\delta}$ represents and how to interpret it.

It is easy to show that $\delta_i$ will not vary along the wage distribution only in the knife-edge case that the VSL has a wage-elasticity exactly equal to one . We start by defining a worker i's VSL as $VSL_i = 10^4\delta_i(2000)wage_i$, then rearranging to get an equation for $\delta_i$ and differentiating with respect to $wage_i$.

$$\begin{aligned}
\delta_i &= \frac{VSL_i}{10^4(2000wage_i)} \\
\frac{\partial \delta_i}{\partial wage_i} &= \frac{10^4(2000wage_i)\frac{\partial VSL_i}{\partial wage_i} - 10^4 VSL_i}{10^8(2000wage_i^2)} \\
&= \frac{1}{10^4(2000wage_i)}\frac{\partial VSL_i}{\partial wage_i} - \frac{VSL_i}{10^4(2000wage_i)^2}
\end{aligned}$$

To simplify and assist in interpretation we can write the left-hand side as an elasticity

$$\begin{aligned}
\frac{wage_i}{\delta_i}\frac{\partial \delta_i}{\partial wage_i} &= \frac{10^4(2000wage_i)}{VSL_i}\frac{1}{10^4(2000wage_i)}\frac{\partial VSL_i}{\partial wage_i} - \frac{10^4(2000wage_i)^2}{VSL_i}\frac{VSL_i}{10^4(2000wage_i)^2} \\
\varepsilon_{\delta,wage} &= \varepsilon_{VSL,wage} - 1
\end{aligned}$$

This equation has a simple interpretation: the wage-elasticity of the wage premium to compensate for risk will be zero only if the wage-elasticity of the VSL is one. If the wage elasticity of the VSL is greater than (less than) one then the wage premium for risk will rise (decline) with the wage. Intuitively, if the VSL scales proportionally with the wage then as people get richer their VSL rises proportionally and the amount of money they require to take on risk (the compensating differential) will rise proportionally. As a result, we would not expect there to be variation in $\delta_i$ because of differences in wages and the problems discussed below are probably

not major concerns. If, however, the wage elasticity of the VSL is above 1, as I find in my empirical estimates below, then the $\delta_i$ will be systematically higher for higher-wage workers and lower for lower-wage workers and it will be important to model this variation.

### 2.3.2 The conditional-variance weighting of OLS estimates

The first problem with estimating VSL when $\delta_i$ varies is that we have to be careful in interpreting how $\hat{\delta}$ is constructed by averaging over the various $\delta_i$ in the sample. In general OLS estimates produce a conditional-variance weighted estimate, $\hat{\delta} = \sum w_i \delta_i$ and in cases where *risk* and all the covariates in $\mathbf{X}$ are discrete, which is always the case in practice even if the number of cells is large, Angrist and Krueger (1999), building on Angirst (1998), show that the weighting can be written analytically as

$$w_i = \frac{var(risk_i|X_i)}{\sum var(risk_i|X_i)}$$

Using data from the CPS MORG discussed in more detail in section 2.5.2, I estimated the conditional variance for the sample and Figure 2.1 shows a binned scatterplot of conditional variance by income excluding the endogenous labor income from a primary job. It is clear from the figure that the weight is much larger for workers with little outside income. This makes intuitive sense because most jobs held by people in the CPS have low risk. It is the handful of high risk jobs that have extremely large residual levels of risk after regression on the covariate vector $X$ that drive the results and most people working in these jobs are relatively unskilled and earn almost all of their income from their job.

It is unclear how elastic we would expect $\delta_i$ to be with respect to other income so this bias might not be that large in practice. Solon et al (2015) discuss how it is possible to use reweighing to identify the full-sample average partial effect in the presence of this kind of unmodeled heterogeneity, but this is difficult and their preferred strategy is instead to explicitly model the heterogeneity. I take a first pass at this type of modeling in section 2.5 by developing a quantile hedonic wage model that allows the partial effects to vary along conditional wage quantiles

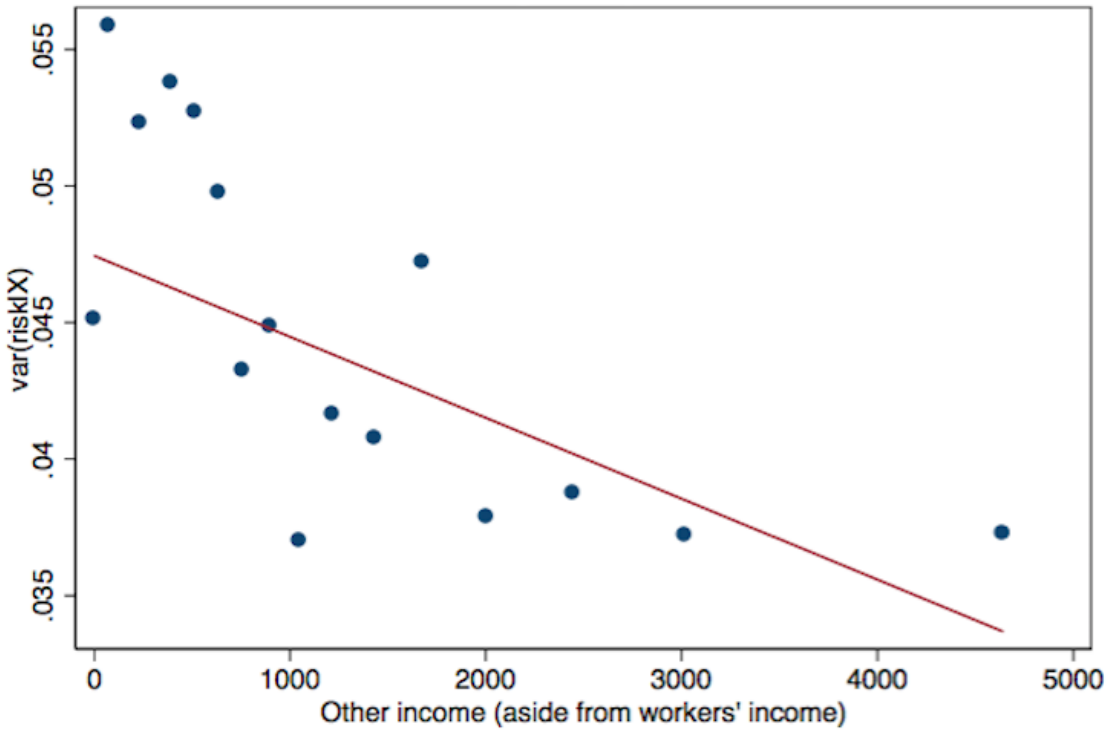


Figure 2.1: Conditional variance in the CPS MORG sample by Other Household Income

and use this model to assess whether the average partial effect for all quantiles diverges substantially from the OLS weighted average.

### 2.3.3 The cov($\delta_i$,$wage_i$) problem

A second problem arises in translating the information contained in $\hat{\delta}$ into an estimate of the population mean VSL. As noted above the standard practice is to use $\hat{VSL} = \hat{\delta}(2,000\bar{wage})10,000$ as the VSL estimate. This formula is an approximation for the sample analog to $E[VSL] = E[\delta wage](2,000)10,000$, which is simply a rearrangement of the definition of $\delta$. One problem is that it assumes $E[wage]E[\delta]$ will equal $E[\delta wage]$, which is true if $\delta$is a constant or does not vary with $wage$ but is not generally true. This is easiest to see if we write $E[\delta wage] = E[\delta]E[wage] + cov(\delta, wage)$ using the standard identity that $cov(X,Y) = E[XY] - E[X]E[Y]$. If $cov(\delta, wage) > 0$ then the standard formula underestimates the true sample mean VSL because it ignores a positive covariance term or, intuitively, it ignores how workers with high wages also have higher wage premiums for risk and thus particularly large VSLs. Conversely, if $cov(\delta, wage) < 0$ as would be the case if the the wage elasticity of the VSL is below 1 then the standard formula will overestimate the sample mean VSL.

Should we expect this bias to be large? Under some plausible assumptions it easily could be. The key question here is how large do we expect the covariance term to be compared to the product of means term. If the covariance term is on a smaller order of magnitude then the bias would be limited to at most 10% which seems tolerable given the sensitivity analyses conducted by regulatory agencies (Knieser and Viscusi, 2005). On the other hand, if the covariance team is as large or larger than the means term the mean VSL could be underestimated by half or more.

We can calibrate the expected size of the bias term by making some reasonable assumptions about the underlying distributions and the wage-VSL elasticity. We start by rewriting $cov(\delta_i, wage_i) = \rho_{\delta,wage}\sigma_{wage}\sigma_\delta$ . Since we have only considered heterogeneity of VSL, and thus of compensating differentials, by wage we have implicitly assumed that $\delta$is a function of $wage_i$ although not necessarily a linear one. This would lead to a large correlation so I use 0.5 for calibration as a lower bound estimate of the plausible correlation. The standard deviation of

$\delta$ will depend critically on how much $\delta$ varies with characteristics. As emphasized earlier, the the wage-elastiicty of the VSL in particular will be important in determining the amount of variation. If the wage-elasticity of the VSL is only slightly above or slightly below 1 then $\delta$will vary only slightly across the population. If, on the other hand, the wage elasticity of the VSL is around 2, as seems plausible on theoretical grounds (reviewed in the next section), then $\delta$will scale proportionally with the wage, and thus having the same coefficient of variation as the wage ($\sigma_\delta / E[\delta_i] = \sigma_{wage}/E[wage_i]$). How large is $\sigma_{wage}/E[wage_i]$? Wages are usually distributed with something like a log-normal distribution and for the log-normal this ratio is straightforward to calculate and depends solely on the variance parameter. In the case of wages the ratio is probably around 1 or perhaps somewhat higher in recent years as wage inequality has increased.[2]

Substituting these estimates back into our expression for the covariance term we get that $cov(\delta_i, wage_i) = \rho_{\delta,wage}\sigma_{wage}\sigma_\delta \approx 0.5E[\delta]E[wage]$ and thus approximately half as large as the other term, $E[\delta]E[wage]$. Leaving out this covariance term is thus plausibly a major bias, lowering the estimated sample mean VSL by around a third from what it would be after taking into account the correlation between the wage premium and the wage.

### 2.3.4 The top-coding problem

A final problem for standard hedonic wage models is that, due to data limitations, research have no choice but to exclude workers with top-coded earnings from their estimation samples. In recent years this is often a significant fraction of the relevant sample–in the CPS MORG data I use for estimation below, for example, just below 5% of full-time workers have top-coded earnings. Generally studies also exclude low earners as well because they report unreliably low wages (e.g. below minimum wage), minimal hours, or are too young, but if the VSL is highly wage elastic the importance of the two group is highly asymmetric and the effects do not cancel out as can be illustrated concretely with a numerical example.

[2] If X is distribution LN($\mu, \sigma$) then $var(X) = (e^{2\mu+\sigma^2})(e^{\sigma^2}-1)$ and $E[X] = e^{\mu+\frac{1}{2}\sigma^2}$ so the ratio $\frac{\sqrt{var[X]}}{E[X]} = \sqrt{e^{\sigma^2}-1}$. For the wage distribution a conservative approximation is $\sigma^2 = 0.7$ giving the results in the text.

Assuming a constant elasticity, the ratio of the VSLs for two workers with different wages will depend on the ratio of their wages. In particular, for elasticity $\epsilon$ the ratio of the VSLs is $\frac{VSL_i}{VSL_j} = \left(\frac{wage_i}{wage_j}\right)^{\epsilon}$. For calibration purposes let us assume, as before, that the VSL wage-elasticity is approximately 2. In that case the VSL of a low wage worker who earns, say, \$6/hour or about 25% as much as the average worker will have a VSL approximately $(1/4)^2 = 0.0625$ as large as the average worker. Even if we round down and assume these workers have VSLs of zero then adding in 5% or even 10% extra observations to represent excluded workers with low wages can onlypull the average VSL down by 5% to 10%. On the other hand, consider top-coded workers which are often assumed to have average wages around 1.5 times the cutoff based on assumptions about the underlying distribution of the upper tail. If the cutoff for censoring is 3 times the sample average wage then these workers will have, on average, a wage about 4.5 times as large as the average worker. With a wage elasticity of 2 these workers will have an average VSL at least $4.5^2 = 20.25$ times as large as the average worker.[3] With such large VSLs, excluding these workers substantially lowers the estimated VSL and adding back in their extrapolated VSLs will substantially increase the average–in this example by almost 100% ($1(0.95) + 20.25(0.05) = 1.9625$). The main lesson then is clear, if the VSL is wage-elastic than it is deeply problematic to leave out the top-coded earners since they are likely to have particularly high VSLs, while leaving out low earners is less of a concern since in the extreme they could only pull the average VSL down by less than their population share.

The rest of the chapter attempts to empirically assess the importance of the three biases discussed above and get a corrected estimate for the population mean VSL. It proceeds as follows. First, since the importance of these biases depends on the wage-elasticity of the VSL, I review the past literature on the income elasticity of the VSL, a closely related concept with a much larger literature, to see if it makes sense to be worried about these biases. Motivated by mixed evidence from that review, I discuss the construction and estimation of a quantile hedonic wage model that can be used to directly assess the first two biases as well as estimate the wage-elasticity of the VSL and thereby account for the top-coding bias.

[3]This is a conservative approximation since the VSL is convex in the wage so by Jensen's inequality the expectation $E[VSL(wage)] > VSL(E[wage])$.

## 2.4 What is the wage elasticity of the VSL?

The previous section emphasized that the wage elasticity of the VSL is critical for determining the quantitative significance of the three biases from heterogeneity. This section briefly reviews the literature on a closely related concept with a larger literature, the labor income elasticity of the VSL (IEVSL) in order to get a sense of whether the aforementioned biases are worth worrying about. There are two main strands of the literature on this subject, one theoretical that emphasizes models and economic intuition, and the other empirical, emphasizing reduced form estimation and correlations. The theoretical literature emphasizes two intuitive mechanisms that cause us to expect safety (i.e. risk reduction) to be a luxury good with a large income elasticity. First, when wages or incomes are low then, under the standard assumption of diminishing marginal utility, the marginal utility of consumption is high. As emphasized in section 2.2, the marginal utility of consumption functions as a conversion factor between utils and dollars; when it is high then a stream of utils from avoiding death translate into relatively few dollars as each dollar spent on consumption can yield a lot of utils. As incomes increase, due to high wages for instance, the marginal utility falls and the the stream of utility from survival is worth more than dollar terms. Hall and Jones (2007) emphasize this mechanism in modeling the rise in demand for health care as the U.S. has become more affluent. A second factor, emphasized in Murphy and Topel (2006), is that the utility from staying alive itself rises with income. The flow of utility from surviving is presumably low if one is on the edge of survival but increases with higher levels of income. As incomes grow the flow utility from survival becomes large, making any reduction in risk worth more utils and thus more when converted to dollars. In essence, these two observations tells us that if a consumer is rich the value of risk reduction is high (in utility terms) and the opportunity cost (in utility terms) is low, so demand for risk reduction will grow rapidly with income.

Kaplow (2005) formalizes these intuitions, showing that each of the two mechanisms above are important but which is dominant depends on parameters. In particular, he showed that there is a tight link between a parameter that controls diminishing marginal utility, the coefficient of relative risk aversion (CRRA), and the IEVLS if utility has sufficiently rapidly dimin-

ishing returns (CRRA > 1). The IEVLS will be larger than the CRRA at low levels of income and then asymptotically approach CRRA as income increases. Some calibration of the model suggests that that for people in affluent countries the IEVLS should be reasonably close to the asymptote with CRRA if CRRA > 1. Alternatively, if CRRA < 1 then $u'(c)$ declines slowly, minimizing the importance of the first mechanism, but increasing the importance of the second mechanism since the flow utility accumulates rapidly with income and consumption. Kaplow showed through illustrative calculations that the IEVLS in that case is likely to be close to but slightly above 1 and his general conclusion is that under standard assumptions it is difficult to find parameter values where safety is not a luxury good.

The coefficient of relative risk aversion is often represented by $\gamma$ and I adopt that convention for the rest of the chapter. A large literature in financial economics and macroeconomics has attempted to estimate $\gamma$ with plausible estimates ranging from slightly above 1 to nearly 10. A common reading of the literature is that values around 2 or slightly above are reasonable (Havranek 2013). Surveys that use stated preferences to get implied values of risk aversion have found even higher values with Barsky et al. (1997) suggesting 3.97 as a reasonable value in their elderly sample. Finkelstein, Hendren, and Luttmer (2016) use a value of 3 for calibrating the value of Medicaid and Lakdawalla, Reif, and Bauer (2017) assume $\gamma = 2$ when investigating how VSLs are affected by various types of old-age insurance such as pensions. Hall and Jones (2007) use more conservative estimates in the range of 1.5 in order to reconcile their model with findings on the growth of optimal health spending. Overall, there is nearly universal agreement that empirically $\gamma > 1$ and thus we should expect the IEVSL > 1 so theoretical analysis makes a strong case that we should be worried about the biases from heterogeneous VSLs.

An alternative approach to studying the IEVSL focuses on reduced-form empirical estimation with four main approaches. The simplest is to look at cross-country variation in income and how this is related with safety regulations and healthcare spending. If richer countries have safer workplaces and greater spending on healthcare then that is *prima facie* evidence that the income elasticity is positive and potentially large. Two pieces of cross country evidence from this line of research are particularly striking. The first is the dramatic decline in workplace fa-

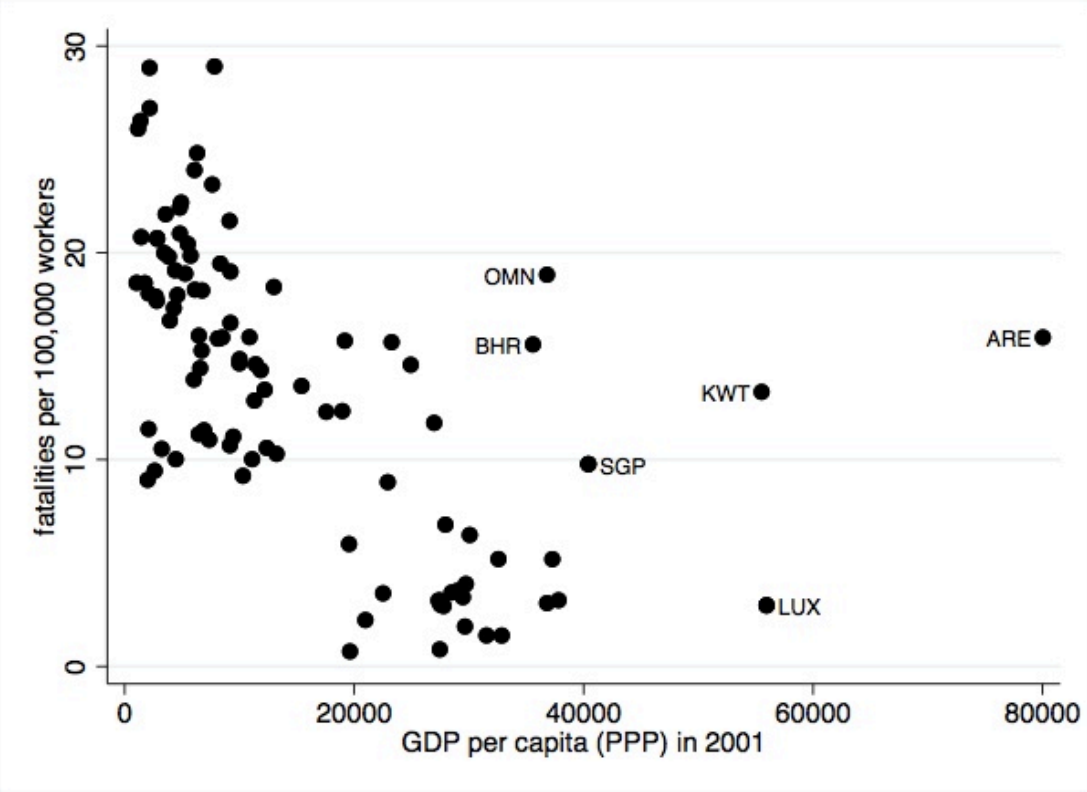


Figure 2.2: Relationship of workplace fatality rates and GDP per capita

tality rates across low and high income countries, illustrated in Figure 2.2 using data from the International Labor Organization (ILO) and World Bank for 2001, the most recent data I could find. Aside from a few oil-rich countries that are outliers the general relationship is robustly elastic with workplace fatalities dropping more than 1% for every 1% increment in income per capita. An additional striking piece of evidence, emphasized by Hall and Jones (2007) is richer countries spend disproportionally larger amounts on healthcare, suggesting that it, and survival, are luxury goods.

The second strategy in widespread use is meta-analysis of a pool of studies, usually ones based on revealed-preference using hedonic wage models, to investigate how estimates of the VSL vary across samples with different average incomes. The most widely cited of such meta-analyses is Aldy and Viscusi (2003) who also survey older meta-analyses. They report that most of the older meta-analyses find income elasticities in the wide range of 0.43-0.96. In their updated sample of studies and using different specifications with a broad set of controls they

find estimates all in the narrow range of 0.5-0.6 and "in none of [their] specifications [does] the income elasticity's 95 percent confidence interval upper bound exceed 1.0." However, many other meta-analyses have found larger IEVSLs. Dionne and Michaud (2002) estimate a confidence interval of 1.0-1.7 for the IEVSL and Viscusi (2018) reports a wider confidence interval of 0.378 to 1.322. In the later paper the wider confidential interval is a result of modeling how selection bias limits the observed sample of VSL estimates to, for the most part, only positive ones, censoring the variation for the meta-analysis and potentially leading to unwarranted precision. Making adjustments for selection bias leads to lower precision in the meta-regressions and helps illustrate why the IEVSL is difficult to pin down using this methodology.

In addition to concerns about sample selection and imprecision, there are several issues that make meta-analysis estimates of the IEVSL difficult to interpret and potentially biased. First, as noted in section 2.3.1, the standard OLS hedonic wage regression does not estimate the mean compensation that workers require to take on a small increase in risk. In practice, under the standard assumptions about selection on observables, the coefficient reflects a weighted average of the wage premiums for risk that different types of workers require. It is reasonable to think that the weighting might vary across samples and this complicates using meta-analysis to estimate the income elasticity. A concrete example can help illustrate these complications. Suppose, as is common, one study uses a sample of all workers with wages above the minimum wage and earnings that are not top coded to estimate a hedonic wage model while another study focuses on high-risk jobs. The full sample of workers likely has a higher average income since, as noted earlier, on the job fatality risk is much lower in high income countries and among high wage earners within a country. Now suppose we get similar VSLs in both studies. We could interpret that as evidence of a small IEVSL since the estimated VSL is not any higher in the sample with higher incomes. But it could instead be an artifact of how in both samples the estimated return to risk is driven by putting more weight on similar workers in high-risk blue collar occupations, thus generating similar VSLs. In order to avoid confounding these effects, studies would need to calculate the conditional variance weights and then use them to find the weighted mean income of the study which would be the appropriate input into

meta-analysis. In addition to this bias, there are a number of concerns with meta-analysis, such as the completeness of income reporting in different surveys and how this could bias the reported sample mean income. For example, it is known that income reporting in the PSID is more complete than in either version of the CPS (Fixler and Johnson, 2014). As a result, studies of compensating differentials using the PSID will have higher average incomes, all else equal, even though they are of essentially the same population (American workers) as studies using the CPS. If studies of the PSID and CPS find similar VSLs, as we would expect given the identical populations, then it would appear that the IEVSL is nearly zero as it does not vary with apparent income. These issues make it hard to be confident in empirical estimates of the IEVSL based on meta-analysis, especially considering how much these estimates diverge from our theoretical expectations.

As alternatives to meta-analyses economists have turned to a number of methods including using repeated cross sections and natural experiments to assess the IEVSL. Miguel and Leon (2017) survey air travelers in Sierra Leone who must get to the international airport in the capital. To reach the airport passengers must cross a body of water and have multiple transportation options with varying levels of safety. They find that richer travelers are much more likely to choose a safer but more expensive mode of transport and estimate a sample IEVSL of 1.77. Unfortunately, their context makes external validity a concern. The VSLs estimated in their study are uniformly low, below $1 million both for Africans and non-Africans, despite most of the travelers having fairly high incomes as one would expect for people taking international flights. This may be because the sample represents a set of highly risk-tolerant people, which would make sense as they are visiting a country that recently emerged from civil war and harbors a variety of deadly pathogens including the ebola virus.

A third method for getting an estimate of the IEVSL is to estimate hedonic wage regressions using repeated cross sections from the same country but at different time periods as safety and wages changed. Hammit, Liu, and Liu (2000) use data from Taiwan province in China to explore how the VSL evolved over several decades, from the 1970s through 1990s, showing that wage premiums rose dramatically even as wages were rising, leading to an estimated IEVSL

of close to 3 in their preferred specification. Costa and Kahn (2004) apply the same strategy to data from the U.S. in the 20th century and find a slightly lower but still substantial IEVSL of 1.4. The U.S. also displays the same general trend of rising wages and broad-based improvements in safety that would be hard to reconcile with a low IEVSL. One concern in both studies is that there may have been time-varying factors that could have influenced the wage structure such as, in Taiwan's case, industrialization or, in the U.S. case, skill biased technological change, that confounds the evolving compensating differential for risk.

To date no method for empirically estimating the IEVSL has gained universal acceptance but my reading of the literature is that there is good reason to expect the VSL is highly income (and wage) elastic and as a result we should be concerned about the biases discussed in the previous section. Theory is unambiguous that safety is a luxury good and as a result it is hard to develop a model where the IEVSL is not at least one. Empirical evidence, while mixed, suggests that the IEVSL could easily be above one, although the data limit the ability to give precise numbers. Having established that it is likely the wage-elasticity of the VSL is large and thus that the heterogeneity bias is important to correct for, I turn to estimating a more flexible hedonic wage model that allows for some heterogeneity. This allows me to take a first pass as assessing the empirical importance of these biases, although only under somewhat restrictive assumptions.

## 2.5 A Quantile Hedonic Wage Model

In this section I estimate a conditional quantile hedonic wage model, extending the model developed by on Evans and Schaur (2010) to recover a joint distribution of wages and risk premiums, which I think use to calculate the implied marginal distribution of VSLs consistent with the conditional quantile model. These methods, as discussed below, build on work by Machado and Mata (2005). I also estimate a standard OLS hedonic wage model on the same sample for comparison purposes so that the following section can assess the quantitative significance of the differences.

### 2.5.1 Empirical Strategy

The standard estimating equation for the OLS quantile hedonic model is

$$\ln(wage_i) = \alpha + \delta risk_i + X\beta + \varepsilon_i$$

where $X$ is a vector of controls. The conditional quantile hedonic wage model simply extends this to allow the coefficients to differ by conditional quantile $q$:

$$Q_q(\ln(wage_i)|risk_i, X_i) = \alpha(q) + \delta(q)risk_i + X\beta(q)$$

Intuitively, the key difference between the models is that the conditional quantile model allows for some heterogeneity in the compensation for risk along conditional wage quantiles. In effect I am estimating $\delta(q)$ for each of 99 quantiles as opposed to a single parameter $\delta$ that represents a weighted average of the compensating differential across the population. This should help mitigate the bias from assuming there is no heterogeneity in compensating differentials, but does not necessarily eliminate the bias since the heterogeneity is required to fit a certain parametric form. In that sense the estimates below are best seen not as providing a new number that is an improved estimate of the VSL for use in policy analysis, but rather as providing evidence on the scale of the bias and importance of finding ways to model heterogeneity. The model allows me to estimate an analog of $\delta$by averaging $\delta(q)$ across the quantiles, which are all equally likely, for comparison purposes, which allows me to get a sense of how much $\delta$ varies across workers.

A few words about the identifying assumptions of this model are important. They include all the standard assumptions needed for the OLS hedonic wage model including the key assumption of selection on observables. While both the OLS and quantile models can be fit with a rich set of demographic covariates as well as covariates for other influences on wages (occupation, industry) this remains a strong and untestable assumption. Nevertheless, this is standard practice in the literature because there is no better alternative and some assumptions are necessary in order to get a number to guide public policy.

The quantile model does not as easily yield estimates of compensation for risk as the OLS hedonic model because the quantile operator is not linear, as widely discussed in the quantile treatment efforts literature (Angirst and Pishke, 2009). Under only the selection-on-observables assumption the quantile model identifies the effect of risk on conditional quantiles of the wage distribution, but not necessarily on any individual person. For instance, if we find that the coefficient on risk at the 10th quantile is 0.01 then that tells us that, after conditioning on the effect of the other covariates, an extra 1 in 10,000th risk of fatality on the job raises wages in the 10th quantile by approximately 1%. This does not mean that it raises the wage for a particular individual who is in the 10th (conditional) quantile by 1% unless the same person would be at that point in the distribution under the lower and higher values for risk. That changes in risk (or any other covariate) do not change the ordering of wages in the distribution is called the rank invariance assumption and it is required in order to interpret the coefficients in the quantile model as treatment effects on individuals they way we generally do with OLS models.

Like the selection on observables assumption, the rank invariance assumption seems unlikely to hold exactly. In particular, changes in the level of risk faced by a worker are likely associated with at least some changes in the actual tasks done on the job. It seems likely that some workers might have strong skills in tasks related to crab fishing, a high risk job, but no particular skills for a low risk job like cooking crabs as a quick service restaurant. This pattern of skills is likely to violate the rank invariance because this worker will be higher up in the wage distribution when risk is high and lower in the distribution when risk is low. That said, it is possible that some relaxation of the rank invariance assumption is possible and this is a topic of active investigation by a number of researchers (Frandsen et al 2015). Furthermore, it is possible that small deviations from the rank invariance assumption do not severely bias the treatment effect estimates but bounding this sort of bias quantitatively remains an active area of research as well. Ultimately, some assumptions must be invoked in order to make headway assessing the importance of heterogeneity in VSLs for estimation purposes. To the extent this assumption is invalid or unreasonable the results below must be taken as suggestive and pending confirmation or reassessment in future research as methods improve.

A related difficulty in interpreting the conditional quantile model comes from the fact that the quantiles are only conditional. I need to recover the joint distribution of $\delta$ and $wage$ in order to calculate the covariance term discussed in section 2.3.3 and the conditional quantiles are not directly helpful for this purpose. I adopt the strategy proposed and implemented in Machado and Mata (2005) to generate an estimate of the joint distribution of the variables of interest that is based on the conditional quantile model. Following their method, I draw a large random sample of vectors $(risk_i, X_i)$ from the empirical distribution in the estimation sample (with replacement) and pair them with a randomly chosen quantile $(q_i)$. I then use the model to predict the wage ($wage_i = \alpha(q_i) + \delta(q_i)risk_i + X\beta(q_i)$) and risk premium ($\delta(q)$) associated with each draw, yielding a large simulated sample of $(wage, \delta)$ pairs. The VSL for each point can then be calculated using the standard formula for an individual's VSL ($VSL_i = (2,000)wage_i\delta_i(10,000)$) and I can calculate any sample statistic of interest such as the mean sample VSL based on this marginal distribution. This method assumes that the quantile regression model does a good job fitting the data, an assumption that is tested below by plotting the marginal distribution of wages generated by the model against the empirical distribution.

### 2.5.2 Data

I follow standard practice in the hedonic wage model literature in assembling my estimation sample and constructing measures of risk, paralleling the widely cited Viscusi (2004) as closely as possible but with updates due to changes in data coding. The main data source for earnings and demographics is the Current Population Survey (CPS) Merged Outgoing Rotation Groups (MORG) sample for 2015. The CPS MORG includes information on hours worked, wages or usual weekly earnings, industry, occupation, and a host of demographics that can be used as controls. I drop observations for part-time workers and then impute an hourly wage for non-hourly wage workers by dividing usual weekly earnings by usual hours worked. Following standard practice in the literature, I drop observations with reported wages below \$6 as implausible and exclude workers with top-coded earnings (\$72/hour or \$2880/week) since censoring will obscure any relationship between risk and earnings for these workers. I restrict attention

to private sector workers because I lack information on non-wage compensation which is disproportionally important for public sector workers. The most recent BLS estimate from the Employer Costs for Employee Compensation data is that private sector workers earn 69.5% of their compensation as wages while state and local government workers receive just 62.6% in wages. Estimates of the size of compensating differentials will be biased for all workers to the extent that higher on the job fatality risk leads to higher (unmeasured) compensation, but this underestimate will be much larger for workers who bargain to obtain so much of their compensation as benefits. As is standard in the literature I use pre-tax wages for analysis due to data limitations and I focus on workers in their fourth month in the sample to avoid repeated observations which rarely add much new information but significantly complicate the calculation of standard errors. [4] I construct standard demographic controls for education groups, race, potential experience, membership in a union, gender, marital status, and an indicator for whether a respond-ant is a veteran using the demographic information in the CPS. Finally, for consistency with prior research I limit the estimation sample to workers between the ages of 18 and 65 inclusive. The final estimation sample has 39,006 observations and summary statistics for the group are reported in Table 2.1 below.

To construct a measure of on-the-job fatality risk I use the Census of Fatal Occupational Injuries (CFOI) collected by the BLS, which regarded as the most accurate source of information for the U.S. and is widely used in estimating hedonic wage models (Viscusi 2013). The CFOI data is, as the name indicates, a census of all fatalities on the job in the U.S. each year, assembled meticulously by the Bureau of Labor Statistics (BLS) and matched to a full NAICS (industry) code and SOC (occupation) code for each deceased worker. The CFOI has been compiled each year since 1992 but the coding of industries changes slightly every five years as industries are recategorized so I limit my use of the data to the years 2011 to 2015 for consistency with the 2015 CPS. The BLS publicly reports fatality rates for 2-digit NAICS by 2-digit SOC cells and further disaggregation could introduce substantial measurement error into the measure of risk,

[4]Standard errors must already be clustered for correlation within industry-occupation cells since this the level of aggregation for the risk variable. If there are repeated observations for some individuals then I would need to account for an additional dimension of clustering and two-way clustering is an active area of research.

Table 2.1: Summary statistics for the CPS MORG sample

| Variables | N | mean | StD | min | max |
|---|---|---|---|---|---|
| wage (2015 $s) | 39,006 | 23.25 | 13.33 | 6.010 | 72 |
| union member | 39,006 | 0.0703 | 0.256 | 0 | 1 |
| risk | 39,006 | 0.277 | 0.480 | 0 | 5.424 |
| experience | 39,006 | 23.12 | 12.06 | 0 | 46 |
| High school | 39,006 | 0.318 | 0.466 | 0 | 1 |
| Some college | 39,006 | 0.290 | 0.454 | 0 | 1 |
| Bachelors | 39,006 | 0.230 | 0.421 | 0 | 1 |
| Masters | 39,006 | 0.072 | 0.258 | 0 | 1 |
| PhD | 39,006 | 0.014 | 0.117 | 0 | 1 |
| Professional Deg | 39,006 | 0.011 | 0.106 | 0 | 1 |
| veteran | 39,006 | 0.0514 | 0.221 | 0 | 1 |
| male | 39,006 | 0.565 | 0.496 | 0 | 1 |
| Hispanic | 39,006 | 0.149 | 0.356 | 0 | 1 |
| Black | 39,006 | 0.100 | 0.301 | 0 | 1 |
| Indian | 39,006 | 0.007 | 0.084 | 0 | 1 |
| Asian | 39,006 | 0.062 | 0.241 | 0 | 1 |
| Other (race) | 39,006 | 0.012 | 0.108 | 0 | 1 |
| married | 39,006 | 0.555 | 0.497 | 0 | 1 |

**Note:** units for risk and further details in the main text.

so I stick to this level of aggregation for measuring risk. There are 25 different 2-digit NAICS codes and 23 major occupational groups (2-digit SOC codes) and thus a total of 575 combined industry-occupation categories that workers in the estimation sample can be matched to. Many of these cells have no fatalities in any given year but most have at least one over the five year period used for averaging and some have substantial levels of risk. For instance, protective service workers in the accommodations industry suffer 5.4 deaths per 10,000 workers each year and and forestry workers in the wood products industry have a comparable risk with 3.48 deaths per 10,000 workers each year. The counts of fatalities by occupation-industry cell were matched to CPS counts of the total number of worker-years in each group over the time period to calculate a fatality rate per worker per year. Finally, I multiplied all of the fatality rates by 10,000 for readability so that we can interpret the rate as the number of deaths per 10,000 workers per year on the job.

A few observations from the summary statistics are worth noting. First, the mean wage in the sample is $23.25 compared with $29.18 estimated in the National Income and Product

Accounts. This is due to the omission of the 2,042 workers with top coded earners who earn a significant fraction of the total earnings in the sample but are excluded. The sample also excludes some low-wage and part-time workers which pushes the average up, but as discussed in the theoretical sections, these effects are highly asymmetric. I draw attention to this because it emphasizes how significant the relatively small sample of top-coded workers can be in influencing the population average for a variable like wage or, as we will see, VSL. Second, the median risk in the sample is just 0.06 deaths per 10,000 workers but the mean is close to 0.3 deaths per 10,000 workers, illustrating the heavy skew in the distribution of risk across workers.

### 2.5.3 Results

The results of the OLS hedonic wage model as well as selected quantiles of the conditional quantile model are shown together in Table 2.2. The OLS coefficient of interest is 0.0244 indicating approximately a 2.44% wage premium for a 1/10,000th risk of death. This is comparable to other estimates in the literature which are typically in the 1-3% range. The standard implied estimate of the population VSL is then this wage premium times the average wage times a scaling factor to get the proper units which yields $11.346 million for an estimate of the mean VSL, slightly above the Department of Transportations preferred VSL of $9.6 million and close to many estimates of around $10 million that are common when using the CFOI data (Moran 2016).

The quantile results reported in the table show substantial variation as expected. At lower quantiles the estimates of compensation are positive but not significant either statistically or economically. The compensation for a 1/10,000th risk of death is below 1% at both the 10th and 25th quantiles. This may have to do with bunching of wages around minimum wage, which suppresses variation in observed wages, or it may be a result of the known imprecision of quantile models for extreme quantiles (Chernozhukov and Fernandez-Val 2005). By the 50th percentile the coefficient is remains insignificant statistically be has risen to 1.84%, similar to typical hedonic wage estimates although perhaps surprisingly still below the OLS

Table 2.2: Quantile Regression Results

| **Dependent:** | **ln(wage)** | | | | | |
|---|---|---|---|---|---|---|
| | **OLS** | **q=0.1** | **q=0.25** | **q=0.5** | **q=0.75** | **q=0.9** |
| risk | 0.0244 | 0.0065 | 0.0076 | 0.0184 | 0.0399** | 0.0695** |
| | (0.0153) | (0.0142) | (0.00535) | (0.0147) | (0.0144) | (0.0203) |
| union | 0.146** | 0.0965** | 0.152** | 0.171** | 0.155** | 0.137** |
| | (0.017) | (0.0153) | (0.0109) | (0.0144) | (0.0124) | (0.0146) |
| high school | 0.201** | 0.118** | 0.167** | 0.212** | 0.250** | 0.261** |
| | (0.0108) | (0.0121) | (0.0118) | (0.00856) | (0.0125) | (0.0204) |
| some college | 0.258** | 0.156** | 0.228** | 0.277** | 0.321** | 0.334** |
| | (0.0115) | (0.0134) | (0.0119) | (0.00823) | (0.0102) | (0.0168) |
| bachelors | 0.474** | 0.322** | 0.429** | 0.513** | 0.558** | 0.566** |
| | (0.0156) | (0.0163) | (0.0162) | (0.0117) | (0.012) | (0.0249) |
| masters | 00.570** | 0.429** | 0.546** | 0.637** | 0.654** | 0.619** |
| | (0.0184) | (0.0192) | (0.0167) | (0.0157) | (0.016) | (0.0185) |
| doctorate | 0.650** | 0.508** | 0.646** | 0.722** | 0.773** | 0.691** |
| | (0.0359) | (0.0485) | (0.0251) | (0.0211) | (0.0149) | (0.0325) |
| experience | 0.0253** | 0.0143** | 0.0202** | 0.0269** | 0.032** | 0.0313** |
| | (0.00166) | (0.00142) | (0.00010) | (0.00010) | (0.00056) | (0.00117) |
| veteran | -0.009875 | 0.0000787 | 0.00259 | -0.00178 | -0.0314** | -0.0295** |
| | (0.0091) | (0.0107) | (0.00839) | (0.0105) | (0.0111) | (0.0147) |
| married | 0.0537** | 0.0419** | 0.0507** | 0.0565** | 0.0566** | 0.0599** |
| | (0.00657) | (0.00564) | (0.00617) | (0.00757) | (0.00536) | (0.00686) |
| Hispanic | -0.0934** | -0.0744** | -0.0916** | -0.0866** | -0.086** | -0.0915** |
| | (0.0142) | (0.0097) | (0.00923) | (0.0066) | (0.00796) | (0.0127) |
| Black | -0.133** | -0.116** | -0.126** | -0.131** | -0.131** | -0.117** |
| | (0.011) | (0.0117) | (0.00813) | (0.0173) | (0.0156) | (0.00775) |
| male | 0.173** | 0.121** | 0.162** | 0.177** | 0.192** | 0.173** |
| | (0.00855) | (0.0071) | (0.00813) | (0.0173) | (0.0156) | (0.00775) |
| Constant | 2.774** | 2.458** | 2.582** | 2.745** | 2.950** | 3.144** |
| | (0.0477) | (0.0471) | (0.0430) | (0.0434) | (0.0514) | (0.0570) |
| N | 39006 | 39006 | 39006 | 39006 | 39006 | 39006 |

Standard errors clustered on industry and occupation in parentheses.
** p<0.01, * p<0.05, + p<0.1

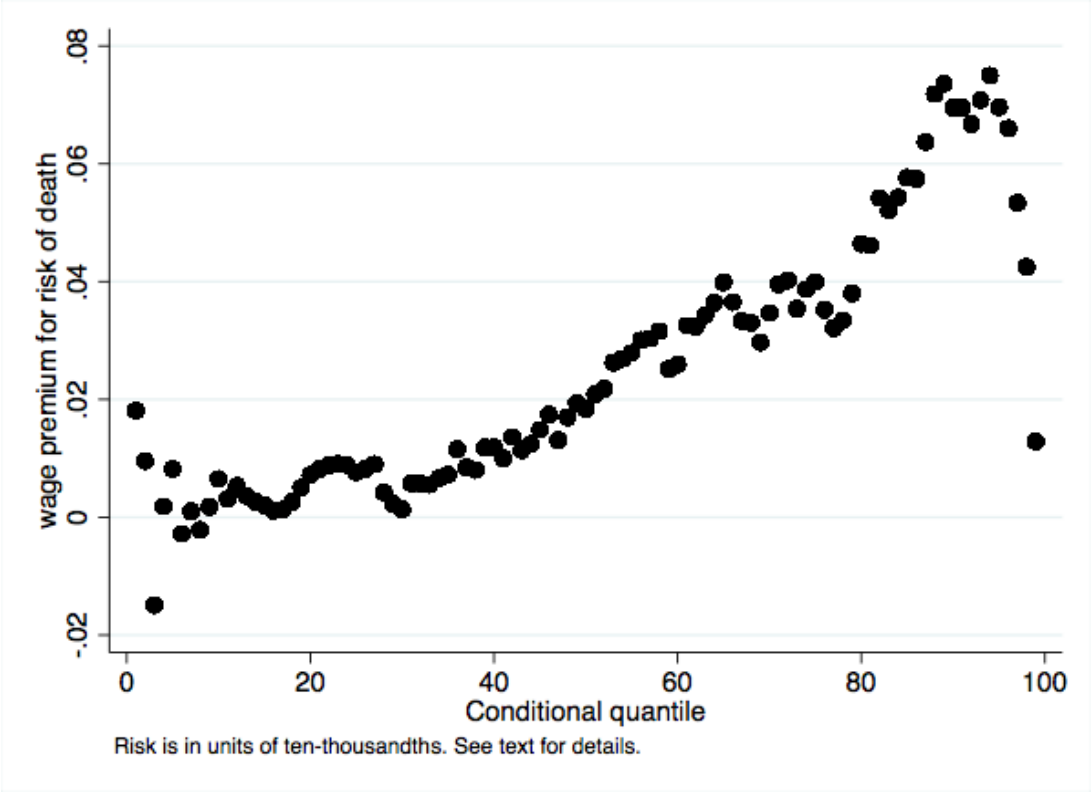


Figure 2.3: Raw estimates of $\beta$ by quantile

estimate. From the 50th percentile to the 90th percentile the coefficient on risk rises dramatically, almost quadrupling to 6.95% and becoming statistically significant at the 75th and 90th quantiles. The steady rise in the coefficients on risk across the quantiles is consistent with the theory articulated earlier and suggests a large wage-elasticity of the VSL. It is also consistent with the quantile estimates reported in Evans and Schaur (2010) who fit a similar model for elderly workers using the HRS and Polat (2013) who estimates a similar model for a sample of workers in Turkey.

The full set of coefficients on risk for each of the 99 quantiles are plotted in Figure 2.3. The general trend of increasing compensation for risk at higher conditional quantiles is clear but a few other trends emerge. First, there is substantial noise at both the top and bottom of the distribution. This is not unusual for quantile regression estimates due to the same biases that plague non-parametric estimates around boundaries, amplified here by censoring. The coefficients at the higher quantiles are especially imprecisely estimated, halting the rising trend

around the 90th quantile and dropping dramatically for the top few quantiles. This is likely also due to censoring since no wages above \$72/hr are included in the estimation sample. This kind of censoring is known to induce a bias toward zero since, to the extent compensating differentials increase wages, they tend to push the wages above the cutoff and thus out of the sample. This forces the coefficients on all characteristics including education and experience to be pushed toward zero in the higher quantiles. For example, the premium for attending college (versus a dropping out of high school) rises steadily with quantiles until hitting at peak of 0.58 at the 89th quantile and dropping to 0.347 for the 99th quantile. The returns to experience drop off by more than half in the top 10 quantiles and the gender bias in favor of men peaks at the 79th quantile before dropping by almost 60% by the 99th quantile.

I fit two models to correct this bias on the estimated compensating differential for risk and the results reported in Table 2.3. Both models project the coefficient on risk for each quantile ($\delta(q)$) on different features of the data. Model 1 regresses these coefficients to the average wage at that conditional quantile and the fraction of projected wages above the cutoff. The later term is highly significant and suggests adjusting the $\delta$s by adding back in the estimated amount they are dragged down due to censoring. Model 2 takes a simpler approach and simply predicts $\delta(q)$ based on the empirical (roughly quadratic) trend in the lower 79 quantiles. The compensation for risk in the top 10 conditional quantiles are then based on the fitted values from the model. Both are these proposed solutions are ad hoc and it is unclear which model is superior or if the more conservative approach of not making adjustments in preferred. Based on these considerations I report results for each model in the following sections.

Following the procedure outlined above I drew 10,000 samples (with replacement) of covariates from the empirical distribution of the estimation sample and paired them with randomly chosen quantiles from 1 to 99. Then for each simulated observation I used the quantile hedonic wage model to predict the associated wages. The (kernel smoothed) empirical distribution of wages from the sample is plotted against distribution of simulated wages in Figure 2.4.

Table 2.3: Models to adjust for bias in the top quantiles

| | Model 1 | Model 2 |
|---|---|---|
| **Dependent:** | $\delta(q)$ | $\delta(q)$ |
| wage | 0.00267*** | |
| | (0.00013) | |
| % above cutoff | -0.875*** | |
| | (0.099) | |
| quantile | | -0.00019*** |
| | | (0.000117) |
| quantile$^2$ | | 9.43e-06*** |
| | | (1.12e-06) |
| constant | -0.034*** | 0.00460* |
| | (0.00285) | (0.00266) |
| N | 99 | 79 |
| $R^2$ | 0.840 | 0.941 |

Robert standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
**Note:** Details on the specifications included in main text.

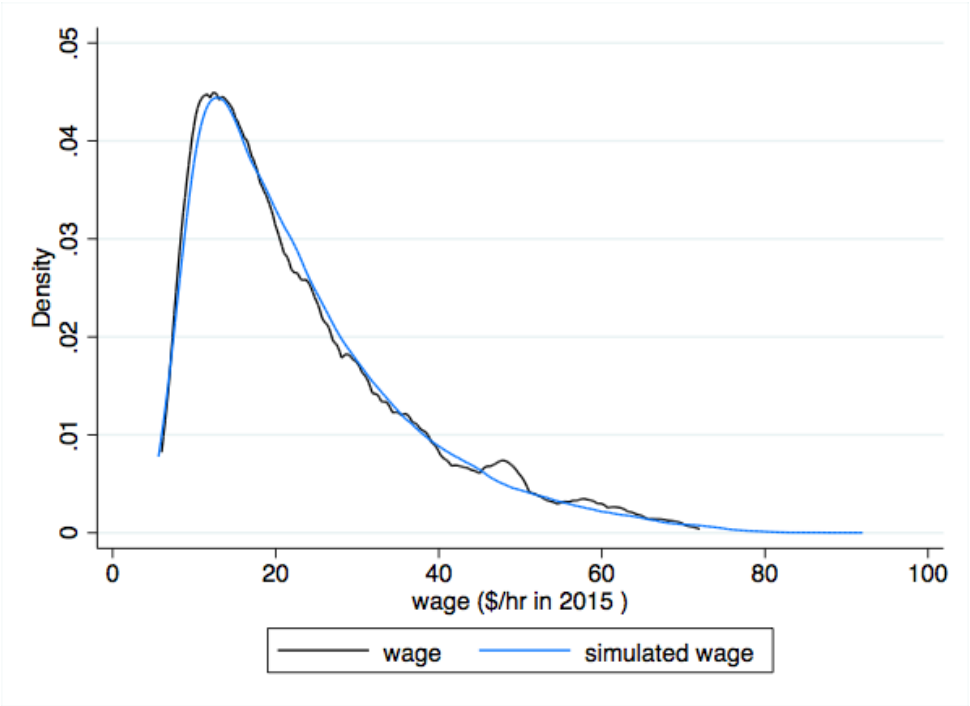


Figure 2.4: Density of simulated wages compared to the CPS

This provides a test of goodness of fit for the model and the model appears to perform reasonably well. The distributions are easy to distinguish on a couple of minor points since only the simulated distribution has any mass above $72/hr and the CPS wage distribution has clumps of excess mass at certain round numbers such as $50/hr, but otherwise the fit seems quite good. This provides confidence in using the model for the next step of calculating implied VSLs and comparing the mean of this distribution to the standard OLS estimate from above.

Using the predicted wage for each observation and the associated compensating differential for its quantile, I calculated a VSL for each draw and plotted the distribution in Figure 2.5. As expected the distribution is highly skewed with a long right tail. This is partly because the underlying distribution of wages is skewed as shown in Figure 2.4 but the effect is reinforced because higher wage earners also tend to have the largest compensating differentials for risk. About 2.9% of the sample has a negative VSL, consistent with the fact that 3 of the quantiles have a negative estimated compensating differential for risk. I opted not to censor these obviously erroneously low values because it would introduce an upward bias in estimates of the population mean VSL if I didn't also simultaneously censor estimates with coefficients too high, which are impossible to identify.

## 2.6 Implications for estimates of Population mean VSL

In this section I focus on interpreting the results of the model estimated in the previous section and decompose the difference between the standard OLS hedonic wage VSL and the quantile model mean VSL into contributions from the three factors identified earlier in the chapter. To begin the analysis we consider some sample statistics, reported in Table 2.4, for the distribution of the simulated VSLs from each of the three models of compensating differentials (raw and adjusted according to Model 1 and Model 2).

Each of the models generates a sample mean VSL substantially larger than the standard OLS hedonic wage model. The standard model sample mean VSL is $11.346 million while the unadjusted quantile model has a mean of $15.12 million (about 33% larger) and the adjusted

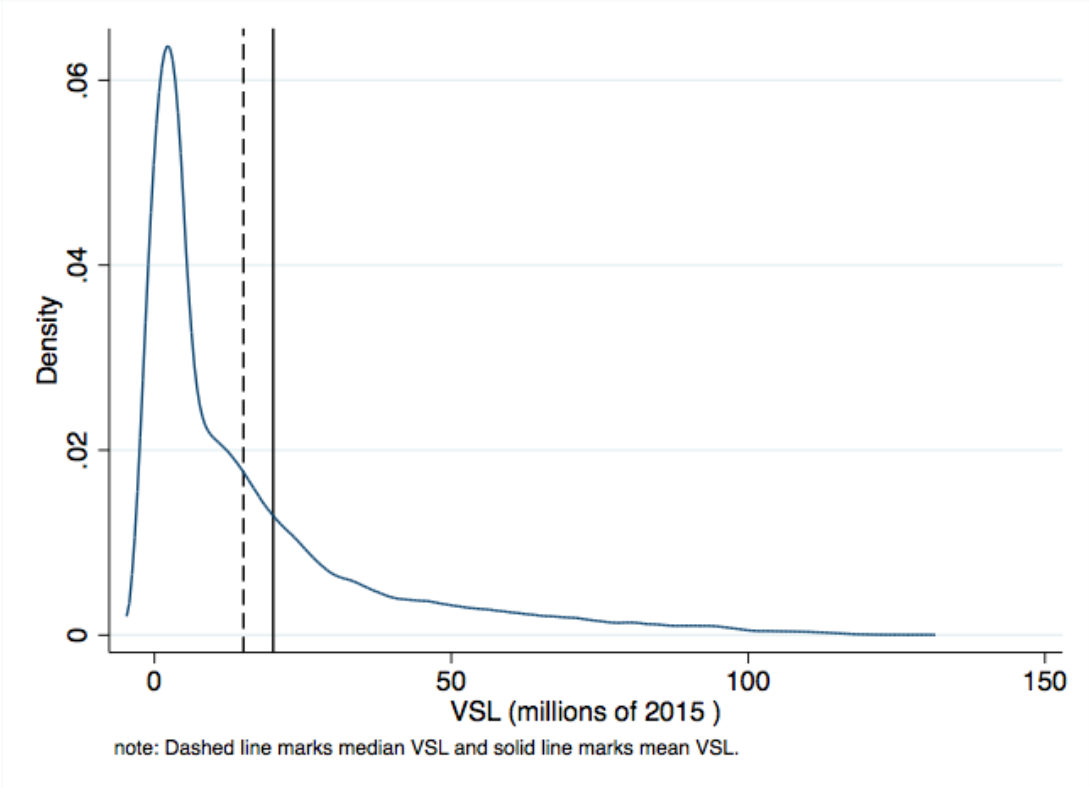


Figure 2.5: Marginal distribution of VSLs

Table 2.4: Summary statistics for simulated VSLs

| | n | mean | SD | min | max |
|---|---|---|---|---|---|
| VSL (raw) | 10,000 | 15.12 | 18.74 | -4.681 | 128.0 |
| VSL (model 1) | 10,000 | 18.04 | 25.83 | -4.681 | 211.0 |
| VSL (model 2) | 10,000 | 16.44 | 21.38 | -4.681 | 129.5 |

**Note:** Units are millions of 2015 $s.

Table 2.5: Relationship between ln$(VSL)$ and ln$(wage)$

| **Dependent:** | Raw VSL | VSL (model 1) | VSL (model 2) |
|---|---|---|---|
| ln(wage) | 1.797*** | 2.005*** | 1.931*** |
| | (0.0206) | (0.0179) | (0.0172) |
| Constant | -3.224*** | -3.782*** | -3.578*** |
| | (0.0648) | (0.0577) | (0.0557) |
| N | 6,186 | 7,084 | 7,084 |
| Standard errors in parentheses | | | |
| *** p<0.01, ** p<0.05, * p<0.1 | | | |

models have even higher averages of $18.04 and $16.44 million (up to 59% larger). The quantile model helps to correct for the uneven weighting of the compensating differentials in the OLS model and the bias from omission of the covariance term so at least one of these sources of bias must be substantial. All three quantile models have standard deviations larger than their mean, consistent with the underlying theory discussed in section 2.2. All three have a common minimum VSL of -$4.681 since the lower quantiles with are not affected by the censoring bias and are not adjusted, but the maxima range considerably since the models differ primarily in the compensating differentials for workers in higher conditional quantiles and this translates to differences in VSLs for some of these workers.

The third bias discussed previously was from omitting top coded earners who remain omitted even in the quantile model due to a lack of data on wages. VSLs for these workers can be extrapolated by using the quantile model to estimate a general wage-VSL relationship. As reviewed in section 2.4, theoretical modeling predicts that this relationship should have a roughly constant elasticity close to the coefficient of relative risk aversion (CRRA), often thought to be in the range from 1.5 to 4. Table 2.5 shows the results of regressing the simulated $ln(VSL)$ on simulated $ln(wage)$.[5]

Results are broadly similar across the three models. In the raw model the wage-elasticity of the VSL is just below 1.8, while it is barely above 2 for Model 1, and 1.93 for Model 2. It is surprising how well these elasticities fit into the narrow range predicted by theory. For brevity,

[5]The bottom 30 quantiles are excluded from this analysis since the log transformation cannot be applied to negative values. The top 10 quantiles are also excluded from the raw model since we know these coefficients are heavily biased due to censoring.

Table 2.6: Extrapolated VSL results for the three models

| Model | $E[VSL\|wage > 72]$ | $E[VSL\|wage < 72]$ | $E[VSL]$ |
|---|---|---|---|
| Raw | \$215.93 | \$15.12 | \$25.16 |
| Model 1 | \$363.70 | \$18.04 | \$35.32 |
| Model 2 | \$302.51 | \$16.44 | \$30.74 |

**Note:** Units are millions of 2015 \$s.

I illustrate how I extrapolated a mean VSL for the top-coded workers using only the raw model and just summarize the results for the other models in the tables below. The regression models in Table 2.5 relate the VSL (in millions of 2015 \$s) to the wage. For example, the raw model is $\ln(VSL) = -3.224 + 1.797\ln(wage)$. We can exponentiate both sides to get the approximate relationship $VSL = 0.0397 wage^{1.797}$ .[6] For example, for a wage of \$72/hour the model predicts a VSL of \$86.5 million, not far below the highest VSLs estimated using the quantile model, which is what we would expect since this model is derived from the quantile one.

Our object of interest is the average VSL for workers above the cutoff, or in symbols $E[VSL|wage > 72]$ which can be rewritten as $E[0.0397 wage^{1.797}|wage > 72]$ using the relationship between $VSL$ and $wage$ in the basic model. It is widely documented that the wage distribution has a long upper tail that has, approximately, a Pareto distribution. The standard technique for imputing wages is to assume that the mean above the top code is about 1.5 times the cutoff, or around \$108 in our case. The mean of a Pareto distribution is always the cutoff times $\frac{\alpha}{\alpha-1}$ where $\alpha$is the Pareto parameter so to match the standard assumption it is implied $\alpha = 3$ and thus we can approximate the wage distribution for top-coded workers as Pareto($x_m$ =72, $\alpha$ =3). To the extent rising inequality has lead the rule of thumb that wages are 1.5 times the cutoff to be conservative then the resulting estimates below will be conservative, although they turn out to substantial nevertheless. For most distributions $E[0.0397 wage^{1.797}|wage > 72]$ would needed to be estimated by simulation since it is hard to get an exact formula for a random variable raised to a random exponent, but, fortunately, the moments of the Pareto distribution have a simple closed form, $E[X^a] = \frac{\alpha}{\alpha-a} x_m^a$. Using the closed-form formula for the moments we have that $E[0.0397 wage^{1.797}|wage > 72] = 0.0397 \frac{3}{3-1.797}(72)^{1.797} = 215.41$ million for the raw model.

[6]Exponentiation is not a linear transformation, so there should be a constant term reflecting uncertainty in the regression parameters that slightly raises the estimated VSL but it is omitted for clarity.

Table 2.6 shows the estimates for the VSL above the cutoff for each of the three models as well as the population mean VSL that is implied by combining the estimates for the sub-populations. Although the top-coded workers only make up 5% of the workers they have an outsized impact because of their large VSLs. For the baseline model, for instance, the average VSL among the top-coded workers is $215 million, which helps to pull the estimated average VSL for workers up to $25.16 million, more than double the standard hedonic wage model estimate and more than 50% larger than the average for the other 95% of the workers.

It is natural to ask whether these nine-figure mean VSLs are plausible since they are an order of magnitude larger than typical estimates in the hedonic wage model literature. I think are for three reasons. First, the order of magnitude makes sense based on the underlying economics. These workers have wages that are, by assumption, $108/hr on average or almost five times the sample average. Since the wage-VSL elasticity is around 2 the VSL will scale up by almost 2% for every 1% increase in wages. As a result, when if a group of workers has wages 400% larger than normal we would expect their VSLs to be more than 800% larger or about an order of magnitude larger.

Another helpful perspective is to compare the weight that the top 5% of earners have in the wage distribution with the weight they have in the VSL distribution. In the CPS sample, which probably suffers from income underreporting and may have a fatter tailed wage distribution than I assumed earlier, the top 5% of workers earn about 20% of total wage income. As a result is seems implausible that with 20% of the total income they would contribute any less than 20% toward total demand for risk reduction and, since safety is a luxury good, plausibly would contribute an oversized share or 30-60% toward the total demand. In my estimated models they contribute 43-56% to the total demand for risk reduction, fitting this reasoning. If the top 5% contributes around 50% toward total demand then it is a mathematical necessity that their average willingness to pay (i.e. VSL) must be roughly an order of magnitude larger than the median worker's.

Finally, while there is no solid empirical evidence on the VSL for high earners there is some anecdotal evidence that some wealthy people have extraordinarily high VSLs. First, there is a

Table 2.7: Breakdown of the relative importance of the biases

| Model | $\Delta\delta$ | $cov(\delta, wage)$ | Excluding top-coded |
|---|---|---|---|
| Raw | +\$0.314 (2.8%) | +\$3.50 (31.0%) | +\$10.04 (88.5%) |
| Model 1 | +\$1.739 (15.3%) | +\$4.99 (44.0%) | +\$17.28 (152.3%) |
| Model 2 | +\$0.942 (8.3%) | +\$4.29 (37.8%) | +\$14.30 (126.0%) |

**Note:** Units are millions of 2015 \$s. Percent changes in parentheses.

thriving medical tourism industry which suggests people with means are willing to pay significant sums for small amounts of mortality risk reduction. The medical care at top academic medical centers is probably only marginally more effective than at other first-world hospitals, but executives from around the U.S. and the ultra-wealthy from around the world pay enormous markups to get such marginally better medical care. Future research, pending data availability, may be able to tease out information on the elasticity of this demand by looking at how it varies with exchange rate fluctuations. Second, there are large markets for ransom and extortion in some countries and ransoms demanded tend to be highly elastic to the victim's resources. Smith (2018) reports that the typical ransom charged after a kidnapping in Mexico is around \$500 but there are documented cases of ransoms of \$30 million or more being paid and there are documented cases of larger ransoms being paid in richer countries. Li KaShing's paid 2 billion (HKD) or 258 million (USD) to ransom his son (Vines 1997) or about half a billion in todays dollars and Argentinian grain traders Juan and Jorge Born were kidnapped and ransomed for \$60 million in 1974 which would be well over \$300 million in today's dollars. These ransoms are hard to interpret since they are paid in highly emotional situations with concentrated risk so they do not correspond to a VSL or any closely related concept, but they do illustrate the fact that differences in ability to pay plausible lead to differences in willingness to pay for safety of several orders of magnitude.

Table 1.7 provides a breakdown of how much each of the three sources of biases contributed to the final difference between the standard VSL estimate and my final quantile model estimates with extrapolated VSLs included. The baseline standard OLS estimate of the VSL is \$11.364 million. This increases by 2.8-15% depending on model when we allow for heterogeneity in the compensation for risk ($\delta(q)$) across conditional quantiles but ignore the possible

covariance between the wage and $\delta$. It was difficult to a priori assess the importance of this bias, so it is reassuring that the bias is not spectacularly large. Factoring in the covariance term is quantitatively more important, increasing the estimated VSL by 31-44% depending on the model. Earlier calibrations suggested this term would increase the estimated VSL by around 50% so these estimates seem reasonable and certainly quantitively significant. Finally, the third and most important source of bias was ignoring high wage earners in the estimates. In the baseline model with an elasticity of 1.8 the top-coded workers pull the average VSL up by 88.5% while in Model 1 with an elasticity slightly above 2 the top-coded workers pull the average up by 152.3%. It is clear that the population mean VSL depends critically on this elasticity so future work pinning it down will be especially valuable. For now, my tentative conclusion is not to emphasize any one number in the table as the best estimate for the population mean VSL, but rather to emphasize the importance of understanding and modeling heterogeneity when estimating hedonic wage models since the bias from not doing so can clearly be substantial.

## 2.7 Conclusion

The theory of the VSL predicts that it will be heterogenous across people since it is a rescaled measure of demand and demand is influenced by a multitude of factors, perhaps most importantly income and preferences. This chapter has emphasized that this heterogeneity is not only important in its own right, but it is also essential to take into account when estimating a population mean VSL. I identified three reasons that standard hedonic wage regressions might yield estimates substantially below true the population mean VSL.

First, the standard OLS regression implicitly weights the observations by the conditional variance of the risk for each combination of covariates. This means that estimates of compensating differentials do not necessarily reflect the preferences of all workers equally and appear to put most of the weight on male workers with relatively little formal education who work in blue-collar jobs. To the extent these workers require more, or less, compensation to take on fatal injury risk than the average person, the estimated VSL will be biased. Second, since we expect

safety to be a luxury good, VSLs will rises rapidly with income and thus the wage premiums demanded by workers will vary significantly across the wage distribution. To the extent there is a correlation between the compensating differentials for risk and wages the standard formula for the sample VSL is biased because it leaves out an important covariance term. Third, most estimates of the population mean VSL ignore top-coded observations due to data limitation. If the VSL is not wage-elastic then systematically leaving out high earners is not particularly problematic, but to the extent the VSL is wage-elastic leaving out even a small fraction of the top earners can severe bias the estimated mean for the population.

I showed that all three of these potential problems are likely to be quantitatively important when there is substantial heterogeneity in the wage premiums that workers require to take on risk. I showed that the condition for this to be the case is that the wage-elasticity of the VSL is not equal to one. The empirical literature of how the VSL varies with wages and income is mixed but theory makes clear that we should expect a large wage-VSL elasticity and thus that these biases could be quantitatively significant. I extended the standard hedonic wage model into a quantile regression framework, building on Evans and Schaur (2010), to estimate a joint distribution of wages, compensating differentials and VSLs and then used this model to assess the quantitative significance of the three biases. The first bias, from variation in compensation wage premiums, is tolerable (2-15%) but the later two biases were substantial. Factoring in the covariance between risk premiums and wages increased the estimated VSL in the sample by 33 to 44% and including extrapolated estimates of the VSLs for top-coded workers approximately doubled the average VSL in all specifications.

My main conclusion is that most VSL estimates from hedonic wage models are likely biased downward significantly due to ignoring heterogeneity so an important goal for future work is to better understand and model this heterogeneity, perhaps by refining the quantile hedonic wage model framework as one approach to modeling heterogeneity and explore to what extent some of the stronger assumptions can be relaxed. Finally, since the key parameter controlling the scale of the bias is the wage-VSL elasticity, it is clear that a better understanding of this parameter would be valuable.

# Chapter 3

# Do physician shortages explain long wait times?

## 3.1 Introduction

There is an ongoing debate about the adequacy of the current supply of physicians in the United States and how supply will need to evolve in the future to meet growing health care needs. Cooper et al. (2002) estimate that there is a substantial shortage based on a model of demand that emphasizes income and demographics and argue that the steady aging of the population and rising incomes will only aggravate this shortage over time. Goodman et al. (2008) criticize this approach to modeling, noting that there is substantial variation in physician supply across the country, suggesting that a proper distribution of current health resources would be sufficient. Bodenheimer and Smith (2013) reconceptualize the problem by arguing that whether the physician supply is adequate or not is hard to determine from studying raw inputs since productivity appears to vary dramatically across the U.S. They suggest that reforms to reduce paperwork, improve coordination, and prevent duplication would substantially increase total factor productivity, potentially turning projected shortages into surpluses in the coming years. This line of argument dates back to at least 1993 when Wennberg et al. used the patient-doctor ratios at successful HMOs such as Kaiser Permanente as a benchmark

to suggest the required supply of physicians for the coming years. Green et al. (2013) use a simulation-based supply-demand model to support this claim, emphasizing how expanding scope of practice and use of technology can increase productivity.

This debate has taken on new urgency as the Affordable Care Act (ACA) has expanded insurance coverage for millions of Americans and there is continued interest in developing a universal system of coverage that would provide coverage for the 30 million residents who lack it in any given year. Due to the well-documented relationship between insurance coverage and demand for medical care these changes are likely to contribute to an ongoing surge in demand for physician services, amplifying higher demand from an aging and increasingly frail population. In this context it is reasonable to ask whether supply constraints will replace lack of insurance as the dominant impediment to getting care. Rhodes et al. (2014) conducted an audit study across 10 states and thousands of primary-care providers in 2013 and found that many doctors were not accepting new patients, particularly if the patient had Medicaid coverage. Furthermore, Massachusetts, which has nearly universal coverage, stood out in their study for its abnormally long wait times and low rates for accepting new patients, despite its large supply of physicians. Commentators have voiced concerns that the rest of the states could suffer from similar problems in the coming decade. Others commentators, however, have noted that most new Medicaid enrollees claim to have a usual source of care (Kenny 2014) and that the long wait times in the Boston metropolitan area pre-date the insurance reform.

In this chapter I investigate how variation in physician supply impacts both physicians and patients on dimensions such as wait times for an appointment, measures of access to care, and physician incomes. While the evidence is noisy and the topic deserves further research I find suggestive evidence that greater physician supply lowers physician incomes and stronger evidence that physician supply impacts the style of medical practice. I find little evidence of an effect of physician supply on access to care, even using methods to account for endogenous sorting of physicians, but the estimates are not precise enough to rule out a substantial effect The rest of the chapter is structured as follows. Section 3.2 discusses several theories for how local physician supply could impact the market for physician services, touching on both a

benchmark competitive model and frictional models with endogenous productivity. Section 3.3 introduces the empirical framework for this reduced form analysis of this chapter and presents the two identification strategies used to deal with endogeneity. Sections 3.4 and 3.5 present results from the estimation of the econometric models, discussing the estimated impacts of physician supply on outcomes for physicians and for patients. Section 3.6 briefly concludes.

## 3.2 Theories

The traditional view of the effects of physician supply emphasizes economic intuition from the theory of competitive markets where a large number of price-taking sellers competes for business from a large number of price-taking customers. In this model growth in the supply of licensed physicians in a local healthcare market leads to lower prices and more total volume of services but each individual physician provides less care. Modest deviations from this framework would have similar implications as the basic mechanism is that limited demand from customers mixed with greater supply weakens the bargaining power of each provider (Friedman and Kuznets 1954). This view helps to explain why the American Medical Association (AMA) long opposed expanding the physician supply and worked to restrict competition through regulations. Petersen et al. (2014) also document that, consistent with this model, many state medical boards explicitly banned non-citizens from the practice of medicine, evidently in order to reduce competition, until their power to do so was eliminated by the In re Griffiths decision (Oyez 2019).

The pattern of support and opposition to scope of practice regulations among healthcare occupations also suggests some validity to this theory. Physician groups have generally advocated limiting the scope of practice for nurse practitioners and others with advanced training and have prevented psychologists from prescribing neuroleptics in most states. The effects of these regulations remains an unsettled and active area of research but the motivation behind them seems clear (Kleiner 2016).

A counterpoint to this view notes that healthcare markets are frictional and, in practice, appear to be far from the ideal of competitive markets (Cutler and Zeckhauser 2000). Medicare sets prices for most physician services administratively and they are uniform across the nation aside from adjustments for local input prices. Private insurance reimbursement rates tend be strongly influenced by the Medicare fee schedule, adjusting almost one for one in responses to updates (Clemens and Gottlieb 2017). Physicians can also pool together in group practices with little concern about antitrust oversight and use this as leverage when bargaining. There is some evidence that such consolidation leads to higher prices, explicitly violating the competitive norm of price taking (Hausman and Lavetti 2018). Consistent with these more sophisticated theories of market organization, the AMA's position on expanding the physician workforce has changed in recent years, switching to supporting expanded opportunities for training in order to meet pending shortages. It remains unclear why the AMA reversed its stance but it suggests that it no longer seems supply as a threat to existing physician's financial interests. The competitive model also runs up against surprising prima facia evidence against its validity: local areas with greater supply do not appear to have any improvements in access to care. Figure 3.1 shows that wait times are actually slightly longer in areas with more physicians per capita and a few states, including Massachusetts and Connecticut. in particular stand out for their larger supply of physicians and abnormally long wait times for appointments.

This upward slope on the graph could be a result of noise and is explored in more detail later in this chapter but there is a simple economic explanation for the pattern. Physicians in the competitive model have a financial incentive to sort based on local demand. As the population grows or becomes older or sicker in an area there will be more demand for medical care, drawing in more physicians, but only enough to roughly balance the greater need. In equilibrium there will be variation across states in the number of physicians per capita as well as other healthcare inputs like hospital beds per capita that balances the variation in demand, with little variation in wait times or utilization rates (e.g. fraction of hospital beds filled).

This spatial equilibrium view, however, contrasts with the enormous variation in supply that appears to have no relationship with needs. Massachusetts and Connecticut, as noted ear-

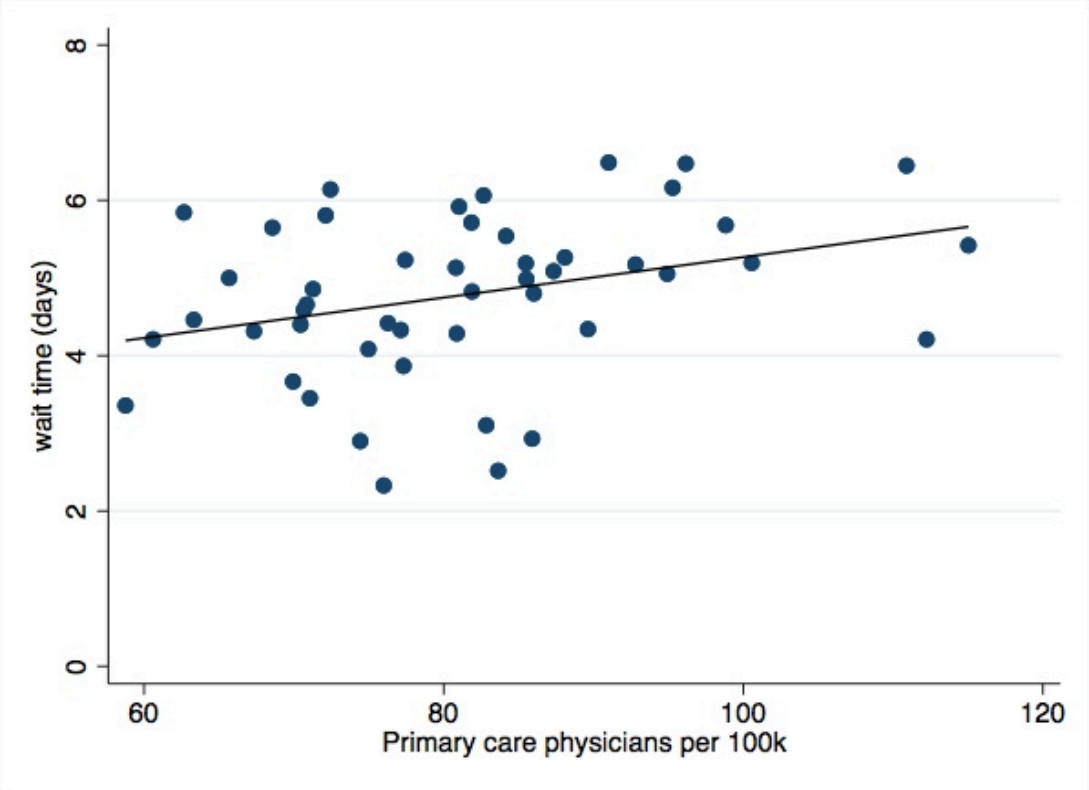


Figure 3.1: Wait times vs physician supply

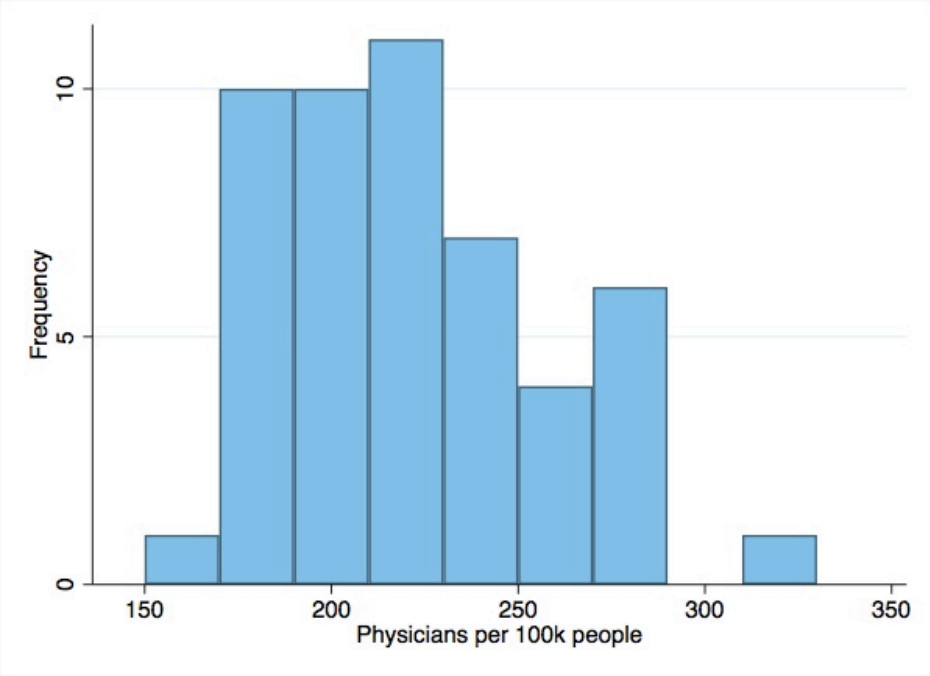


Figure 3.2: Physician density varies significantly across states

lier, have particularly large numbers of physicians per capita but are two of the richest and healthiest states. The patterns are even less explicable when looking at sub-state patterns. As Wennberg et al (1998) note, an affluent suburb of New York (White Plains) has the most physicians per capita of any hospital referral region in the country. This strongly contrasts with intuition based on obesity rates, rates of drug overdose deaths, and other population health indicators which would suggest parts of South Dakota, Mississippi, or West Virginia would garner the greatest number of physicians in a spatial equilibrium dominated by demand-side factors.

In addition to potential maldistribution, the underlying variation in physician supply across states appears difficult to reconcile with variation in demand. There is nearly three to one variation in the number of primary care physicians (PCPs) per capita across states and similar variation in physicians in general per capita as illustrated in Figure 3.2. The patterns of distribution among specialties are also hard to reconcile with either the spatial equilibrium model as we would expect less variation in basic inputs required by all people for primary care and more variation in specialty inputs only needed for particularly sick patients. Surprisingly, however,

the coefficient of variation is roughly the same for PCPs (CV = 0.160) and specialists (CV = 0.181).

Another approach to the market for physician services emphasizes variation in productivity (Bodenheimer and Smith 2013). Potentially endogenous variation in the way in which physicians practice—their standards for diagnostic testing or the typical frequency of check up visits—influences the efficiency of providing health care in different areas and then indirectly the local supply. In effect, sine some places have low levels of productivity they require more physicians to provide the same care. In this framework there can be substantial variation in inputs across localities without corresponding variation in pay or measures of access to care. Furthermore, there is little reason for physicians to worry about supply as changes could lead to endogenous adjustments in practice styles, becoming more efficient when required by negative supply shocks and less efficient when positive supply shocks allow a slow pace of work or more aggressive treatment. This theory ties in with the hypothesis that physicians can induce demand for their own services, recommending more aggressive treatment and more frequent follow-up visits when their schedules become slack and these patterns of induced demand might become ossified as variation in practice styles over time.

These theories are not necessarily mutually exclusive and so in this chapter I focus on a preliminary investigation to answer reduced form questions about how changes in physician supply influence doctors and patients. Does a larger supply of physicians lower physician incomes as the competitive model might predict or does it appear unimportant as induced demand would suggest? Does having more physicians in an area lead to shorter wait times or more people to have a usual sources or care? These questions are important as an input into the policy planning process and can also provide a basis for future structural work that explores mechanisms in competing models.

## 3.3 Empirical Strategy

The basic empirical approach is this chapter is to use regression analysis to undercover associations and, under identifying assumptions, treatment effects of changes in physician supply. All specifications under in the empirical analysis sections take the form:

$$Y_{ij} = \alpha + \beta \ln(\text{physicians per 100k}) + \gamma X_{ij} + \delta Z_i + \varepsilon_{ij}$$

where $i$ indexes the state and $j$ indexes either physicians or patients, depending on the source of the measure, $X$ is a vector of patient or doctor level controls, $Z$ is a vector of state-level controls, and $\varepsilon$ is an error term. $\beta$ is the coefficient of interest.

Standard errors are clustered at the state level since the predictor of interest, ln(physicians per 100k) is measured at the state level and, as discussed in chapter 1, this can lead severe underestimation of standard errors without accounting for correlation across observations within a state.

The basic threat to validity in interpreting $\beta$ as a treatment effect is the possibility of omitted variable bias. Suppose, for example, that some states have sicker patients due to local diet and its impact on health. Following Angrist and Pischke (2009) we can quantify the size of the bias by contrasting two regressions, the long form that includes the omitted variable $O$, $Y_{ij} = \alpha + \beta_{long} \ln(\text{physicians per 100k}) + \gamma X_{ij} + \delta Z_i + \mu O_i + \varepsilon_{ij}$ and the short form, above, which excludes it.

An auxiliary regression helps illustration the relationship between between $\beta_{long}$ and $\beta$: $\ln(\text{physicians per 100k}) = \pi_1 X_{ij} + \pi_2 Z_i + \pi_3 O_i + \nu_{ij}$. In this setup it can be shown that $\beta = \beta_{long} + \mu\pi_3$. If the long regression is the correct specification then $\mu\pi_3$ quantities the bias of $\beta$. For the example above it seems likely that areas with sicker patients should draw in more physicians ($\pi_3 > 0$) and if the outcome is mean physician income, will lead to more business and higher incomes ($\mu > 0$) so the omitted variable bias will lead to an overestimate of the effect of supply on incomes. Since that effect is likely to be negative this would bias the coefficient estimate toward zero and finding no effect.

Unfortunately, there are many potential omitted variables, both at the market (state) level and the individual level, that could potentially bias the treatment effects estimates either up or down so it is hard to know *a priori* whether estimates are too high or too low. To some extent these biases can be mitigated with the inclusion of controls for factors such as the fraction of the area that is obese, a proxy for health status, but these controls and mostly imperfect proxies and some omitted factors cannot be proxied for. Factors at the physician or household level are likely to be problematic in particular since we would expect wide variation in many traits that cannot easily my measured and included in the regression models. Physicians presumably differ significantly in abilities, such as the ability to do specific procedures or the speed at which they can do procedures, and this will influence their incomes, but I do not have data to proxy for ability beyond the most basic controls of experience, specialty, and board certification. I use two approaches in this chapter to deal with the omitted variables problem, each discussed in turn below.

### 3.3.1 Instrumental variables

The first and simplest approach is to find a natural experiment, represented by an instrumental variable, that causes variation in state-level physician supply but plausibly has no impact downstream outcomes except through its effect on supply. Such instruments are difficult to find and especially difficult to validate, but I propose one based on the economic theory of regulatory capture. In particular, some states have autonomous, or self-financing medical boards, due to conditions during their initial chartering in the late 19th and early 20th centuries. Boards have wide scope to regulate medical licensing, particularly for physicians trained overseas, called international medical graduates (IMGs), and the autonomous boards have less supervision in how they use their power to limit supply. As a result, the seemingly arbitrary characteristic of medical board financing should have a significant influence shaping local supply. The rest of this subsection lays out the theory behind how medical boards, in general, can limit physician supply and why we would expect autonomous medical boards to be more aggressive in limiting supply and then shows some evidence on the relevance and validity of the

proposed instrument. Most of the discussion on the mechanism for medical board function builds on Petersen et al (2014) who kindly provided the data on medical board characteristics.

IMGs are licensed after a multiyear process used to prove their competence to practice medicine. The first step is passing the three stages of the U.S. Medical Licensing Exam (USMLE), a standardized exam given in three parts on basic science, clinical knowledge and clinical skills that is used to "ensure that all licensed MDs have passed the same assessment standards – no matter in which school or which country they have trained" (Federation of State Medical Boards and National Board of Medical Examiners 2019). Just 42.6% of IMGs pass the exam.

The second and longer step is to complete an internship year training at an academic medical center in the United States. Obtaining an internship is difficult as there are far more applicants than positions available each year in this Medicare-funded program. Like step 1 this step applies to both IMGs and American-trained physicians, but most states allow American physicians to practice independently after completing this stage of their training. In contrast, IMGs may or may not have to undergo further training before they can practice independently, applying for a residency in a specialty of medicine and competing at least two more years of training before being granted a full license. This is the primary mechanism by which I expect autonomous medical boards to be able to influence supply, although there might be other mechanisms since boards have wide latitude in interpreting foreign credentials such as degrees that are needed before an applicant can even quality to start the licensure process.

Why would state medical boards vary in their requirements at the last stage of IMG licensure? It seems likely that they vary in the extent to which they are subject to regulatory capture (Stigler 1971). The theory of regulatory capture posits that regulatory bodies charged with the public interest will often instead work to advance a special interest of the groups that capture power over the agency. The classic examples are the so-called revolving doors at federal regulatory agencies as staff members of the agencies depart for lucrative offers in the industries they formerly regulated creating incentives to favor potential future employers while working as a regulator. [Further evidence comes from the fact that many heads of regulatory agencies are former executives of the industry they regulate. A recent example is appointment of

Robert G. Cameron, an executive at the Pennsylvania Higher Education Assistance Agency, a federal loan servicer reportedly "at the center of every major industry scandal" and known for lax compliance with consumer protection laws, was recently appointed the student loan ombudsman within the Consumer Financial Protection Bureau (Kreighbaum 2019). Of course, another explanation for these patterns are these only a limited number of people have expertise in administrative law and regulations specified to any industry, so naturally companies and the government compete for this limited pool of experts.]

Capture of a regulator and defense against it both rely on collective action, on behalf of an interest group or on behalf of the general public, which is difficult due to incentives to free ride (Olson 1971). It is generally easier, however, for the special interest to organize since benefits tend to be concentrated for them while costs for the general public are diffuse, amplifying their incentive to not bother getting involved in the issue. Furthermore collective action on behalf of the special interest is even easier if a small board is charged with oversight, as with medical licensing and sanctions. These boards have only 6-25 members with an average of just 11.5.

So far all of these factors suggest that regulatory capture should be a problem in states regardless of how their medical board is organized or financed. Weingast and Moran (1983) further refined the theory in a study that focused on self-financing regulatory agencies. Since they are self-funded they have less interaction with and oversight by a more publicly accountable body such as the legislature and can fly under the radar even if severely captured. Weingast and Moran (1983) tested these predictions with empirical evidence across industries to validate the theory. Under this theory if medical boards are autonomous, meaning they are financed by charging for renewal of licenses, continuing education, and other services as opposed to with annual appropriations from a state legislature, then they should have more scope to limit entry by IMGs. I use a strict rule, following Petersen et al. (2014) and code a medical board as autonomous only if it receives no appropriations from a state regardless of the share of its budget that this appropriation makes up. The intuition is that it is not so much the amount of money appropriated, which is trivial within a modern state budget, but rather the oversight and potential for investigation of board practices that come with appropriation requests that mitigates

Table 3.1: Summary statistics for the state level variables

| | N | mean | sd | min | max | p25 | p75 |
|---|---|---|---|---|---|---|---|
| autonomous | 50 | 0.600 | 0.495 | 0 | 1 | 0 | 1 |
| Physicians per 100k | 50 | 221.8 | 36.37 | 164.3 | 324.2 | 194.2 | 243.2 |
| Primary care per 100k | 50 | 81.96 | 13.10 | 58.77 | 115.0 | 71.28 | 89.63 |
| IMGs per 100k | 50 | 44.81 | 26.74 | 8.21 | 130.75 | 30.95 | 67.77 |
| seniors (%) | 50 | 13.32 | 1.671 | 7.700 | 17.40 | 12.40 | 14.30 |
| White (%) | 50 | 71.40 | 15.48 | 22.70 | 94.40 | 60.30 | 83.10 |
| Black (%) | 50 | 10.33 | 9.557 | 0.400 | 37 | 2.900 | 15.40 |
| Hispanic (%) | 50 | 10.61 | 9.980 | 1.200 | 46.30 | 4.200 | 12.40 |
| income per capita | 50 | 44,083 | 6329.6 | 33072 | 60491 | 39293 | 47573 |
| obesity (%) | 50 | 25.42 | 2.935 | 21 | 34 | 23 | 27.00 |
| smoking (%) | 50 | 19.84 | 3.625 | 10.60 | 28.30 | 17.30 | 22.50 |
| UE rate | 50 | 5.978 | 1.396 | 2.600 | 8.700 | 4.900 | 6.900 |
| uninsured (%) | 50 | 14.32 | 4.087 | 3.757 | 24.26 | 11.76 | 16.85 |

regulatory capture. The rest of this section is focused on testing whether the theory above has validity, namely whether states with autonomous medical boards actually have fewer physicians. Summary statistics for the instrument, state level controls, and the number of physicians per 100,000 residents are presented in table 3.1. The state level controls were obtained from IPUMS USA and the counts of physicians per state and state populations were obtained from the American Association of Medical Colleges' 2009 State Physician Workforce Data Book. A few things about the data are worth noting. First, thirty of the 50 states have autonomous medical boards so there are large groups of states with and without self-financing boards. Second, the number of IMGs per 100,000 residents is, consistent with the theory, significantly more variable as indicated by the coefficient of variation.

State-level first stage regressions of the instrument on the logged number of physicians per 100,000 residents along with various sets of controls are shown in Table 3.2. As the theory predicts, autonomous medical boards are associated with fewer physicians per capita. In the basic specification the effect is about 15 log points fewer or about a 15% decrease in supply due to having a self-financing medical board. This grows to 18.2 log points when some basic demographic controls are added but then shrinks to just under 10 log points when further economic and health controls are added. The pattern is similar when the data are restricted to

Table 3.2: First stage for medical board autonomy instrument

| **Dependent:** | **ln(physicians per 100k)** | | | **ln(primary care per 100k)** | | |
|---|---|---|---|---|---|---|
| autonomy | -0.151** | -0.182** | -0.0957** | -0.125** | -0.146** | -0.0997** |
| | (0.0411) | (0.0408) | (0.0302) | (0.0422) | (0.0380) | (0.0297) |
| seniors (%) | | 0.0366** | 0.0326** | | 0.0332** | 0.0295** |
| | | (0.0121) | (0.00858) | | (0.0113) | (0.00845) |
| White (%) | | -0.00107 | 0.000523 | | -0.00220 | -0.000450 |
| | | (0.00208) | (0.00135) | | (0.00194) | (0.00133) |
| Hispanic (%) | | 0.000847 | -0.00233 | | -0.00295 | -0.00574* |
| | | (0.00308) | (0.00232) | | (0.00287) | (0.00228) |
| Black (%) | | -0.00354 | -0.00174 | | -0.00802** | -0.00811** |
| | | (0.00253) | (0.00199) | | (0.00236) | (0.00196) |
| mean income | | | 3.16e-05** | | | 3.59e-05** |
| | | | (5.49e-06) | | | (5.41e-06) |
| obesity (%) | | | -0.0303 | | | 1.050 |
| | | | (0.947) | | | (0.933) |
| smoking (%) | | | -0.473 | | | -0.605 |
| | | | (0.664) | | | (0.654) |
| UE rate | | | 0.0132 | | | 0.00538 |
| | | | (0.0106) | | | (0.0104) |
| uninsured (%) | | | -0.802+ | | | -1.592** |
| | | | (0.436) | | | (0.429) |
| Constant | -6.033** | -6.399** | -6.949** | -7.044** | -7.202** | -7.657** |
| | (0.0319) | (0.240) | (0.351) | (0.0327) | (0.224) | (0.345) |
| N | 50 | 50 | 50 | 50 | 50 | 50 |
| $R^2$ | 0.219 | 0.387 | 0.789 | 0.154 | 0.453 | 0.789 |
| F-stat | 13.5 | 19.9 | 10.1 | 8.8 | 14.8 | 11.3 |

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

PCPs which is important as some of the measures of access to care used later are only applicable for primary care.

A general issue with the two stage least squares (2SLS) method of estimation is that while the coefficient estimates are consistent they are biased in finite samples. The bias comes from the fact that the method requires fitting a first stage which can be overfit in small samples leading to a biased second stage. The bias shrinks with the sample size since large samples can limit overfitting but in this study the samples are quite limited since there are only 50 states and thus, even in later regressions where technically there are thousands of observations

from individual physicians and households, there are effectively only 50 observations after clustering. The size of the bias is also reduced when the instrument is a powerful predictor of the variable of interest since in that case the fit is less influenced by noise and reflects the genuine correlation of the variables. The rule of thumb is that if the squared t-statistic on the instrument in the first stage, or equivalently the F-statistic, is greater than 10 then the bias from finite samples is unlikely to be problematic (Stock, Wright, and Yogo 2002). The F-statistics for the instrument are shown at the bottom of Table 3.2 and are slightly larger than 10 when the full set of controls is included. Overall then it appears bias is not necessarily a problem but the instrument does not pass this test by a wide margin either so the potential is worth keeping in mind.

The pattern of the coefficient shrinking toward zero as more controls are added is also concerning as, if the instrument is exogenous, the point estimates should not be impacted by the inclusion of controls—the controls mainly serve the function of improving the precision of the estimates by shrinking the unexplained variation in the dependent variable. This could indicate that the instrument is not exogenous, a genuine concern, or be a byproduct of the small samples. I think the later is at least as likely as the former since low power and imprecise estimates are problems throughout this study and the estimates have overlapping confidence intervals, indicating that they are unlikely to be different in statistical terms.

One way to test the validity of the instrument is to look at the effect of autonomous medical board specifically on IMGs. Based on the hypothesized mechanism discussed earlier we should see substantial effects of self-financing on the supply of IMGs and minimal effects on locally trained physicians. Table 3.3 shows that this is indeed the case. IMG density is reduced by over 30 log points while non-IMG density is largely unaffected with negative but very small and insignificant estimated effects.

A final note about the instrumental variables method is important. This econometric method estimates a particular type of treatment effect known as a Local Average Treatment Effect or LATE. Intuitively, if treatment effects vary, it tells us about the effect of increasing supply in the kinds of states influenced by the instrument and for the kinds of supply changes the treatment

Table 3.3: Validity test for medical board autonomy instrument

| **Dependent:** | **ln(IMG per 100k)** | | **ln(non-IMG per 100k)** | |
|---|---|---|---|---|
| autonomy | -0.311+ | -0.310** | -0.012 | -0.046 |
| | (0.161) | (0.107) | (0.0455) | (0.0431) |
| seniors (%) | | 0.219** | | -0.0106 |
| | | (0.0306) | | (0.0123) |
| White (%) | | 0.000576 | | -0.000255 |
| | | (0.00480) | | (0.00192) |
| Hispanic (%) | | 0.0413** | | -0.0118** |
| | | (0.00825) | | (0.00331) |
| Black (%) | | 0.0147* | | -0.00573+ |
| | | (0.00709) | | (0.00284) |
| income | | 6.18e-05** | | 1.92e-05* |
| | | (1.96e-05) | | (7.84e-06) |
| obesity (%) | | 5.149 | | -1.261 |
| | | (3.374) | | (1.352) |
| smoking (%) | | 4.367+ | | -1.029 |
| | | (2.367) | | (0.949) |
| UE rate | | 0.0633 | | -0.00571 |
| | | (0.0377) | | (0.0151) |
| uninsured | | 5.042** | | -1.681* |
| | | (1.552) | | (0.622) |
| Constant | -7.556** | -17.84** | -6.318** | -5.152** |
| | (0.125) | (1.250) | (0.0353) | (0.501) |
| N | 50 | 50 | 50 | 50 |
| $R^2$ | 0.072 | 0.793 | 0.101 | 0.596 |

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.1

induces. This could be important for interpretation as IMGs completed part of their training overseas and thus might have a different style of practice than American-trained physicians. They could be, for instance, more or less likely to practice defensive medicine and more or less likely to response to financial incentives created by payment reforms. To some extent this theory can be tested as the HTS physician surveys (discussed below) can be used to build models where IMG status is a predictor of various behaviors. For the most part such models show significant but modest differences between IMGs and other physicians. IMGs, for instance, have lower incomes conditional on specialty and spend about 5% more time on patient care, all else equal. As a result, the effects of supply shocks estimated here are not necessarily the same as the effects of a supply shock from training more physicians at American medical schools.

### 3.3.2 Coefficient stability and selection

My second approach to making inferences about treatment effects is based on the bounding approach developed in Oster (2019) which builds on Altonji, Elder, and Taber (2005). The intuition for this approach comes from the omitted variable bias formulas discussed at the start of section 3.3 which showed that comparing estimates from short and long regressions is informative about the bias from a known omitted variable. The short and long regressions cannot, of course be estimated for variables known to be important but for which data does not exist, but Oster shows that comparison of the short and long regressions for known controls can be informative about the bias from excluded controls under certain assumptions. Specifically, she shows that her assumption imply that the change in a coefficient between the short and long regression can be used to calculate a limit on how much more the coefficient could change if further controls were added. I briefly review the basic framework below.

Consider a regressor of interest $X$, that is correlated with some other factors that influence $Y$: variables $C$, the control, and $O$, the omitted variable. Assume that the full, correct model is $Y = \beta X + \gamma C + \alpha O + \epsilon$ . The feasible short and long regressions are $Y = \beta_{short} X$ and $Y = \beta_{long} X + \gamma C$. Oster's proportional selection assumption is that $\delta = \frac{\sigma_{OX}/\sigma_O^2}{\sigma_{CX}/\sigma_C^2}$ is constant and small, meaning that the effect of $O$ on $X$ is proportional to the effect of $C$ on $X$ and not too much larger.

In intuitive terms $\delta$ tells us how much more important $O$ is as a predictor of $X$ compared to $C$. Oster suggests that often a plausible assumption is $\delta = 1$.

Defining $R_{max}$ as the $R^2$ from the full model, $R_l$ as the $R^2$ from the long regression and $R_{ss}$ as the $R^2$ from the short regression, Oster shows that under some regularity assumptions $\beta_{long} - \delta(\beta_{long} - \beta_{short})(\frac{R_{max}-R_l}{R_l-R_s})$ is a consistent estimator for $\beta$, the unbiased treatment effect of $X$ on $Y$ with the full set of controls. A similar results holds up under weaker assumptions. She suggests using the formula not as an estimator for $\beta$ but as a way to bound the plausible size of the bias baseline estimates. For example, if the correction term $(\beta_{long} - \beta_{short})(\frac{R_{max}-R_l}{R_l-R_s})$ is large then that suggests we should have low confidence that $\beta_{short}$ is representative of the $\beta$ we would expect in a better model with more controls. One limitation on this method is that $R_{max}$ cannot be estimated since full regression including $O$ cannot be run without data on $O$. Oster suggests using knowledge about the data generating process to pick a reasonable number or using $R_{max} = 1.3R_l$ which appears to work well in practice based on her case studies.

### 3.3.3 Data Description

Some of the data used in this study were discussed in section 3.3.1 including the key predictor, instrument and state-level controls. The outcomes for physicians and patients come from two surveys that are part of the Health Tracking Surveys (HTS) conducted by the Inter-university Consortium for Political and Social Research (ICPSR). These surveys are successors to the Community Tracking Surveys that began in 1996 but with revisions to create a new sampling frame and focus. The older study sampled intensively from a set of 60 communities thought to be representative of the U.S., with particularly large samples from 12 of the communities. This meant that the older studies are not representative of health care markets in the nation as a whole so the HTS was redesigned to sample from all 50 states in proportion to their population. The public use dataset does not include geographic identifiers, critical for matching physicians and households with state characteristics, but the restricted access data, obtained by application, includes this information.

Table 3.4: Summary statistics for HTS physician variables

| | N | mean | StD | min | max | p25 | p75 |
|---|---|---|---|---|---|---|---|
| income ($s) | 4,621 | 213,861 | 107,515 | 75,000 | 400,000 | 125,000 | 275,000 |
| total hours (patients) | 4,621 | 2,014 | 728.5 | 500 | 4,992 | 1,500 | 2,400 |
| total hours (admin) | 4,621 | 406.0 | 391.5 | 0 | 4,851 | 144 | 500 |
| *Specialty:* | | | | | | | |
| - internal medicine | 4,622 | 0.134 | 0.341 | 0 | 1 | 0 | 0 |
| - general medicine | 4,622 | 0.175 | 0.380 | 0 | 1 | 0 | 0 |
| - pediatrics | 4,622 | 0.09 | 0.287 | 0 | 1 | 0 | 0 |
| - medical specialty | 4,622 | 0.228 | 0.448 | 0 | 1 | 0 | 1 |
| - surgery | 4,622 | 0.19 | 0.392 | 0 | 1 | 0 | 0 |
| - psychiatry | 4,622 | 0.066 | 0.248 | 0 | 1 | 0 | 0 |
| - ob-gyn | 4,622 | 0.067 | 0.250 | 0 | 1 | 0 | 0 |
| experience | 4,621 | 18.22 | 10.46 | 1 | 68 | 10 | 26 |
| IMG | 4,622 | 0.217 | 0.412 | 0 | 1 | 0 | 0 |
| male | 4,622 | 0.737 | 0.440 | 0 | 1 | 0 | 1 |
| Hispanic | 4,622 | 0.051 | 0.220 | 0 | 1 | 0 | 0 |
| Black | 4,622 | 0.039 | 0.193 | 0 | 1 | 0 | 0 |
| White | 4,622 | 0.779 | 0.415 | 0 | 1 | 1 | 1 |
| Asian | 4,622 | 0.146 | 0.354 | 0 | 1 | 0 | 0 |
| Board Certified | 4,622 | 0.897 | 0.305 | 0 | 1 | 1 | 1 |

The physician survey, conducted in 2008 samples only physicians active in patient care, so it excludes those who are retired or focus on management, consulting, or research. A small subset of the sample (1.7%) nevertheless reported working far below half-time with patients so I trimmed observations reporting fewer than 500 hours with patients in the previous year to ensure the sample conforms to the active in patient care criteria. Residents and fellows are also excluded from the sampling frame since residents do not practice independently and fellows are nominally focused primarily on research. ICPSR matched responses to the survey with information from the AMA on practice location, specialty, experience and training background in order to enrich the set of demographic controls available. Most of the questions in the survey involve the physician's perspectives on aspects of their business including how they allocate time, what forms of care they provide, sources of income, and the organization of their practice including ownership structure.

Summary statistics for the HTS physician survey are presented in Table 3.4. The physician income variable in particular is worth discussing. Income on the survey is binned into cat-

egories: less than $100,000, $50,000 increments up to $300,000 and more than $300,000. For regression analysis is important to be able to treat income as a continuous variable so impute values for each category using the mean of the $50,000 increments, $75,000 for the lowest category (just under 12% of physicians), and $400,000 for the top category, to reflect the highly skewed distribution of earnings in almost all professions. The average income with these imputations is $213,861 which seems reasonable, although a bit low, with a large standard deviation ($107,515) reflecting the significant variation across respondents.

The data confirm that physicians tend to work long hours, spending just over 2000 hours with patients and an additional 400 on administrative tasks, although this varies significantly with a few doctors claiming to work almost 100 hours weeks. I do not censor any observations aside from those who appear to be only marginally involved in patient care as noted above although some of these numbers appear unreasonable. About 20% of the sample are IMGs and underrepresented groups are, of course, underrepresented. Black and Hispanic physicians, for instance, make up just under 10% of the sample. About 90% of the physicians are board certified which is higher than the national rate of about 79% at the time (Young et al. 2017) suggesting that it is mildly unrepresentative.

A few outcome variables related to access to care are drawn from the complementary HTS household survey, which focuses on gathering information about insurance coverage and interactions with health care providers, including questions about satisfaction, trust, and problems paying bills. For background the survey also asks about demographics including the standard indicators about race, income, and geography, as well as additional questions about health background, such as presence of chronic conditions. These surveys were conducted a few years after the physician surveys, in 2010, which could be a problem, but there were no major changes to the health care system during this period as the Affordable Care Act's provisions primarily went into effect several years later. The sampling frame sampled households but then asked questions about all members of the family at that address so I use responses from all adults as individuals.

Table 3.5: Summary statistics for HTS household variables

| | N | mean | StD | min | max | p25 | p75 |
|---|---|---|---|---|---|---|---|
| age | 13,858 | 50.55 | 17.25 | 20 | 91 | 37 | 63 |
| male | 13,861 | 0.459 | 0.498 | 0 | 1 | 0 | 1 |
| college grad. | 13,861 | 0.332 | 0.471 | 0 | 1 | 0 | 1 |
| employed | 13,861 | 0.499 | 0.500 | 0 | 1 | 0 | 1 |
| below poverty line | 13,861 | 0.198 | 0.399 | 0 | 1 | 0 | 0 |
| Hispanic | 13,861 | 0.102 | 0.302 | 0 | 1 | 0 | 0 |
| White | 13,861 | 0.721 | 0.448 | 0 | 1 | 0 | 1 |
| Black | 13,861 | 0.113 | 0.317 | 0 | 1 | 0 | 0 |
| M.D. visits | 13,858 | 4.173 | 5.206 | 0 | 30 | 1 | 5 |
| Other visits | 13,858 | 0.606 | 1.747 | 0 | 13 | 0 | 0 |
| has usual source | 13,861 | 0.605 | 0.489 | 0 | 1 | 0 | 1 |
| wait time | 6,254 | 5.069 | 7.198 | 0 | 84 | 1 | 7 |
| *Insurance:* | | | | | | | |
| - Medicare | 13,861 | 0.258 | 0.437 | 0 | 1 | 0 | 1 |
| - Medicaid | 13,861 | 0.0651 | 0.247 | 0 | 1 | 0 | 0 |
| - Private | 13,861 | 0.519 | 0.500 | 0 | 1 | 0 | 1 |
| - Military | 13,861 | 0.0127 | 0.112 | 0 | 1 | 0 | 0 |

One potential problem with these surveys is that lots of information about the household is provided by a single respondent who proxies for other adults. In some cases this is likely innocuous and in instances where the proxy would be unlikely to know the answers to questions there is a self-response module which each adult fills out and mails back. Still, it is possible that on some questions the proxy answers but was inaccurate. This could be a contributing factor for why the results using this sample are so imprecise as seen in section 3.5. Response rates, as with most surveys, were low at 42% and the pattern of non-response is non-random as seen in the summary statistics in Table 3.5. Adults that responded are more likely to live below the poverty line, about 19.8% of sample of adults versus around 15.1% according to the Census Bureau for 2010, and slightly less likely to be employed compared to the general population. According to the survey the respondents average 31% more doctors' visits than comparable surveys of physicians would indicate (CDC 2018) but this appears not to be a result of the response pattern as reweighting with weights constructed by IPSCR to match the CPS yield a similar estimate.

## 3.4 Physician Supply and Spending

This section presents results for how changes in physician supply influence physician outcomes. I start by presenting basic OLS models for physician income that likely suffer from omitted variable bias, which leads to an application of the proportional selection bounding procedure disused earlier and then to instrumental variables estimates. Results from the OLS models are presented in Table 3.6 where the columns add progressively more controls. All of the specifications use a log-log specification for physician income and the physician workforce per 100,000 residents so that the effect of supply increases can be interpreted as an elasticity. In the most basic model where supply is the only predictor the elasticity is significant and negative although not particularly large in absolute value. Adding physician demographics as controls shrinks the coefficient by 40% but it remains significant at conventional levels. The fit of both of these models is poor as indicated by the $R^2$ but adding controls for physician specialty significantly improves the fit, raising the $R^2$ from 0.133 to 0.305. This makes sense as a physician's specialty determines the rates they can charge for evaluations and the procedures for which they specialize in and bill for. The controls for specialty raise the elasticity slightly and it remains significant, but further expanding the set of controls to include state level factors shrinks the coefficient enough to make it significant only at the 10% level.

These results make some sense in the competitive market framework discussed earlier, showing a robust negative association between physician income and local supply. In a supply-demand framework where we think of the physician density as exogenous, conditional on covariates, then the coefficient is the inverse of the price elasticity of demand for physician services. Most of the coefficient estimates are clustered in the range of -0.2, indicating that the implied price elasticity of demand for physician services is around 5. This seems extraordinarily high as we would expect medical care demand to be relative inelastic, intuition confirmed by the RAND health insurance experiment (Manning et al. 1987). This suggests that either the OLS estimates suffer from substantial bias or that there is more to this market than the simple frictionless model would suggest.

Table 3.6: Regression models for ln(physician income)

| **Dependent:** | **ln(income)** | | | |
|---|---|---|---|---|
| ln(physicians per 100k) | -0.259** | -0.155* | -0.192** | -0.165+ |
| | (0.0655) | (0.0581) | (0.0481) | (0.0920) |
| experience | | 0.0231** | 0.0223** | 0.0223** |
| | | (0.00245) | (0.00236) | (0.00239) |
| experience$^2$ | | -0.000671** | -0.000617** | -0.000620** |
| | | (5.29e-05) | (5.37e-05) | (5.42e-05) |
| IMG | | -0.0949** | -0.0289 | -0.0299 |
| | | (0.0189) | (0.0189) | (0.0189) |
| male | | 0.349** | 0.274** | 0.274** |
| | | (0.0199) | (0.0160) | (0.0159) |
| Hispanic | | -0.0754** | -0.0423 | -0.0505+ |
| | | (0.0274) | (0.0269) | (0.0274) |
| Black | | -0.149** | -0.0981+ | -0.0829+ |
| | | (0.0464) | (0.0488) | (0.0472) |
| White | | -0.0148 | 0.00703 | 0.0158 |
| | | (0.0354) | (0.0439) | (0.0427) |
| Asian | | -0.0212 | -0.000315 | 0.00100 |
| | | (0.0332) | (0.0369) | (0.0366) |
| Board Certified | | | 0.148** | 0.147** |
| | | | (0.0182) | (0.0181) |
| Constant | 13.55** | 12.65** | 12.53** | 12.24** |
| | (0.353) | (0.320) | (0.273) | (0.460) |
| *Additional Controls:* | | | | |
| specialty | no | no | yes | yes |
| state controls | no | no | no | yes |
| N | 4,621 | 4,621 | 4,621 | 4,621 |
| $R^2$ | 0.005 | 0.133 | 0.305 | 0.308 |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

Does the bounding exercise described in section 3.3.2 help to clarify the emerging picture? Comparing model 1, with the fewest controls, and model 4, with the largest set of controls, the coefficient shrinks from -0.259 to -0.165, suggesting the omitted factors shrink the treatment effect, but the $R^2$ rises significantly from 0.005 to 0.308 suggesting there is limited scope for further shrinkage unless $R_{max}$ is particularly large. The upper bound assuming $\delta = 1$ and$R_{max} = 1.3R_l = 0.4004$, the parameters suggested in Oster (2019), is -0.136, still substantially below zero. Perhaps a more informative way to look at the data is that with this rate of selection the omitted variables would have to push the $R^2$ up to 0.835 before the $\beta$ of interest would drop to zero. That seems highly implausible on two grounds. First, it is roughly double the variance in physician income that can be explained in the most saturated models. Second, even overfitting the model with dummies for subspecialty codes and states of practice can only explain about 40% of the variance in incomes. In summary it seems unlikely that omitted variable bias could completely wipe out the negative association between physician supply and incomes observed in these models. I now turn to the instrumental variables estimates for the effects of physician supply on outcomes and focus on the IV method for the rest of the chapter. IV estimates for the effect of supply on incomes are presented in Table 3.7. The point estimates are similar to the OLS results for all doctors and for a subsample of only PCPs (columns 3-4) as seen by comparing columns 1 to 2 and 3 to 4. Unfortunately the moderate strength of the instrument combined with the small number of clusters result in much wider standard and preclude statistical significance or even being able to rule out large positive effects of physician supply on incomes. For example, in the PCP only IV model (column 4) the point estimate is modestly negative but with a standard error of 0.256 the 95% confidence interval includes coefficients as high as 0.177 meaning a 10% increase in local physician supply would lead to a 1.77% increase income.

The estimates for the effect of supply on time allocation are clearer and presented in Tables 3.8 and 3.9. The estimates are much more precise with both of the IV estimates achieving statistical significance at the 5% level or beyond. These specifications use the log of time spent as the dependent variable so, like with the models of income, the coefficients can be interpreted

Table 3.7: Instrumental variables estimates for effect of supply on income

| **Dependent:** | **ln(income)** | | | |
|---|---|---|---|---|
| ln(physicians per 100k) | -0.165+ | -0.126 | | |
| | (0.0923) | (0.197) | | |
| ln(primary care per 100k) | | | -0.136 | -0.325 |
| | | | (0.128) | (0.256) |
| experience | 0.0223** | 0.0223** | 0.0182** | 0.0183** |
| | (0.00238) | (0.00236) | (0.00263) | (0.00255) |
| experience$^2$ | -0.000620** | -0.000620** | -0.000479** | -0.000481** |
| | (5.41e-05) | (5.39e-05) | (6.20e-05) | (6.07e-05) |
| IMG | -0.0298 | -0.0299 | -0.00499 | -0.00374 |
| | (0.0199) | (0.0195) | (0.0246) | (0.0251) |
| male | 0.274** | 0.274** | 0.248** | 0.248** |
| | (0.0159) | (0.0158) | (0.0185) | (0.0180) |
| Hispanic | -0.0507+ | -0.0503+ | -0.0220 | -0.0243 |
| | (0.0288) | (0.0287) | (0.0436) | (0.0432) |
| Black | -0.0836* | -0.0837* | -0.0435 | -0.0420 |
| | (0.0342) | (0.0339) | (0.0493) | (0.0486) |
| White | 0.0151 | 0.0152 | 0.0172 | 0.0177 |
| | (0.0198) | (0.0195) | (0.0275) | (0.0274) |
| Board certified | 0.147** | 0.147** | 0.0844** | 0.0879** |
| | (0.0181) | (0.0180) | (0.0270) | (0.0272) |
| Constant | 12.24** | 12.06** | 10.63** | 9.268** |
| | (0.468) | (0.921) | (0.894) | (1.827) |
| *Additional controls:* | | | | |
| specialties | yes | yes | no | no |
| state controls | yes | yes | yes | yes |
| N | 4,621 | 4,621 | 1,847 | 1,847 |
| $R^2$ | 0.308 | 0.308 | 0.118 | 0.118 |
| sample | all | | PCPs only | |
| instrument | n/a | autonomy | n/a | autonomy |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

Table 3.8: Instrument variables estimates for effect of supply on hours worked

| **Dependent:** | **ln(total hours with patients)** | | | |
|---|---|---|---|---|
| ln(physicians per 100k) | -0.216** | -0.412** | | |
| | (0.0494) | (0.117) | | |
| ln(primary care per 100k) | | | -0.167 | -0.543* |
| | | | (0.109) | (0.223) |
| experience | 0.00762** | 0.00760** | 0.0108** | 0.0110** |
| | (0.00183) | (0.00179) | (0.00242) | (0.00233) |
| experience$^2$ | -0.000298** | -0.000297** | -0.000325** | -0.000329** |
| | (4.42e-05) | (4.32e-05) | (5.06e-05) | (4.85e-05) |
| IMG | 0.0528** | 0.0533** | 0.0447 | 0.0471 |
| | (0.0157) | (0.0156) | (0.0300) | (0.0291) |
| male | 0.198** | 0.197** | 0.204** | 0.203** |
| | (0.0127) | (0.0125) | (0.0204) | (0.0203) |
| Hispanic | -0.0261 | -0.0284 | -5.46e-05 | -0.00453 |
| | (0.0244) | (0.0242) | (0.0262) | (0.0267) |
| Black | 0.0602+ | 0.0606+ | 0.0745+ | 0.0775* |
| | (0.0330) | (0.0329) | (0.0403) | (0.0395) |
| White | 0.0204 | 0.0200 | 0.0289 | 0.0300 |
| | (0.0158) | (0.0157) | (0.0192) | (0.0190) |
| Board certified | -0.00457 | -0.00305 | 0.0240 | 0.0310 |
| | (0.0203) | (0.0200) | (0.0264) | (0.0259) |
| Constant | 8.394** | 9.290** | 6.251** | 3.535* |
| | (0.248) | (0.529) | (0.831) | (1.589) |
| *Additional controls:* | | | | |
| specialties | yes | yes | no | no |
| state controls | yes | yes | yes | yes |
| N | 4,621 | 4,621 | 1,847 | 1,847 |
| $R^2$ | 0.112 | 0.111 | 0.107 | 0.103 |
| instrument | n/a | autonomy | n/a | autonomy |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

as elasticities. In the OLS specifications these elasticities are modestly negative, -0.216 for all physicians and -0.167 for PCPs. The former is significant at the 1% level while the later just misses the 10% significant level due to the small sample of PCPs. The instrumental variables estimates are roughly double in magnitude, -0.412 for all physicians and -0.543 for PCPs, and both significant at the 5% level. This is consistent with the suggestive evidence that supply lowers income as it indicates that physicians have fewer patients and as a result can schedule fewer appointments and bill for fewer hours. The drop-off is not one for one, suggesting scope for either inducing demand or, perhaps, scheduling longer visits per patient, but indicate that the induced demand cannot fully offset the dilution of demand for individual physicians.

If physicians spend less time with patients what do they use the extra time for? One possibility, noted earlier and emphasized by Bodemheimer and Smith (2013) is that productivity might be endogenous to the availability of clinical inputs. In areas where more physicians practice then more time can be wasted on administrivia without crowding out time needed evaluate and treat patients. Table 3.9 shows OLS and instrumental variables estimates for the effect of physician supply on time spent on administrative tasks and the estimates are consistent with the endogenous productivity hypothesis. The elasticity for the OLS specification is 0.343 and slightly larger in the IV specification for all physicians, 0.486, indicating that for every 10% more physicians that see patients each physician spends a little less than 5% more time on administrative tasks. It should be noted that this is not necessarily bad for patient health as documentation can be important for quality of care if it facilitates smoother transitions, eliminates duplication of tests, or enables research that improves standards of care and clinical processes, so further work should investigate how marginal hours spent on administrative work impact health outcomes. Another caveat with these results is that they do not hold up in the PCP subsample. The point estimates are similar and slightly larger, consistent with the larger administrative burden of primary care in general, but lacks significance due to a lack of power.

Table 3.9: Does variation in physician supply influence practice styles?

| **Dependent:** | **ln(total hours on administrative)** | | | |
|---|---|---|---|---|
| ln(physicians per 100k) | 0.343* | 0.486* | | |
| | (0.150) | (0.224) | | |
| ln(primary care per 100k) | | | 0.451+ | 0.587 |
| | | | (0.259) | (0.497) |
| experience | 0.0110* | 0.0110* | 0.0164* | 0.0164* |
| | (0.00473) | (0.00468) | (0.00770) | (0.00760) |
| experience$^2$ | -0.000324** | -0.000325** | -0.000357* | -0.000356* |
| | (0.000108) | (0.000107) | (0.000158) | (0.000156) |
| IMG | 0.0355 | 0.0352 | 0.0889+ | 0.0880+ |
| | (0.0332) | (0.0330) | (0.0457) | (0.0453) |
| male | 0.0619 | 0.0625+ | 0.0218 | 0.0222 |
| | (0.0370) | (0.0364) | (0.0581) | (0.0572) |
| Hispanic | 0.0515 | 0.0530 | 0.0517 | 0.0526 |
| | (0.0527) | (0.0522) | (0.0795) | (0.0792) |
| Black | 0.260** | 0.260** | 0.437** | 0.436** |
| | (0.0610) | (0.0605) | (0.128) | (0.126) |
| White | 0.0897* | 0.0900* | 0.119 | 0.119 |
| | (0.0376) | (0.0373) | (0.0815) | (0.0800) |
| Board certified | -0.0414 | -0.0423 | -0.0339 | -0.0365 |
| | (0.0453) | (0.0449) | (0.0655) | (0.0647) |
| Constant | 3.838** | 3.181** | 7.928** | 8.915* |
| | (0.803) | (1.134) | (1.972) | (3.730) |
| *Additional controls:* | | | | |
| specialties | yes | yes | no | no |
| state controls | yes | yes | yes | yes |
| N | 4,294 | 4,294 | 1,708 | 1,708 |
| $R^2$ | 0.019 | 0.019 | 0.040 | 0.040 |
| instrument | n/a | autonomy | n/a | autonomy |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

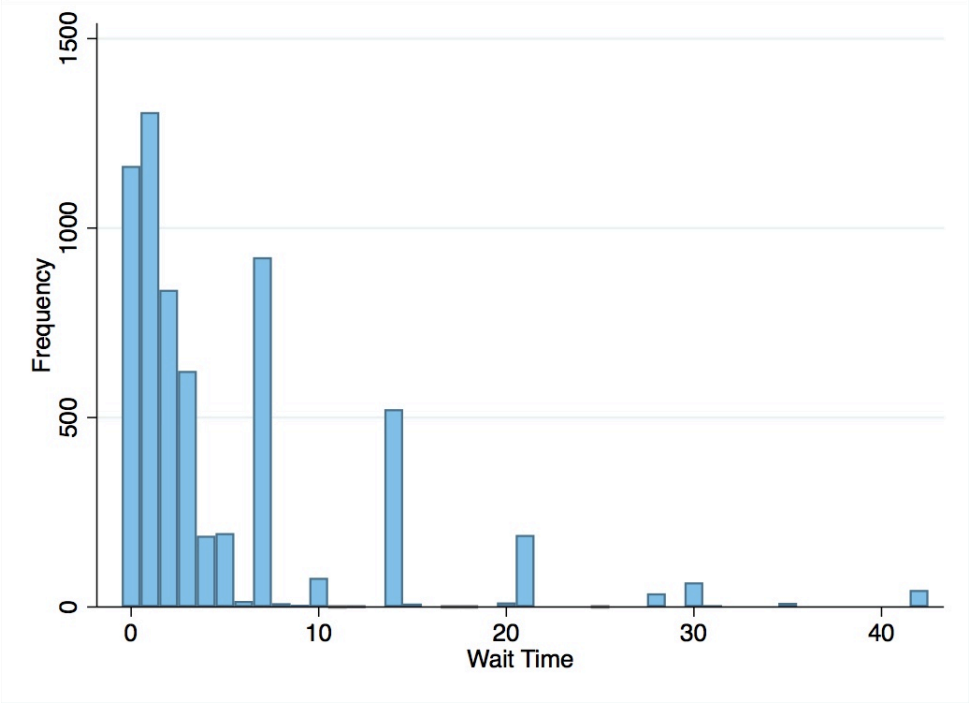


Figure 3.3: Histogram of wait times

## 3.5 Physician Supply and Access to Care

In this section I briefly report results from models of the effect of physician supply on access to care, using outcomes drawn from the HTS patient surveys. It turns out that all of these estimates have wide confidence intervals making them largely uninformative and the models have such poor fit that the proportional selection analysis also provides wide bounds on plausible causal effects. The basic underlying problem is a combination of weak power from the instrument and small sample of clusters combined with substantial noise in recall by respondents on the surveys. Reported wait times, seen in Figure 3.3, illustrate the problem. The distribution is uneven, with significant excess mass on round numbers such as 10 and 30 as well as multiples of seven including, notably 21 and 42, indicating that respondents had poor recall of their wait times and gave approximate, rounded answers. The reported numbers of medical visits suffer from similar problems with excess mass at 2, 5, 10, and 15 visits. This rounding, in addition to genuine recall biases that lead to under and over-reporting visits, creates substantial measurement error.

With that in mind I present the results for the effect of physician supply on access to care in Table 3.10. The first two columns show the OLS and IV estimates for the effect of primary care physician supply on whether a respondent has a primary care physician that manages their care. I restrict this sample to patients with insurance as it is unclear what standard respondents without insurance use for having a usual source of care and it appears that many have in mind that they visit an emergency department.

I would expect that having more PCPs per capita in an area leads to a larger fraction of patients having a PCP and the OLS point estimate shows a large positive association. A ten percent increase in PCP supply is associated with a 1.5 percentage point increase in the probability of having a PCP from a baseline of 65%. The IV point estimate is similar but not significant due to a wide confidence interval.

The HTS survey follows up with respondents that have a PCP (or other usual source of care) and asks about the wait time for the last visit they requested. These estimates are noisy and counterintuitive, with PCP supply associated with significant increases in wait times in the OLS model and associated with even larger increases in wait times in the IV model, although the standard error for that estimate includes large negative effects on wait times. I think it is best to avoid reading too much into these results but, taken at face value, they represent a challenge to the view that training more physicians will improve difficulties with access to care.

My final estimates are presented in Table 3.11 which shows the relationship between the number of physician visits and physician supply as well as the association of physician supply with visits to other medical providers such as nurse practitioners and advanced practice nurses. The OLS results fit with intuition as having more physicians in an area is associated with more visits to doctors and slightly fewer visits to other providers, but the magnitudes are miniscule. The counterpart IV estimates are so noisy as to be uninformative. Future research with billing data on office visits or another more reliable data source may be able to reassess this relationship using my proposed instrument or another natural experiment.

Table 3.10: Instrument variables estimates for effect of supply on access to care

| **Dependent:** | **has a PCP** | | **wait time** | |
|---|---|---|---|---|
| ln(primary care per 100k) | 0.142* | 0.178 | 2.506* | 5.682 |
| | (0.0611) | (0.149) | (1.128) | (3.610) |
| age | 0.00762** | 0.00760** | 0.0108** | 0.0110** |
| | (0.00183) | (0.00179) | (0.00242) | (0.00233) |
| age$^2$ | -0.000298** | -0.000297** | -0.000325** | -0.000329** |
| | (4.42e-05) | (4.32e-05) | (5.06e-05) | (4.85e-05) |
| male | 0.0528** | 0.0533** | 0.0447 | 0.0471 |
| | (0.0157) | (0.0156) | (0.0300) | (0.0291) |
| college grad | 0.198** | 0.197** | 0.204** | 0.203** |
| | (0.0127) | (0.0125) | (0.0204) | (0.0203) |
| employed | -0.0261 | -0.0284 | -5.46e-05 | -0.00453 |
| | (0.0244) | (0.0242) | (0.0262) | (0.0267) |
| below poverty line | 0.0602+ | 0.0606+ | 0.0745+ | 0.0775* |
| | (0.0330) | (0.0329) | (0.0403) | (0.0395) |
| Hispanic | 0.0204 | 0.0200 | 0.0289 | 0.0300 |
| | (0.0158) | (0.0157) | (0.0192) | (0.0190) |
| White | -0.00457 | -0.00305 | 0.0240 | 0.0310 |
| | (0.0203) | (0.0200) | (0.0264) | (0.0259) |
| Black | 8.394** | 9.290** | 6.251** | 3.535* |
| | (0.248) | (0.529) | (0.831) | (1.589) |
| *Insurance:* | | | | |
| - Medicare | 0.00561 | 0.0718 | 0.159 | 0.0195 |
| | (0.0223) | (0.0513) | (1.082) | (1.200) |
| - Private | -0.0787** | -0.0128 | 0.183 | 0.0356 |
| | (0.0254) | (0.0545) | (1.194) | (1.320) |
| - Military | -0.0672 | 0.0647 | 0.767 | 0.685 |
| | (0.0563) | (0.0552) | (1.498) | (1.505) |
| Constant | 1.456** | 1.648 | 20.90** | 43.46+ |
| | (0.446) | (1.055) | (7.347) | (26.41) |
| N | 11,846 | 11,846 | 5,705 | 5,705 |
| $R^2$ | 0.045 | 0.045 | 0.012 | 0.009 |
| instrument | n/a | autonomy | n/a | autonomy |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

Table 3.11: Instrument variables estimates for effect of supply on number of visits

| **Dependent:** | **M.D. visits** | | **other visits** | |
|---|---|---|---|---|
| ln(physicians per 100k) | 0.978+ | -1.259 | -0.118 | 0.265 |
| | (0.510) | (2.139) | (0.156) | (0.505) |
| age | 0.165** | 0.165** | 0.0273** | 0.0275** |
| | (0.0255) | (0.0258) | (0.00964) | (0.00957) |
| age$^2$ | -0.00180** | -0.00179** | -0.000314** | -0.000315** |
| | (0.000314) | (0.000317) | (0.000101) | (9.99e-05) |
| male | -0.938** | -0.933** | -0.117* | -0.118* |
| | (0.132) | (0.133) | (0.0493) | (0.0486) |
| college grad | -0.0460 | 0.00244 | -0.0480 | -0.0563 |
| | (0.148) | (0.169) | (0.0586) | (0.0586) |
| employed | -1.357** | -1.362** | -0.303** | -0.302** |
| | (0.126) | (0.123) | (0.0686) | (0.0678) |
| below poverty line | 0.342 | 0.298 | 0.114 | 0.122 |
| | (0.228) | (0.245) | (0.112) | (0.115) |
| Hispanic | 0.245 | 0.202 | -0.0652 | -0.0578 |
| | (0.485) | (0.494) | (0.105) | (0.0996) |
| White | 0.171 | 0.139 | 0.0556 | 0.0610 |
| | (0.287) | (0.282) | (0.0923) | (0.0928) |
| Black | 0.547 | 0.462 | -0.0143 | 0.000200 |
| | (0.330) | (0.360) | (0.106) | (0.104) |
| *Insurance:* | | | | |
| - Medicare | 1.383* | 0.602 | 0.0488 | -0.340 |
| | (0.564) | (0.895) | (0.221) | (0.396) |
| - Private | -0.715 | -1.470+ | -0.197 | -0.590 |
| | (0.524) | (0.770) | (0.209) | (0.387) |
| - Military | 0.853 | 0.659 | 0.376 | 0.409 |
| | (0.932) | (0.900) | (0.434) | (0.421) |
| Constant | -3.045 | 9.876 | 1.071 | -0.617 |
| | (3.060) | (11.75) | (0.823) | (2.708) |
| N | 11,846 | 11,846 | 11,846 | 11,846 |
| $R^2$ | 0.060 | 0.057 | 0.018 | 0.017 |
| instrument | n/a | autonomy | n/a | autonomy |

Standard errors clustered by state in parentheses
** p<0.01, * p<0.05, + p<0.1

## 3.6 Conclusion

Many commentators are concerned that as the number of insured Americans increases and the population continues to age that the health care system does not have an adequate supply to provide timely care for all patients. In a recent Wall Street Journal survey of 19 health policy experts, seven indicated that the response to the putative physician shortage should be an increase in payments to physicians (i.e. student debt relief) or the number of physicians (i.e. more funding for residencies). Five experts suggested reforming practice styles, emphasizing coordination and technology, to improve productivity, while eight recommended expanding scope of practice for mid-level providers such as nurse practitioners (Potempa et al. 2013). This chapter examined the evidence on whether a greater physician supply, as recommended, is associated with better access to health care for patients and found mixed evidence. A larger physician supply is associated with worse access to care as judged by longer wait times but also associated with a higher probability of having a PCP. I found stronger evidence that a larger physician supply impacts physicians, leading them to spend less time with patients and more time on administrative tasks, and some evidence that greater supply leads to lower incomes. Further study is needed on this topic, both to validate these tentative results and to explore how the reduced form evidence maps onto mechanisms within the market for physician services.

# References


AAMC (2009) 2008 state physician workforce data book, Tech. rep., American Association of Medical Colleges.

AAMC (2018) 2017 state physician workforce data book, Tech. rep., Association of American Medical Colleges.

Adams, J., Bledsoe, G. H. and Armstrong, J. H. (2016) Are pain management questions in patient satisfaction driving the opioid epidemic?, *American Journal of Public Health*, **106**, 985–986.

Altonji, J., Elder, T. and Taber, C. (2005) Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools, *Journal of Political Economy*, **113**, 151–184.

Angirst, J. D. and Pischke, J.-S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Angrist, J. D. (1998) Estimating the labor market impact of voluntary military service using social security data on military applicants, *Econometrica*, **66**, 249–288.

Apfelbaum, J., Chen, C., Mehta, S. and Gan, T. (2003) Postoperative pain experience: results from a national survey suggest postoperative pain continues to be undermanaged, *Anesthesia and Analgesia*, **97**, 534–40.

Arthur, W. B. (1981) The economics of risks to life, *The American Economic Review*, **71**, 54–64.

Bachhuber, M., Saloner, B., Cunningham, C. and Barry, C. (2014) Medical cannabis laws and opioid analgesic overdose mortality in the united states, 1999-2000, *JAMA Internal Medicine*, **174**, 1668–1673.

Baker, D. (2017) History of the joint commission's pain standards: Lessons for today's prescription opioid epidemic, *JAMA*, **317**, 1117–1118.

Bao, Y., Pan, Y., Taylor, A., Radakrishnan, S., Luo, F., Pincus, H. and Schackman, B. (2016) Prescription drug monitoring programs are associated with sustained reductions in opioid prescribing by physicians, *Health Affairs*, **35**, 1045–1051.

Barsky, R. B., Juster, F. T., Kimball, M. S. and Shapiro, M. D. (1997) Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study, *Quarterly Journal of Economics*, **112**, 537–79.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004) How much should we trust differences-in-differences estimates?, *The Quarterly Journal of Economics*, **119**, 249–275.

Birk, J. and Waddell, G. (2017) The mitigating role of prescription drug monitoring programs in the abuse of prescription drugs, Working paper, IZA.

Blanchflower, D. and Oswald, A. (2019) Unhappiness and pain in modern america: A review essay and further evidence on carol graham's unhappiness for all?, *Journal of Economic Literature*, **57**, 385–402.

Blumenthal, D. (2004) New steam from an old cauldron: The physician supply debate, *New England Journal of Medicine*, **350**, 17.

Bodenheimer, T. S. and Smith, M. D. (2013) Primary care: Proposed solutions to the physician shortage without training more physicians, *Health Affairs*, **32**, 1881–1886.

Booth, M. (1996) *Opium: A History*, Simon and Schuster, London.

Buchmueller, T. and Carey, C. (2018) The effect of prescription drug monitoring programs on opioid utilization in medicare, *American Economic Journal: Economic Policy*, **10**, 77–112.

Cameron, C. A. and Miller, D. L. (2015) A practitioner's guide to cluster-robust inference, *Journal of Human Resources*, **50**, 317–372.

Cameron, T. A. (2010) Euthanizing the value of a statistical life, *Review of Enviornmental Economics and Policy*, **4**, 161–178.

Carlson, K. and McDonald, D. (2013) Estimating the prevalence of opioid diversion by 'doctor shoppers' in the united states, *PLOS One*, **8**, e69241.

Cauley, C. E., Anderson, G., Haynes, A. B. *et al.* (????) Predictors of in-hospital postoperative opioid overdose after major elective operations: A nationally representative cohort study, *Annals of Sugery*, **265**, 702–708.

CDC (2018) National ambulatory medical care survey: 2016 national summary tables, mimeo, CDC.

Center for Studying Health Systems Change (2010) Health tracking physician survey, 2008, http://doi.org/10.3886/ICPSR27202.v1.

Center for Studying Health Systems Change (2012) Health tracking household survey, 2010.

Chen, H. *et al.* (2016) Citizen petition to the centers for medicare and medicaid survey requesting changes to the hcaphs survey, mimeo, Physicians for Responsible Opioid Prescribining.

Chou, R. *et al.* (2015) The effectiveness and risks of long-term opioid treatment of chronic pain, Evidence Report/Technology Assessment No. 218, Agency for Healthcare Research and Quality.

Chumpitazi, C. E., Rees, C. A., Camp, E. A. and Bernhardt, M. B. (2017) Opioid prescribing in a pediatric emergency department after the rescheduling of hydrocodone, *Journal of Emergency Medicine*, **52**, 547–553.

Cicero, T. J., Ellis, M. S., Surratt, H. L. and SP., K. (2014) The changing face of heroin use in the united states: A retrospective analysis of the past 50 years, *JAMA Psychiatry*, **71**, 821–826.

Clemens, J. and Gottlieb, J. D. (2017) In the shadow of a giant: Medicare's influence on private physician payments, *Journal of Political Economy*, **125**, 1–39.

Cooper, R. A., Getzen, T. E., McKee, H. J. and Laud, P. (2002) Economic and demographic trends signal an impending physician shortage, *Health Affairs 21*, **21**, 140–154.

Costa, D. L. and Kahn, M. E. (2004) Changes in the value of life, 1940-1980, *Journal of Risk and Uncertainty*, **29**, 159–80.

Currie, J. and Schnell, M. (2018) Addressing the opioid epidemic: Is there a role for physician education?, *American Journal of Health Economics*, **4**, 383–410.

Cutler, D., Skinner, J. S., Stern, A. D. and Wennberg, D. (2019) Physician beliefs and patient preferences: A new look at regional variation in health care spending, *American Economic Journal: Economic Policy*, **11**, 192–221.

Cutler, D. M. and Zeckhauser, R. J. (2000) The anatomy of health insurance, in *Handbook of Health Economics* (Eds.) A. J. Culyer and J. P. Newhouse, Elseiver, pp. 563–643.

Davis, C. S., Lieberman, A., Hernandez-Delgado, H. and Suba, C. (2018) Laws limiting the prescribing or dispensing of opioids for acute pain in the united states: A national systematic legal review, *Drug and Alcohol Dependence*, **194**, 166–172.

Day, W. R. (1919) Webb v. united states, 249 U.S. 96.

DEA (2000) Arcos registrant handbook, mimeo, DEA.

DEA (2018) Arcos retail drug summary reports, mimeo, DEA.

Deyo, R. A., Hallvik, S. E., Hildebran, C., Marion, M., Springer, R., Irvine, J. M., O'Kane, N., Otterloo, J. V., Wright, D. A., Leichtling, G., Millet, L. M., Carson, J., Wakeland, W. and McCarty, D. (2018) Associated of prescription drug monitoring program use with opioid prescribing and health outcomes: A comparison of program users and nonusers, *The Journal of Pain*, **19**, 166–177.

Dineen, K. K. and DuBois, J. M. (2016) Between a rock and a hard place: Can physicians prescription opioids to treat pain adequately while avoiding legal sanction?, *American Journal of Law and Medicine*, **42**, 7–52.

Dionne, G. and Michaud, P.-C. (2002) Statistical analysis of value-of life estimates using hedonic wage method, Working paper02-01, Ecole des HEC Montreal.

Dowell, D., Haegerich, T. and Chou, R. (2016a) Cdc guideline for prescribing opioids for chronic pain–united states, *JAMA*, **315**, 1624–1645.

Dowell, D., Haegerich, T. and Chou, R. (2019) No shortcuts to safer opioid prescribing, *New England Journal of Medicine*, **380**, 2285–2287.

Dowell, D., Zhang, K., Noonan, R. and Hockenberry, J. (2016b) Mandatory provider review and pain clinic laws reduce the amounts of opioids prescribed and overdose death rates, *Health Affairs*, **35**, 1876–1883.

Ehley, B. (2018) *How the opioid crackdown is backfiring*, Politico.

Fass, J. A. and Hardigan, P. C. (2011) Attitudes of florida pharmacists toward implementing a state prescription drug monitoring program for controlled substances, *Journal of Managed Care and Specialty Pharmacy*, **17**, 430–438.

Federation of State Medical Boards and National Board of Medical Examiners (2019) Who is usmle?

Fernandez, W., Hackman, H., McKeown, L., Anderson, T. and Hume, B. (2006) Trends in opioid-related fatal overdoses in massachusetts, 1990-2003, *Journal of Substance Abuse Treatment*, **31**, 151–156.

Finkelstein, A., Hendren, N. and Luttmer, E. (2016) The value of medicaid: Interpreting results from the oregon health insurance experiment, mimeo, Harvard University.

Fishman, S. M., Papazian, J. S., Gonzalez, S., Riches, P. S. and Gilson, A. (2004) Regulating opioid prescribing through prescription monitoring programs: Balancing drug diversion and treatment of pain, *Pain Medicine*, **5**, 309–324.

Fixler, D. and Johnson, D. S. (2014) Accounting for the distribution of income in the us national accounts, in *Measuring Economics Sustainability and Progress* (Eds.) D. W. Jorgenson and J. S. Landefeld, University of Chicago Press, Chicago.

Fortin, N., Lemieux, T. and Firpo, S. (2011) Decomposition methods in economics, in *Handbook of Labor Economics* (Eds.) D. Card and O. Ashenfelter, Elsevier, pp. 1–102.

Frandsen, B. R. and Lefgren, L. J. (2015) Testing rank similarity, mimeo, BYU.

Freyer, F. J. (2016) *Strict opioids law hit chronic pain sufferers hard*, Boston Globe.

Friedman, M. and Kuznets, S. (1954) *Income from Independent Professional Practice*, NBER.

Gage, T. (1982) Drug substitution and triplicate prescription forms, *Texas Dental Journal*, **99**, 16–17.

GAO (2003) Prescription drugs: Oxycontin abuse and diversion and efforts to address the problem, Tech. Rep. GAO-04-110.

Goodman, D. C. and Fischer, E. S. (2008) Physician workforce crisis? wrong diagnosis, wrong prescription, *New England Journal of Medicine*, **358**, 1658–1661.

Green, L. V., Savin, S. and Lu., Y. (2013) Primary care physician shortages could be eliminated through use of teams nonphysicians and electronic communication, *Health Affairs*, **32**, 11–19.

Haffajee, R. L., Jena, A. B. and Weiner, S. G. (2015) Mandatory use of prescription drug monitoring programs, *JAMA*, **313**, 891–892.

Hall, R. E. (1988) Intertemporal substitution in consumption, *Journal of Political Economy*, **96**, 339–357.

Hall, R. E. and Jones, C. I. (2007) The value of life and the rise in health spending, *The Quarterly Journal of Economics*, **122**, 39–72.

Hammitt, J. K., Liu, J.-T. and Liu, J.-L. (2000) Survival is a luxury good, manuscript, Harvard University.

Han, B., Compton, W., Bianco, C., Rane, E., Lee, J. and Jones, C. (2017) Prescription opioid use, misuse, and use disorders in U.s. adults: 2015 national survey on drug use and health, *Annals of Internal Medicine*, **167**, 293–301.

Hargan, E. (2017) Determination that a public health emergency exists, mimeo, U.S. Department of Health and Human Services.

Hausman, N. and Lavetti, K. (2018) Physician concentration and negotiated prices: Evidence from state law changes, mimeo, The Ohio State University.

Herzig, S., Rothberg, M., Cheung, M., Ngo, L. and Marcantonio, E. (2012) Opioids and opioid-related adverse events, *Journal of Hospital Medicine*, **2**, 73–81.

Horowitz, J., McClelland, L., Fordon, R. and Meara, E. (2018) The problem of data quality in analyses of opioid regulation: The case of prescription drug monitoring programs, Working Paper 24947, NBER.

Howard, R. B. F., Gunaseelan, V., Lee, J., Walijee, J., Brummett, C., Campbell, D., Seese, E., Englesbe, M. and Vu, J. (2019) Association of opioid prescribing with opioid consumption after surgery in michigan, *JAMA Surgery*, **154**, e184234.

Huber, M. (2018) *Unintended targets: How pain patients suffer in the opioid crisis*, The Statesman.

Hwang, H.-s., Reed, W. R. and Hubbard, C. (1992) Compensating wage differentials and unobserved productivity, *The Journal of Political Economy*, **100**, 835–858.

Inciardi, J. A., Surratt, H. L., Lugo, Y. and Cicero, T. J. (2007) The diversion of prescription opioid analgesics, *Law enforcement executive forum*, **7**, 127–141.

IOM (2011) 2. pain as a public health challenge, in *Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education and Research*.

Irvine, J. M. S. E. H., Hildebran, C., Marino, M., Beran, T. and Deyo, R. A. (2014) Who uses a prescription drug monitoring program and how? insights from a statewide survey of oregon clinicians, *Journal of Pain*, **15**, 747–755.

Jeffrey, M. M., Henk, H., Hess, E., Meara, E., Ross, J. and Shah, N. (2018) Trends in opioid use in commercially insured and medicare advantage populations in 2007-16: retrospective cohort study, *BMJ*, **362**, k2833.

Jena, A., Goldman, D. and Karaca-Mandic, P. (2016) Hospital prescribing of opioids to medicare beneficiaries, *JAMA*, **176**, 990–997.

Johnson, S. (1996) Disciplinary actions and pain relief: analysis of the pain relief act, *Journal of Medical Ethics*, **24**, 319–327.

Jones, C. M., Paulozzi, L. J. and Mack, K. A. (2014) Sources of prescription opioid pain relievers by frequency of past-year nonmedical use: United states, 2008-2011, *JAMA Internal Medicine*, **174**, 802–803.

Jones, M. R., Viswanath, O., Peck, J., Kaye, A. D., Gill, J. S. and Simopoulos, T. T. (2018) A brief history of the opioid epidemic and strategies for pain medicine, *Pain and therapy*, **7**, 13–21.

Jung, E., Srivastava, K., Abouljoud, M., Keller, R., Okoroha, K. and Davis, J. (2018) Does hospital consumer assessment of healthcare providers and systems survey correlate with traditional metrics of patient satisfaction? the challenge of measuring patient pain control and satisfaction in total joint replacement, *Arthroplasty Today*, **4**, 740–474.

Kanno-Youngs, Z. (2018) *'Drug Dealers in White Coast'*, Wall Street Journal.

Kenan, K., Mack, K. and Paulozzi, L. (2012) Trends in prescriptions for oxycodone and other commonly used opioids in the united states, 2000-2010, *Open medicine*, **6**, e41–e47.

Kenney, G., Saloner, B., Anderson, N., Polsky, D. and Rhodes, K. (2014) Access to care for low-income medicaid and privately insured adults in 2012 in the national health interview survey: A context for findings from a new audit study, mimeo, Urban Institute.

Kertesz, S. G. (2017) Turning the tide or riptide? the changing opioid epidemic, *Substance Abuse*, **38**, 3–8.

Kertesz, S. G. and Gordon, A. J. (2017) *Strict limits on opioid prescribing risk the 'inhumane treatment' of patients*, Stat News.

Kilby, A. (2015) Opioids for the masses? welfare tradeoffs in the regulation of narcotic pain medications, mimeo, MIT.

Kleiner, M. M. (2016) Battling over jobs: Occupational licensing in health care, *American Economic Review*, **106**, 165–70.

Kleiner, M. M. and Krueger, A. B. (2013) Analyzing the extent and influence of occupational licensing on the labor market, *Journal of Labor Economics*, **31**, S173–S202.

Knieser, T. J. and Viscusi, W. K. (2005) Value of a statistical life: Relative position vs. relative age, *American Economic Review*, **95**, 142–146.

Koepke, E., Manning, E., Miller, T., Ganesh, A., Williams, D. and Manning, M. (2018) The rising tide of opioid use and abuse: the role of the anesthesiologist, *Peirooperative Medicine*, **7**, 16.

Lakdawalla, D., Rief, J. and Bauer, D. (2017) Mortality risk, insurance, and the value of life, mimeo, UIUC.

Lee, J. S., Hu, H., Brummett, C., Syrjamaki, J., Dupree, J., Englesbe, M. and Waljee, J. (2017) Postoperative opioid prescribing and the pain scores on hospital consumer assessment of healthcare providers and systems survey, *JAMA*, **317**, 2013–2015.

Lemeneh, T., Lehrman, W. and Conway, P. (2016) Measurement of the patient experience: Clarifying facts, myths, and approaches, *JAMA*, **315**, 2167–2168.

Liang, K.-Y. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

Libby, R. T. (2005) Treating doctors as drug dealers: The dea's war on prescription pain killers, CATO Policy Analysis 545, CATO.

Lubin, G. and Sprung, S. (2012) *The 18 Largest Ransoms Ever Paid*, Business Insider.

Machado, J. A. F. and Mata, J. (2005) Counterfactual decomposition of changes in wage distributions using quantile regression, *Journal of Applied Econometrics*, **20**.

Mallat, J. (2017) The efffect of prescription drug monitoring programs on opioid prescriptions and heroin crime, mimeo, Purdue University.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B. and Leibowitz, A. (1987) Health insurance and the demand for medical care: Evidence from a randomized experiment, *American Economic Review*, **77**, 251–277.

McQuay, H. (1999) Opioids in pain management, *Lancet*, **353**, 229–32.

Meinhofer, A. (2015) The war on drugs: Estimating the effect of prescription drug supply-side interventions, mimeo, Brown University.

Meldrum, M. (2003) A capsule history of pain management, *JAMA*, **290**, 2470–2475.

Moran, M. J. (2016) Guidance on treatment of the economic value of a statistical life (vsl) in u.s. department of transportation analyses - 2016 adjustment, mimeo, U.S. Department of Transportation.

Mrozek, J. R. and Taylor, L. O. (2002) What determines the value of life? a metaanalysis, *Journal of Policy Analysis and Management*, **21**, 253–70.

Murphy, K. M. and Topel, R. H. (2006) The value of health and longevity, *Journal of Political Economy*, **114**, 871–904.

Murray, C. and Lopez, A. (1997) Global mortality, disability, and the contribution of risk factors: Global burden of disease study, *The Lancet*, **349**, 1436–1442.

National Academies of Sciences Engineering and Medicine (2017) *Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use*, National Academies Press.

Nedelman, M. (2017) *Doctors increasingly face charges for patient overdoses*, CNN Health.

Oehler, R. L. D., Elizabeth C., Robinson, D. B. and Brown, L. H. (2016) Has the rescheduling of hydrocodone changed ed prescribing practices?, *The American Journal of Emergency Medicine*, **34**, 2388–2391.

Olafsdottir, T., Asgeirsdottir, T. L. and Norton, E. C. (2019) Valuing pain using the subjective well-being method, Working Paper 23649, NBER.

Olsen, Y., Daumit, G. L. and Ford, D. E. (2006) Opioid prescriptions by u.s. primary care physicians from 1992 to 2001, *The Journal of Pain*, **7**, 225–235.

Olson, M. (1971) *The Logic of Collective Action: Public Goods and the Theory of Groups Cambridge*, Harvard University Press, MA.

Oster, E. (2019) Unobservable selection and coefficient stability: Theory and evidence journal of business and economic statistics, **37**, 187–204.

Owen, F. (2003) *The DEAs War on Pain Doctors*, The Village Voice.

Oyez (2019) In re griffiths.

Pacheco, R. (2002) The use and misuse of oxycontin, mimeo, Harvard Law School.

Pardo, B. (2016) Do more robust prescription drug monitoring programs reduce prescription opioid overdose deaths?, *Addiction*, **112**, 10.

Paulozzi, L. J., Budnitz, D. S. and Xi, Y. (2006) Increasing deaths from opioid analgesics in the united states, *Pharmacoepidemiology and Drug Safety*, **15**, 618–627.

Paulozzi, L. J. and Ryan, G. W. (2006) Opioid analgesics and rates of fatal drug poisoning in the united states, *American Journal of Preventive Medicine*, **31**, 506–511.

PDMP TTAC (2018) History of prescription drug monitoring programs, mimeo, Brandeis University.

Peterson, B., Pandya, S. and Leblang, D. (2014) Doctors with borders: occupational licensing as an implicit barrier to high skill migration, *Public Choice*, **160**, 45–63.

Pizzi, L., Toner, R., Foley, K., Thomson, E., Chow, W., Kim, M., Couto, J., Royo, M. and Viscusi, E. (2012) Relationship between potential opioid-related adverse effects and hospital length of stay in patients receiving opioids after orthopedic surgery, *Pharmacotherapy*, **32**, 502–14.

Pletcher, M., Kertesz, S., Kohn, M. and Gonzales, R. (2008) Trends in opioid prescribining by race/ethnicity for patients seeking care in us emergency departments, *JAMA*, **99**, 70–78.

Polat, S. (2013) Wage compensation for risk: The case of turkey, mimeo, IKTISAT.

Portenoy, R. K. and KM., F. (1986) Chronic use of opioid analgesics in non-malignant pain: report of 38 cases, *Pain*, **25**, 171–186.

Porter, J. and Jick, H. (1980) Addiction rare in patients treated with narcotics, *New England Journal of Medicine*, **302**, 123.

Potempa, K. *et al.* (2013) *The Experts: What Should be Done to Fix the Predicted U.S. Doctor Shortage?*, The Wall Street Journal.

Powell, D., Pacula, R. and Jacobson, M. (2018) Do medical marijuana laws reduce addictions and deaths related to painkillers?, *Journal of Health Economics*, **58**, 29–42.

Powell, L. F. (1975) United states v. moore, 423 u.s. 122.

Rankin, B. (2018) *Prosecutors notify 30 doctors about excessive opioid prescriptions*, Atlanta Journal-Constitution.

Rhodes, K. V., Kenney, G. M., Friedman, A. B. *et al.* (2014) Primary care access for new patients on the eve of health care reform, *JAMA Internal Medicine*, **174**, 861–869.

Roodman, D., MacKinnon, J., Nielsen, M. and Webb, M. (2018) Fast and wild: Bootstrap inference in stata using boottest, mimeo, Queens University.

Rosen, S. (1988) The value of changes in life expectancy, *Journal of Risk and Uncertainty*, **1**, 285–304.

Rosenblum, A., Marsch, L. A., Joseph, H. and Portenoy, R. K. (2008) Opioids and the treatment of chronic pain: Controversies, current status, and future directions, *Experimental and Clinical Psychopharmacology*, **16**, 405–416.

Rubin, R. (2019) Limits on opioid prescribing leave patients with chronic pain vulnerable, *JAMA*, **321**, 2059–2062.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. and Sobek, M. (2019) Ipums usa: Version 9.0 [dataset].

Rutkow, L., Turner, L., Llucas, E., Hwang, C. and Alexander, G. C. (2015) Most primary care physicians are aware of prescription drug monitoring programs, but many find the data difficult to access, *Health Affairs*, **34**, 484–492.

Sacco, L., Duff, J. and Sarta, A. K. (2018) Prescription drug monitoring programs, Congressional Research Report 7-75700, Congressional Research Service.

Sansone, R. and Sansone, L. (2012) Doctor shopping: A phenomenon of many themes, *Innovations in Clinical Neuroscience*, **9**, 42–46.

Schnell, M. (2017) Physician behavior in the presence of a secondary market: The case of prescription opioids, mimeo, Princeton University.

Schnell, M. and Currie, J. (2018) Addressing the opioid epidemic: is there a role for physician education?, *American Journal of Health Economics*, **4**, 383–410.

Scholl, L., Seth, P., Kariisa, M., Wilson, N. and Baldwin, G. (2019) Drug and opioid-involved overdose deaths — united states, 2013–2017, *MMWR Morb Mortal Wkly Rep*, **67**, 1419–1427.

Shepard, D. and Zeckhauser, R. (1984) Survival versus consumption, *Management Science*, **30**, 423–439.

Sherry, T., Sabety, A. and Maesta, N. (2018) Documented pain diagnoses in adults prescribed opioids: Results from the national ambulatory medical care survey, 2006–2015, *Annals of In-*

*ternal Medicine*, **169**, 892–894.

Simeone, R. (2017) Doctor shopping behavior and the diversion of prescription opioids, *Substance Abuse: Research and Treatment*, **11**.

Simoni-Wastilla, L. and Tompkins, C. (2001) Balancing diversion control and medical necessity: the case of prescription drugs with abuse potential, *Substance Use and Misuse*, **36**, 1275–96.

Smith, R. (2018) *Hundreds of people in Mexico are kidnapped every year. And the problem's getting worse*, Vox.

Stigler, G. J. (1971) The theory of economic regulation, *Bell Journal of Economics*, **2**, 3–21.

Stock, J. and Watson, M. (2007) *Introduction to econometrics*, Pearson/Addison Wesley, Boston.

Stock, J. H., Wright, J. J. and Yogo, M. (2002) A survey of weak instruments and weak identification in generalized methods of moments, *Journal of Business and Economic Statistics*, **20**, 518–529.

Stone, W. (2018) *Patients with chronic pain feel caught in an opioid-prescribing debate*, Health Leaders.

Sunstein, C. (2013) The value of a statistical life: some clarifications and puzzles, *Journal of Benefit-Cost Analysis*, **4**, 237–261.

Taylor, P. (2018) *Are doctors cutting back on opioids too much and too quickly?*, healthydebate.com.

Thaler, R. and Rosen, S. (1976) The value of saving a life: evidence from the labor market, in *Household Production and Consumption* (Ed.) N. E. Terlecky, NBER.

Tompkins, D., Hobelmann, J. and Compton, P. (2017) Providing chronic pain management in the 'fifth vital sign' era: historical and treatment perspectives on a modern-day medical dilemma, *Drug and Alcohol Dependence*, **173**, S11–S21.

Vines, S. (1998) *Luck runs out for Big Spender gangster*, The Independent.

Viscusi, W. K. (2013) The value of individual and societal risks to life and health, in *Handbook of Economics of Risk and Uncertainty* (Eds.) M. Machina and W. Viscusi, Elseiver, pp. 385–452.

Viscusi, W. K. (2018) Best estimate selection bias in the value of a statistical life, *Journal of Benefit Cost Analysis*, **9**, 205–246.

Viscusi, W. K. and Aldy, J. E. (2003) The value of a statistical life: a critical review of market estimates throughout the world, *Journal of Risk and Uncertainty*, **27**, 5–76.

Weiner, S. G., Griff, C. A., Mitchell, P. M., Langlois, B. K., Friedman, F. D., Moore, R. L., Lin, S. C., Nelson, K. P. and Feldman, J. A. (2013) Clinician impression versus prescription drug monitoring criteria in the assessment of drug-seeking behavior in the emergency department, *Annals of Emergency Medicine*, **62**, 281–289.

Weingast, B. and Moran, M. (1983) Bureaucratic discretion or congressional control? regulatory policymaking by the federal trade commission, *Journal of Political Economy*, **91**, 765–800.

Wennberg, J., Goodman, D., Nease, R. and Keller, R. (1993) Finding equilibrium in u.s. physician supply, *Health Affairs*, **12**, 89–103.

Wennberg, J. E. and the Dartmouth Atlas of Health Care Working Group (1998) *The Dartmouth Atlas of Health Care*, The Center for the Evaluative Clinical Sciences, Dartmouth Medical School.

World Health Organization (1986) *Cancer Pain Relief*, The World Health Organization, Geneva, Switzerland.

Young, A., Chaudhry, H., Pei, X., Arnhart, K., Dugan, M. and Snyder, G. (2017) A census of actively licensed physicians in the united states, 2016, *Journal of Medical Regulation*, **103**, 7–21.