The Economics of Fraud and Corruption

by
Jetson Leder-Luis
B.S., California Institute of Technology (2014)

Submitted to the Department of Economics in Partial
Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2020

Author: ....
Department of Economics
May 15, 2020

Certified by: ....
James Poterba
Mitsui Professor of Economics
Thesis Supervisor

Certified by: ....
Benjamin Olken
Professor of Economics
Thesis Supervisor

Accepted by: ....
Amy Finkelstein
John & Jennie S. MacDonald Professor of Economics
Chairman, Departmental Committee on Graduate Studies

The Economics of Fraud and Corruption

by
Jetson Leder-Luis

Submitted to the Department of Economics
on May 15, 2020 in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Economics

ABSTRACT

Fraud and corruption are serious issues which undermine the provision of public goods. This thesis consists of three papers which analyze the economics of fraud and the mechanisms by which it can be detected and averted. An introductory chapter presents an overview of the economic ideas surrounding these topics. In the first paper, I analyze a US federal law that incentivizes whistleblowers to litigate against fraud and misreporting committed against the Medicare program. I provide a theoretical framework for understanding the economic tradeoffs associated with privatized whistleblowing enforcement and then empirically analyze the deterrence effects of whistleblower lawsuits. In the second paper, conducted as joint research, we consider the incentives for misreported enrollment statistics in Israeli public school data and the way in which data manipulation undermines economic estimates of the returns to smaller class sizes. We provide evidence of enrollment manipulation and show that smaller class sizes have no effect on student achievement, overturning earlier literature. In the third paper, we develop a mechanism for detecting misreported financial data and apply it to reports from a World Bank project. Our results are consistent with strategic and profitable falsification of data, and our method matches the results of an audit conducted independently by the World Bank on the same project.

Thesis Supervisor: James Poterba
Title: Professor of Economics

Thesis Supervisor: Benjamin Olken
Title: Professor of Economics

Acknowledgements

I would like to thank my advisors, Benjamin Olken and James Poterba, for their immense support through my doctoral program. I am incredibly lucky to have had the opportunity to study under these two scholars, without whom this work would not have been possible. James Poterba has gone above and beyond his role as advisor and has served as a mentor for me through many life challenges, for which I am deeply grateful. Joshua Angrist, Jonathan Gruber, Alberto Abadie, Daron Acemoglu, and Jean Ensminger have also been very generous with their time throughout my graduate school career.

I have been very fortunate to have great peers who have enriched my graduate school career. I thank Pablo Azar, Cory Smith, Ryan Hill, and Evan Soltas, among many others.

My family and friends have been the foundation of my life. I thank Caroline Howard, Jayda Vera Siegler, Charles Siegler, Stamps Howard, Brandon Archambault, Jason Kuster, Sara Wildavsky, John Abel, and my parents Julio Leder-Luis and Judith Herzfeld for their unending love and support.

The financial and research support of MIT, The Ford Foundation, The National Academies of Science, and the National Bureau of Economic Research made it possible for me to complete this work.

Chapter 1

Introduction

The provision of public goods relies on a government's ability to correctly expend public funds on its intended targets. The diversion of funds due to corruption or fraud undermines the link between the government's planned spending and its execution. Despite the importance of these issues, fraud in governmental spending has been poorly measured; in fact fraud faces fundamental measurement challenges as those committing bad acts seek to hide their behavior. Proxy measures indicate that the diversion of funds is a serious issue: for example, US government "improper payments" (a catchall term to describe payments made without confidence of their correctness), totaled to $141 billion in 2017 [1]. The potentially large scope of this issue merits economic analysis of the economics of fraud and corruption as well as an exploration of the methods that may be used to detect it.

Besides undermining federal expenditure, fraud can have broader downstream economic consequences. To understand the economic consequences of fraud, we have to consider where the money comes from, how it is used once stolen, and the incentives that fraud provides to bureaucrats running governments. In the best case scenario, fraud exists without distortionary behavior by bureaucrats, and acts as a transfers from the government to private individuals. In this case, the net economic loss is the difference between the social value of

[1] Government Accountability Office Report Highlights, GAO-18-377, May 2018. Estimates such as this one do not include fraud that avoids detection, and do include payment mistakes that may ultimately be corrected

the intended use of the funds and the private value of the illicit recipient. This efficiency loss therefore depends on the marginal value of the funds in federal expenditure. When fraud affects programs for needy recipients, these costs can be high; conversely, if the money is stolen from wasteful, low-value programs, fraud could be of little economic cost. However, the opportunity to commit fraud can also encourage rent-seeking behavior on the part of bureaucrats and firms working for and with the government. In more severe examples of regulatory capture, a profit motive can encourage government policymakers to direct funds towards areas where they are best able to seek rents through abuse of their position.

Fraud and corruption are exacerbated by the complicated system dynamics surrounding federal payments. The process of government expenditure involves many different actors, from bureaucrats responsible for the administration of the funds to private firms responsible for the execution of certain projects and individual citizens that are intended to benefit from the spending. Each of these actors faces different incentives and has different information about the realization of the intended expenditure. From a principal-agent purview, the principal in the case of government expenditure is the public, and the agents are those individuals entrusted by the public to execute governance. The differences in these objectives, and the inability to efficiently monitor and contract on the behavior of bureaucrats, provides opportunities for bad behavior that undermines governmental provision of goods and services.

One of the mechanisms by which fraud arises in federal expenditures is through misreporting. Governmental processes rely on data that flows upward, from governmental employees and private firms that execute spending toward those who hold the purse strings. Information asymmetries between different actors in the governmental process therefore provide opportunities for fraud: by misreporting data that determine funding, unscrupulous actors can divert federal funds toward seemingly appropriate uses. Yet this also presents opportunities to correct this behavior. By understanding the mechanisms by which data are misreported, and the statistical properties of misreported data, governments can potentially eliminate some portion of fraud from the source.

In this dissertation, I present three papers that give examples of misreporting in the context of governmental expenditures and provide a set of mechanisms by which misreporting can be detected or averted. In my first paper, “Whistleblowers, The False Claims Act, and the Behavior of Medicare Providers,” I examine the economics of a whistleblower law that seeks to combat misreporting and fraud to Medicare, the US federal old-age health insurance program. Public health insurance is plagued with profiteering by private healthcare providers who face incentives to profitably misreport their claims for payment to the federal government-as-insurer. The False Claims Act provides a mechanism for whistleblowers, individuals with private information about illegal behavior by healthcare providers, to sue those providers in civil court on behalf of the government and receive a share of the recovery rewarded to the government. This paper presents a theoretical framework for the economics of this private enforcement mechanism and also reports key measurements of its efficacy. The major empirical findings are two: first, that lawsuits by whistleblowers have a large and valuable deterrence component through changes to provider behavior; and second, that these changes to provider behavior appear to benefit patients. This lends support to the idea that whistleblowing, and privatized enforcement more generally, may be an effective way to combat misreporting.

In the second paper, “Maimonides Rule Redux”, coauthored with Joshua Angrist, Victor Lavy, and Adi Shany, we examine misreported public school enrollment statistics in Israel. Per the funding rules of the Israeli government, classes are capped at 40 students, and so schools with 41 or more students in a grade are split into two smaller classes. This rule had been used by earlier economic studies in a Regression Discontinuity design to study the effect of smaller classes on student achievement. This new paper uncovers the fact that this funding rule provided incentives for misreporting, particularly among schools near the cutoff, for whom having an extra few students would generate an additional class (with funding for an additional teacher). A plot of the histogram of the grade-level enrollments shows excess mass just after the cutoff, and qualitative documents from the Israeli education system describe efforts by principals to generate extra classes by manipulating enrollment. Unlike the other examples in this dissertation, this misreporting appears to be motivated by the desire of some principals to provide better educational quality to their students, an altruistic rather than rent-seeking motive. This paper documents the enrollment manipulation, providing statistical analysis through the McCrary test for manipulation of the running variable. We also revisit the economic questions about class size and student achievement explored in earlier literature. Our results contradict earlier findings and show that smaller classes do not produce any achievement gains for students. This highlights the way in which fraud can also undermine economic analyses that rely on governmental data.

The third paper of this dissertation is “Measuring Strategic Data Manipulation: Evidence from a World Bank Project,” co-authored with Jean Ensminger. We develop a method to detect strategic misreporting among financial expenditure data. This method relies on the fact that individuals are bad random number generators with limited numeracy and therefore encode patterns when falsifying data. We apply this method to a World Bank development aid project and demonstrate high levels of strategic and profitable misreporting. Our metric matches an audit independently conducted by the World Bank, demonstrating that this method could be a viable way to provide scalable, real-time monitoring of expenditure data in a variety of governmental and private contexts.

Chapter 2

Whistleblowers, The False Claims Act, and the Behavior of Healthcare Providers

Abstract

This paper studies the effects of litigation by whistleblowers against healthcare providers for misreporting claims for payment to the Medicare program. Under the U.S. False Claims Act, whistleblowers bring lawsuits on behalf of the government in exchange for a share of recovered payments. I combine a new dataset on whistleblower cases from the Department of Justice with the universe of Medicare Fee-for-Service claims from 1999-2016. First, I measure the deterrence effects of successful whistleblowing lawsuits using a synthetic control design. I find that whistleblower settlements totaling to $1.9 billion in recovery generated future cost savings of more than $18 billion over 5 years. Next, I examine how whistleblowing impacts care decisions by providers. Using a case study of spine procedures for osteoporotic patients, I find that after a whistleblower settlement, care shifted from inpatient to less-expensive outpatient treatment and towards patients with the greatest expected benefit.

2.1 Introduction

Misreporting, overbilling, and fraud hinder governmental provision of public goods and services. The American federal health insurance programs are particularly susceptible to misreporting by medical providers, driven by the information asymmetry between the federal insurer and the private healthcare providers these insurers reimburse. This information asymmetry, compounded by the sheer size of the federal healthcare programs, makes monitoring both expensive and challenging. With the United States government spending more than a trillion dollars per year on health care, even small shares of financial impropriety can be very expensive, motivating the concern that healthcare has become a source of illicit profit for those providers willing to misreport.

Whistleblowing is a potentially valuable method to curb improper behavior by providing incentives for individuals to come forward with private information about this behavior. The False Claims Act (FCA) is a federal law that since 1986 has allowed whistleblowers to recover over-billed money for the government and receive a share of the recoveries. This creates strong financial incentives for whistleblowing. Uniquely, FCA whistleblowers conduct litigation on behalf of the federal government in federal civil court, without the need for government approval nor necessarily receiving government support in conducting the case. As such, the FCA combines a reward for private information with financial incentives for private enforcement. The FCA, which is frequently used to combat fraud in healthcare, has become a powerful tool for uncovering impropriety, generating thousands of whistleblowing cases and tens of billions of dollars of recovered money over the past 30 years. In fiscal year 2018 alone, whistleblowers recovered almost $2 billion from FCA healthcare lawsuits, for which whistleblowers were awarded $266 million (Civil Division, U.S. Department of Justice, 2018).

There has been substantial disagreement in the public sphere over the value of the FCA.

Proponents of whistleblowing point to the volume of settled cases and the billions of dollars recovered as evidence of the effectivity of the Act. Attorney General Eric Holder said in a 2012 press release: “In the last quarter century, the False Claims Act’s success has been unparalleled with more than $30 billion dollars recovered...and $8.8 billion since January 2009” (United States Department of Justice, 2012). Yet detractors suggest that profit-seeking whistleblowers use civil litigation to force settlements from providers, regardless of the validity of their allegations. In an *amicus* brief to the Supreme Court, the U.S. Chamber of Commerce, a pro-business organization, wrote: “[whistleblowers] can extract settlements from defendants averse to high discovery costs, the risk of large losses, and...reputational harms” (Chamber of Commerce, 2015).

The private enforcement of law has many potential benefits and costs. By compensating whistleblowers for conducting the enforcement, the False Claims Act has been very effective at producing lawsuits and recovering funds for the federal government. In addition, whistleblowing can have deterrence effects, both by changing spending on the types of conduct litigated against by whistleblowers, which I call direct deterrence, and also by preventing potential fraud from being committed, which I call ex-ante deterrence. In contrast to these benefits, private enforcement faces the risk of unnecessary and costly over-enforcement. Lawsuits have both public and private costs, including to the court system, to both plaintiff’s attorneys and defense attorneys, and to medical providers who face increased litigation risk and may change their care decisions.

In this paper, I examine the economics of whistleblowing under the False Claims Act and conduct empirical analyses of two effects of whistleblowing. First, I measure the direct deterrence effects of whistleblower cases, and second, I study the effects of whistleblowing on provider care decisions. Data on whistleblowing cases were obtained through a Freedom of Information Act request filed with the Department of Justice, as well as hundreds of press releases from the Department of Justice archives and federal court documents. These are paired with large samples of Medicare claims data from 1999-2016, which detail medical procedures conducted, federal expenditures, and some patient health outcomes.

To measure direct deterrence effects, I conduct a series of case studies of successful whistleblowing lawsuits and analyze their effects on Medicare claims and spending. First, I create categories of whistleblower lawsuits that have similar allegations of misreporting or fraud. For example, one category is the profitable manipulation of the Medicare outlier payment system; this category contains 11 Department of Justice press releases. I analyze the 4 highest-recovery categories as case studies, looking at post-lawsuit claims and spending. These categories combined contain 29 press releases detailing $1.9 billion in total settlements. I apply a synthetic control methodology to these case studies to estimate counterfactuals in the absence of whistleblowing. The standard synthetic control methodology is augmented by an additional parameter, a time-shift for each control, which I estimate. This method allows the comparison between similar trends in spending that occur at different points in time. Using the time-shifted controls, I compare spending on treated types of medical care to a synthetic control group constructed of similar, untreated types of care. The difference between spending on treated care and the synthetic control group provides the measure of direct deterrence in the post-whistleblowing periods.

My results point to a high direct deterrence value of whistleblower cases. The total direct deterrence effect of the 4 case studies exceeds $18 billion in the first 5 years after the cases were filed. On average, direct deterrence is about 6.7 times the case's settlement value, with wide variation in this ratio. For comparison, total whistleblower compensation for all healthcare related cases in my data from 1986-2012 totals to $4.29 billion, indicating that the deterrence effects of just the few largest cases outweigh the collective costs of paying whistleblowers. Importantly, these direct deterrent effects do not count the ex-ante deterrence value of fraud that is never committed by providers who face increased litigation risk, and therefore these amounts constitute a lower bound of the total deterrence effects of these cases.

Next, I examine the effects of whistleblowing on provider care decisions. Whistleblowing has the potential to change provider care decisions by changing the compliance requirements, litigation risk, and profitability of care that doctors conduct. I conduct a case study on kyphoplasty, a spinal procedure for patients with osteoporosis. A set of whistleblower lawsuits alleged that many hospitals fraudulently admitted patients for inpatient kyphoplasty rather than perform this procedure outpatient. Kyphoplasty is linked to decreased mortality in the medical literature, and there was a large effect of whistleblowing on treatment. This motivates a case study of kyphoplasty as it is uniquely positioned for understanding changes in care decisions. I model the effects of kyphoplasty on Medicare patient death among two cohorts of roughly 8 million patients each, from before and after the lawsuits. My model interacts patient covariates and medical history with treatment, which produces estimated treatment effects of kyphoplasty for each patient. These effects provide a scale by which to measure how beneficial it is for a patient to receive treatment.

This modeling exercise finds that, in the case of kyphoplasty, whistleblowing had positive overall effects on patient care. Following whistleblowing, there was better targeting of the procedures to those patients for whom greater benefit is predicted. Patients for whom it is expected that kyphoplasty increased mortality were 7% less likely to be treated, while patients with reduced mortality if treated were 7% more likely to be treated. This targeting change was concurrent with a substitution from inpatient kyphoplasty to less expensive outpatient kyphoplasty and vertebroplasty, a close substitute. This indicates that whistleblowing can have positive effects on care delivery by changing the incentives in the care decision process by providers, even while reducing spending.

This paper relates to a variety of literature on the False Claims Act, private enforcement, whistleblowing, and Medicare fraud. Despite the volume of lawsuits and funds recovered under the FCA, there has been little empirical economic analysis of this law. Engstrom

(2012; 2013) presents descriptive statistics on FCA case length, settlement values, defendant characteristics, and more, although he does not measure any of the effects of whistleblowing. Depoorter and De Mot (2006) present a theoretical analysis of FCA whistleblowing and government intervention. In the accounting literature, Heese and Cavazos (2019) show that firms which settle under the FCA receive reduced procurement contracts from the federal government, and Heese (2018) shows that hospitals prosecuted under the FCA are less likely to participate in broad measures of overbilling. Related to whistleblowing more generally, Dyck, Morse, and Zingales (2010) analyze the types of whistleblowers who come forward under the Securities and Exchange Commission whistleblower program, which provides compensation but does not allow direct private enforcement. This paper also relates to the legal literature on private enforcement, of which Polinsky (1980) is a seminal work; and on deterrence, for which Becker and Stigler (1974) began a wide literature. In the medical fraud literature, Silverman and Skinner (2004) and Dafny (2005) describe the financial incentives for upcoding inpatient DRGs among for-profit hospitals, which relates to the analyses in Section 2.4 that describe overbilling for inpatient care. Nicholas et al. (2019) present observational evidence that patients treated by Medicare providers subsequently excluded from the program for fraud or abuse face increased mortality risk, which relates to the analysis in Section 2.5 on provider care and patient mortality. This paper fills a gap in the existing literature by providing empirical evidence of the effects of whistleblowing and private enforcement in the context of Medicare fraud.

This paper is organized as follows. Section 2.2 describes the institutional details of the False Claims Act and gives a deeper treatment of the economics of private enforcement. Section 2.3 describes the data and provides stylized facts about FCA lawsuits and recoveries. Section 2.4 presents the time-shifted synthetic control method, describes a series of case studies of successful whistleblower lawsuits and measures direct deterrence. Section 2.5 presents the effects of the kyphoplasty lawsuits on patient care, and Section 2.6 concludes.

2.2 The False Claims Act: Background and Economic Framework

2.2.1 Background and Institutional Details

Medical care has a fundamental information asymmetry between providers, insurers, and patients (Arrow, 1963), which creates opportunities for misreporting. A patient who receives medical care rarely observes the billing process, which is conducted by the provider. Conversely, insurers have limited means of directly observing care, relying on provider's claims for payment. This asymmetry provides opportunities for misreporting by providers, whose billing practices tie directly to their profits. It is difficult to uncover this misreporting using top-down enforcement, as insurers generally lack other sources of information besides the provider's claim and supporting documentation. As such, private information from providers or their staff is useful for uncovering any misreporting or fraud.

When the insurer is the federal government, as is the case with Medicare and Medicaid, these problems are exacerbated. Medicare and Medicaid are massive programs, spending respectively around \$700 and \$400 billion per year (Congressional Budget Office, 2019), creating bureaucratic issues due to the sheer volume of claims. Indeed, the Government Accountability Office (GAO) estimates that around 8% of Fee-for-Service Medicare expenditures in 2017 were "improper," i.e. lack necessary documentation to ensure the correct amount was paid to the right person for a valid claim (United States Government Accountability Office, 2019). Though most improper payments are not fraudulent, this underscores the opportunism that may arise from expensive and overwhelmed federal programs. Even small shares of fraud in Medicare spending can amount to tens of billions of dollars per year.

With these issues in mind, in 1986 Congress amended[1] the False Claims Act to enable

[1] The FCA was amended in 1986, but originally existed during the Civil War to combat fraud. It was ineffective and out of use in the 20th century before the 1986 amendments.

whistleblowers to directly conduct lawsuits against those who over-bill the government (United States Department of Justice, 2012). Though not restricted to healthcare, the False Claims Act has largely been used against healthcare-related fraud, overbilling, and misreporting. Under the False Claims Act, individuals who uncover misreporting against the US government, themselves often healthcare workers (e.g. a disgruntled nurse), hire their own attorneys and sue those committing the impropriety in federal civil court. The whistleblower sues *qui tam*, i.e. on behalf of the US government. These civil court cases have 3 parties: the whistleblower, the defendant, and the US government. In some cases, the Department of Justice intervenes in what it believes to be a lucrative lawsuit by assigning its own attorneys and conducting the investigation and litigation. In other cases, the whistleblower does not receive federal support, and either pursues the case alone or drops it. All cases are filed under seal, meaning the defendant is not immediately notified of the filing, giving the government an opportunity to investigate and elect to intervene before the defendant is made aware. The Department of Justice must also approve any settlements between the whistleblower and the defendant, regardless of their intervention status.

False Claims Act lawsuits can be high stakes for all parties involved. These cases are conducted in civil court, and the burden of proof is the preponderance of the evidence, i.e. "more likely than not." Because litigation is expensive, few cases go to trial; unsuccessful cases are often voluntarily dismissed by the whistleblower, and clear-cut cases are settled. In successful cases, the federal government can recover up to 3 times the amount of the proven false claims from the defendants, plus potentially large criminal fines. Upon settlement, the whistleblower is entitled to 15-25% of the recovery amount if the government intervened, and 25-30% if the government did not intervene. Whistleblowers regularly earn 6-figure payouts and above from these cases, of which their attorneys, working on contingency, take around 30%. Defendants are also often hit with criminal fines and can furthermore be sued for legal fees by successful whistleblowers. Enforcement is compounded by the use of Corporate

Integrity Agreements, where settling providers agree to additional federal oversight, or by the use of exclusion of the provider from the Medicare and Medicaid programs.

2.2.2 The Economics of Private Enforcement

The private enforcement of law faces a tradeoff between the benefits of privatization, which include incentives for enforcement and deterrent effects, and the downsides of privatization, which include spurious litigation and distortionary effects on medical provision.

The False Claims Act creates a bounty program for the private enforcement of law. The opportunity for a large payout creates incentives for a whistleblower to come forward with their private information about fraud or misconduct, which can alleviate personal and professional costs arising from whistleblowing on one's employer. Furthermore, the ability for the whistleblower to conduct the case in lieu of the government creates a profit motive for rooting out impropriety that may be otherwise lacking in the federally-administered programs. This profit motive is in contrast to the usual incentives of federal bureaucrats, and thus can alleviate principal-agent problems within the government that can cause inefficient investment in monitoring and enforcement. Prosecution conducted by the government has capacity constraints due to the limited resources of the Department of Justice, while privatized enforcement creates a market for whistleblowing information and generates substantially more litigation than the federal government conducts alone. [2]

Whistleblower cases also have the potential for valuable direct deterrence effects. Following a lawsuit, both the defendants and other providers of the same care face incentives to change their behavior to avoid further litigation or to comply with the terms of their settlement

[2] The False Claims Act has a provision for enforcement by the Department of Justice without whistleblowers, if for some reason the government has information about misreporting or fraud against federal programs without a whistleblower filing a lawsuit. Since 1993, FCA lawsuits filed by whistleblowers have exceeded FCA lawsuits by the government; in 2016, there were 501 new whistleblower suits to 69 non-whistleblower suits (Civil Division, U.S. Department of Justice, 2018). Non-whistleblower suits are not used in any analysis for this paper; these statistics are included as a point of comparison.

agreements and avoid exclusion from the Medicare and Medicaid programs. Because the defendants may be only a small share of those committing impropriety, these direct deterrent effects have the potential to affect providers far exceeding the scope of the settlement. One might expect that providers who commit "rational fraud" do so having fully internalized the expected costs of their behavior, and observing settlements would not affect their decisions. However, observing settlements can either update other providers' beliefs about being caught, or increase the salience of the expected costs, thus causing behavioral changes and direct deterrence effects. Finally, behaviors that constitute litigable FCA violations may be "gray areas" of billing or care, in which case settlements can draw a clear line on what is acceptable behavior, and can prompt rule changes and clarifications from the Medicare administrators.

In addition to direct deterrence, whistleblowing can cause ex-ante deterrence by increasing litigation risk. Because anyone can file a whistleblower lawsuit, and whistleblowers regularly receive large and well-publicized payouts, providers face the threat of whistleblowing from their entire staff as well as any contractors with whom they interact. This increased risk may cause providers who have an opportunity to commit fraud to forgo the overbilling in the first place. These ex-ante deterrent effects are difficult to quantify, as they come from fraud opportunities never pursued by a provider. Even without knowing the magnitude, the ex-ante deterrence effects must weakly decrease spending on fraudulent or misreported procedures, under the assumption that providers are weakly less likely to commit fraud given increased scrutiny.

The value of deterrence effects is policy-relevant in evaluating the compensation of whistleblowers. Whistleblowers are paid a portion of the settlement recovery, which is itself proportional to the amount of damages due to pre-settlement overbilling. Therefore, whistleblowing compensation is purely retrospective. However, the value of whistleblowing depends on both the settlement and the deterrence effects, the latter of which does not factor into whistleblower compensation.

Figure 2-1 shows the relationship between the levels of damages and direct deterrence. The damages due to fraud are the difference between fraudulent and non-fraudulent spending, integrated up to the time of the lawsuit. Whistleblowers are paid 15-30% of the federal recovery, which is 2-3 times the damages. The direct deterrence effect is the integrated difference between spending without whistleblowing and spending with whistleblowing in the post-whistleblowing period. In some circumstances, the prior damages may make the expected settlement value too small to be worth pursuing. Yet from a social welfare perspective, these cases might be valuable to litigate if the direct deterrence value is large. This disconnect between whistleblower compensation and whistleblower value added may indicate that whistleblowers are inefficiently compensated in some circumstances.

There are other potential circumstances in which the direct deterrence values are small. Direct deterrence is the difference between spending with and without whistleblowing, and when these values are similar then direct deterrence is small. This could occur when the increase in spending due to fraud all occurs before the whistleblower files, and future spending would look the same with or without whistleblowing. In this circumstance, the settlement serves as a transfer from the defendant to the government and whistleblower for past bad actions, but there is no direct deterrence. However, this still retains the potential prospective benefit of deterring others from committing fraud in the first place, if observing this transfer changes their beliefs about their own enforcement probability or about the profitability of fraud. Another circumstance with little direct deterrence effect is one in which whistleblowing is not meaningful; for example, if fraud continues to be profitable even following a settlement, whistleblowing may not deter future bad behavior. In these circumstances, whistleblowing is potentially inefficient because the settlement only serves to correct retrospective damages, and the lawsuit incurs its full costs without providing social value into the future.

The timing of whistleblowing also factors into its social benefits as well as the whistleblower's compensation. The faster that fraud is litigated against, the smaller the retrospective damages and, therefore, the smaller the whistleblower's share. This could in theory cause whistleblowers to increase their payout by waiting before filing their lawsuit. However, these effects are mitigated by a priority race, in which the first-to-file whistleblower generally receives the bulk of the compensation. The False Claims Act also has a statute of limitations of at most 10 years from the date of the fraud to the filing of the whistleblower lawsuit (31 U.S. Code Section 3731, 1986). From a social welfare perspective, the timing of the whistleblower lawsuit is ambiguous, because smaller damages are reflected in greater deterrence amounts. In practice, plaintiff attorneys report that they tend to file the lawsuit as quickly as they are able to put together a good case.

Private enforcement also comes with many potential costs. Unlike other whistleblower programs, such as the IRS or SEC whistleblower programs, False Claims Act prosecution is conducted directly by the whistleblower. This eliminates any prosecutorial discretion by the government, and may lead to the litigation of cases for which there is little social harm or even explicit misconduct. Because civil lawsuits can be filed by anyone for any reason, there is a potential for spurious litigation by profit-motivated whistleblowers seeking a settlement. Defending against FCA lawsuits can be expensive, and defendants should settle if the expected cost of settlement is below the expected costs of fighting the lawsuit (and potentially losing), regardless of the truthfulness of the whistleblower's claims. Litigation is also costly for other parties, expending public resources as well as the time and costs of whistleblowers' attorneys. Even in circumstances where the government chooses not to intervene in a lawsuit, Department of Justice officials spend time reviewing all cases, and the judicial system expends resources on the litigation process. Furthermore, in the face of increased litigation risk, medical providers must undertake greater investment in compliance measures to ensure their conduct does not inadvertently violate the FCA.

There are some institutional barriers to whistleblowing that deter low-quality cases. First, whistleblowers are not allowed to represent themselves in court (United States District Court, D.C., 2003). Due to the costs of litigation, and the fact that plaintiffs' attorneys work on contingency, plaintiffs' attorneys have incentives not to take on low-quality lawsuits. This provides a barrier to filing spurious cases. Furthermore, FCA cases are most likely to be successful if the government intervenes, due to the resources and investigatory power the federal government brings when litigating a case. Since low-quality lawsuits are unlikely to generate an intervention from the federal government, this further exacerbates the unwillingness of plaintiffs' attorneys to self-fund any low-quality cases. Empirically, Kwok (2013) studies data on whistleblower attorneys and finds no evidence for "filing mills", i.e. law firms pursuing a large volume of low-quality cases.

In light of the costs of whistleblowing cases, the net efficiency of the law relies on the extent to which deterrence effects outweigh the costs of private enforcement. For this reason the measurement of deterrence effects, and the ratio of future deterrence to retrospective damages, is necessary (though not sufficient) for understanding the overall efficiency of the False Claims Act. Section 2.4 undertakes an exercise to measure these effects.

In addition to all of the cost and benefit analyses above, whistleblowing cases may also have impacts on patient care. Following lawsuits, providers may change actual care and not just the way in which it is billed. Healthcare providers have immense compliance requirements, and False Claims Act cases may inform providers' care decisions as they seek to comply with the shifting landscape of regulation and litigation risk. These changes in provider behavior can be consequential to patient health outcomes, as whistleblowing has touched such critical types of care as acute inpatient hospitalization and spine surgery. These patient health outcomes may be either an additional cost or a benefit to whistleblowing, depending on whether whistleblowing changes care in a way that benefits patients or in a way that disrupts valuable care. I expect that the patient effects differ between cases. In section 2.5, I conduct one such case study and present an example where whistleblowing changed provider care decisions for the better.

2.3 Data

The data for this project come from a variety of complementary sources which aggregate information on whistleblower cases and their downstream impacts on medical care provision and patient health outcomes.

Data on Medicare claims and payment are necessary for the analysis of the medical and fiscal impacts of whistleblowing cases. My available data include 100% samples of Fee-for-Service Medicare, i.e. Parts A and B, from 1999-2016, of all types. This includes inpatient and outpatient claims, the MedPar files that aggregate inpatient care at a hospital stay level, data on durable medical equipment, hospice care, skilled nursing facilities, home health data, Part D drug data from 2006-2016, and beneficiary information through the base files and chronic condition segments. Outpatient data are only available as cleaned files from 2002 onwards. Death dates are available at a patient level through the base files. These data, containing 100% samples of each type of care over nearly 20 years, cover tens of billions of claims from hundreds of millions of patients. Section 2.4 presents the methodology by which I selected whistleblowing cases for analysis, which translates into the usage of these data. Medicare data are used only as they related to each case presented there, and for the analysis of patient health outcomes in Section 2.5. As such, only a portion of the available data is used in these analyses.

Data on whistleblower lawsuits was compiled from multiple federal sources. Overview data on whistleblowing at a case level comes from a FOIA request I conducted on the Department of Justice for data on all completed (settled or dismissed) *qui tam* FCA cases.[3] These data describe more than 5,000 whistleblowing cases and include information on the defendant, whistleblower, filing date, federal agency to which the case relates, federal court district of filing, government intervention election status and date, settlement amount, and

[3] This data set is similar to that used in (Engstrom 2012; 2013), which also came from a DOJ FOIA request. However, to access the most recent data, I conducted an original FOIA request.

whistleblower share. These data start with the introduction of the law in 1987, and the coverage declines after 2012, as many newer cases are still under seal. These data are used for descriptive statistics and stylized facts in section 2.3.1, as well as for providing supplementary information on whistleblower lawsuits for each case study in Section 2.4. See Appendix A.1 for more details on the data cleaning process.

For substantive information on whistleblower cases related to Medicare, I scraped the Department of Justice website for all press releases related to Medicare and whistleblowing. Generally, each press release corresponds to one settlement, and my data contain 262 Medicare-related press releases through 2014. I hand-coded these press releases for the type of care and type of fraud as well as settlement value reported. I group lawsuits of similar nature into categories, and I find settlement totals within each category. As an example, one category of enforcement is the manipulation of the outlier payment system of inpatient hospitalization; this category contains 11 press releases for a total of $923 million in settlements. Each lawsuit and press release in this category contains nearly identical allegations against the defendants. Section 2.4 describes this process and presents its results.

For more detailed information on the whistleblower lawsuits for which I conduct case studies, I collected whistleblowers' original court filing documents (complaints), settlement agreements, and other court documents from a variety of sources. These documents detail exact filing dates, settlement timing, allegations of fraud, and the conduct covered by the settlement agreements. Sources for these documents include the federal court record system (PACER), the Department of Justice digital archives, SEC filings of publicly traded companies, and the legal database of Taxpayers Against Fraud, a not-for-profit supporting whistleblowers' attorneys. Combined with the press release and FOIA data, the court filings give a complete picture of the allegations and outcomes of a subset of the whistleblower lawsuits.

2.3.1 Stylized Facts About False Claims Act Lawsuits

An analysis of the DOJ case-level whistleblowing data underscores the strong incentives it provides for whistleblowers to litigate, and also raises questions about over-enforcement.

Since the 1986 introduction of qui tam whistleblowing, the volume of FCA litigation has grown immensely. The number of cases rose from 32 filed in 1987 to nearly 400 in 2012. There were 5,949 completed cases from 1999-2012, of which 3,262 come from the healthcare sector. Of these healthcare cases, only 36% result in a recovery of funds; the rest were dismissed by the whistleblower, the judge, or the Department of Justice. This points to a high level of cases for which the federal government receives no compensation, yet bears the burden of administrative costs on the court system and Department of Justice.

Figure 2-2 shows the trend of healthcare whistleblowing cases by year of filing and whether they end in a settlement. Settlements rose between 1990 and 1995 to around 50 cases per year, and have stayed rather constant since. Conversely, the total number of cases and the share of dismissed cases have both risen substantially since 1987, and continue to grow. Cases that are ultimately dismissed now constitute the majority share of whistleblowing. In terms of the raw case count, the high volume of dismissed cases reinforces the issue that privatized enforcement allows for the filing of spurious and costly cases.

Despite the relatively stable level of settled cases, settlement dollar amounts have grown substantially. Successful healthcare cases have had settlements as high as billions of dollars, although the median is substantially lower at around $1.5 million. Total healthcare settlement values have increased vastly, dominated by a few extremely large settlements. Total settlements were just $85 million in 1995, when the number of settled cases grew to its current stable levels. However, settlement totals exceed $3.5 billion in 2012, the last year of the data. The 2012 total was in a large part due to a single $1.5 billion settlement against GlaxoSmithKline for allegedly promoting its pharmaceuticals for non-FDA-approved uses. Appendix Figure A1 plots the histogram of healthcare-related settlement values. The healthcare-related settlements in my data total to $26.4 billion of recovery.

Whistleblowers face a potential for a very large payout. Among settled healthcare whistleblowing cases, the mean whistleblower payout is $3.8 million. However, the distribution has a long right tail; median whistleblower payout is only $250,000, while 4 cases have had whistleblower payouts in excess of $100 million. Appendix Figure A2 plots the histogram of healthcare-related whistleblower shares. Total whistleblower payouts for all of the healthcare-related cases in my data is $4.29 billion. These awards are split between the whistleblower and their attorney, who usually take 30%.

The Department of Justice Data also include lawsuits from outside of the medical field, and exhibit the broad use of the False Claims Act. Medical-related suits, those categorized by the DOJ as relating to the Department of Health and Human Services, the Food and Drug Administration, or the Center for Medicare and Medicaid Services, constitute 55% of cases. But suits regarding the Department of Defense account for 11% of the nearly 6,000 whistleblower lawsuits, and cases have arisen from nearly all parts of the federal government, including the Department of Education (3% of cases) and the Goods and Services Administration (2% of cases). The use of FCA whistleblowing outside of the medical field is beyond the scope of this paper and poses an opportunity for future research.

2.4 Deterrence Effects

The economic effect of False Claims Act cases depends not only on the money they recover, but also the savings to the government from fraud or misreporting that is not committed due to the deterrent effects of these lawsuits. This deterrence takes two forms: direct deterrence, from changes in spending on types of care named as improper in whistleblower lawsuits, and ex ante deterrence, from fraud that is never committed in the first place due to litigation risk. This analysis measures the dollar value of the direct deterrence of the largest categories of whistleblower cases. However it does not consider the ex ante deterrence, which is substantially more difficult to measure, and therefore provides a lower bound of the total deterrence effects.

2.4.1 Method

Motivation

The goal of this analysis is to estimate the amount of direct deterrence caused by FCA lawsuits that change provider behavior and Medicare spending. The measurement of deterrence requires an analysis of a counterfactual, between the real world in which enforcement happened and one in which it did not. Synthetic controls, first introduced in Abadie and Gardeazabal (2003), provide a mechanism by which to produce such a counterfactual. Here, the outcome of interest is spending on the type of medical care treated by whistleblowing, and the treatment effect of interest is the change in spending following whistleblowing. The idea of synthetic controls is to use untreated control groups to construct a series that most closely matches the treated unit in the pre-treatment periods. Then, the difference between the treated unit and the synthetic control group in the post-treatment periods can be used to measure treatment effects.

My analysis builds on the traditional synthetic control method with the inclusion of a time-shift parameter for each control. Under traditional synthetic controls, control groups are used to estimate the counterfactual of the treated unit under the assumption of shared time fixed effects. In many circumstances, the assumption of shared time fixed effects is valuable. For the whistleblowing cases estimated here, as well as in other situations, these assumptions do not accurately reflect the circumstances. Whistleblowing often affects types of care with unusual trends: they exhibit high growth in spending and claims, potentially driven by the improper conduct of the defendants. Yet there are many reasons why certain types of medical care may exhibit a rise in spending besides fraud, such as technological changes or changes in best practice. When the treated unit is on a different time trend than the untreated units, traditional synthetic controls can fail to find a sufficient fit in the pre-treatment periods.

The addition of a time-shift parameter allows for the treated unit to be compared to control units that exhibit similar patterns at different points in time. When a treated unit experiences a large pre-treatment increase and a subsequent post-treatment fall, the relevant counterfactual question is whether the increase would have persisted in the absence of treatment. Rather than comparing the treated unit to untreated units at the same time, which may not have seen the same pre-treatment increase, the relevant control units are other treatments that saw a similar rise at other points the available data. In practice, this is implemented with a two step procedure: first, I shift the control units forward or back in time to match the pre-treatment series of the treated type of spending; and second, I construct a synthetic control group by assigning weights to the time-shifted controls. The resulting synthetic control unit is used to estimate the post-treatment trends of the treated unit in the absence of treatment. The following model details the econometrics of this procedure and closely resembles that of Abadie et al. (2010).

Model

Synthetic controls with time shifts are motivated by the assumption that treated and untreated units have common trends at different times.

Consider spending on a type of medical care that could be affected by whistleblowing treatment. For $i = 1, \dots, N$ and time t, we would observe the spending level Y_{it}^{U} in the absence of treatment and Y_{it}^{I} following treatment. Call T_i the treatment period for unit i, which is the filing of the whistleblower's lawsuit. Assume that the treatment has no effect on periods $t < T_i$; then $Y_{it}^{U} = Y_{it}^{I}$ for all $t < T_i$. Let δ_{it} be the effect of treatment at time

t. Because Y_{it} represents spending, δ_{1t} represents the change in spending on a procedure following whistleblowing. Thus the spending level can be written as:

$$Y_{it}^{I} = Y_{it}^{U} + \delta_{it}\mathbb{I}_{t \geq T_i}$$

Let $i = 1$ be the unit treated by whistleblowing, which is the only unit subject to treatment; thus $Y_{it}^{U} = Y_{it}^{I}$ for all $i > 1$ for all t, and units $i > 1$ serve as control groups. The treatment effect of interest is δ_{1t} , which is given by:

$$\delta_{1t} = Y_{1t}^{I} - Y_{1t}^{U}$$

in periods $t \geq T_i$.

Because Y_{1t}^{I} is always observed for all times $t \geq T_i$, estimation of the treatment effect relies on estimation of Y_{1t}^{U}, which is not observed in those periods.

Suppose that control units exhibit similar time trends at different points in calendar time, beginning at t_{0i}, which varies between units. Suppose that for all units, Y_{it}^{U} is given by the factor model:

$$Y_{it}^{U} = \kappa_{\tau} + \lambda_{\tau}\mu_i + \epsilon_{it} \tag{2.1}$$

Here, $\tau = t - t_{0i}$ is the time after the start of the control unit's trend begins; κ_{τ} is a common shock across all units at time τ relative to the starting point; λ_{τ} is a vector of common factors describing the trajectory of an outcome along a common trend; the parameter μ_i is an unknown vector describing the individual factor weights; and ϵ_{it} is a set of unobserved shocks of 0 mean.

Consider a $(N - 1 \times 1)$ vector of weights $\vec{W} = (w_{2,}w_{3,} \dots, w_N)$, such that $w_i \geq 0$ for $i = 2, \dots, N$ and $\sum_{i=2}^{N} w_i = 1$. These values represent weights on the untreated control units, and every value of the vector $\vec{W}$ represents a possible synthetic control. Then, a weighted average of the control units is given by:

$$\sum_{i=2}^{N} w_i Y_{it} = \kappa_\tau \sum_{i=2}^{N} w_i + \lambda_\tau \sum_{i=2}^{N} w_i \mu_i + \sum_{i=2}^{N} w_i \epsilon_{it}$$

If weights w_i^* can be constructed such that:

$$\sum_{i=2}^{N} w_i^* \mu_i = \mu_1$$

Then it holds that

$$E[\sum_{i=2}^{N} w_i^* Y_{it}] = E[\kappa_\tau + \lambda_\tau \sum_{i=2}^{N} w_i^* \mu_i + \sum_{i=2}^{N} w_i^* \epsilon_{it}]$$

$$= E[\kappa_\tau + \lambda_\tau \mu_1] + \sum_{i=2}^{N} w_i E[\epsilon_{it}] = E[Y_{1t}^U]$$

Therefore, the weighted average of the control units provides an unbiased estimator of the untreated counterfactual of the treated unit:

$$\widehat{Y_{1t}^U} = \sum_{i=2}^{N} w_i^* Y_{it}$$

and we can estimate $\widehat{\delta_{1t}} = \widehat{Y_{1t}^U} - Y_{1t}^I$.

By integrating $\widehat{\delta_{1t}}$ over the post-treatment periods, we can estimate a discounted deterrence effect:

$$D = \int_{t=T_1} (\widehat{Y_{1t}^U} - Y_{1t}^I) \beta^{t-T_1} dt \tag{2.2}$$

where T_1 is the treatment period and β^{t-T_1} is a discount factor starting at the treatment period.

Implementation

The practical estimation of this model can be performed as a two-step procedure: shifting control units in time to match the pre-treatment spending on the treated unit, and then finding synthetic control weights w_i.

First, I align the control units with the treated unit to ensure they are on common trends in the pre-treatment period. For each control, I construct a set of leads and lags, and find the lead or lag with the best fit to the treated series in the pre-period. With any fixed set of data, producing leads and lags creates missing data at the front or back of the series: in a monthly series, if one uses a 5 month lag, the first 5 months of available data have no value. In practice, this means that shifting the control units too far forward or back in time leaves a limited set of data for the evaluation of pre-treatment fit and post-treatment effects. Here, I bound the time shifts to ensure that there are 36 months of pre-treatment data, used to construct the synthetic control weights, and 60 months of post-treatment data, used to compute the deterrence effect.[4] Within these bounds, I select the appropriate lead or lag for each control unit that minimizes average square distance from the treated unit in the pre-period:

$$\min_{d} \frac{\sum_t^M (Y_{1t} - Y_{it+d})^2}{M} \tag{2.3}$$

for control unit i, where d indexes the different leads and lags, and M is the number of pre-treatment periods in which the shifted control and the treated unit overlap. [5] Figure

[4] In circumstances where the treatment period is too close to the start of the data, 36 pre-treatment periods are unavailable, and all of the available periods are used.

[5] Average square distance and not total square distance must be used because the number of points over which this sum is evaluated depends on the time shift.

2-3 provides a simple graphical explanation of the time-shifting process for two controls, one shifted forward in time and one shifted backward.

After associating each control with a time shift, I conduct the standard synthetic control process to choose weights as per Abadie et al. (2010). Weights are chosen to minimize mean-square error over all pre-treatment periods in which all of the controls overlap:

$$\min_{\vec{W}} \sum_{t \in M^*} (Y_{1t} - \sum_{i=2}^{N} w_i Y_{it+d^*})^2 \tag{2.4}$$

where M^* is the set of periods for which all of the time-shifted controls overlap with the treated unit; d^* is the optimal time shift found by 2.3; and $\vec{W}$ is the set of all potential $(N-1 \times 1)$ vectors of weights $(w_{2,}, \ldots, w_N)$ where $w_i \geq 0$ and $\sum_{i=2}^{N} w_i = 1$. Given the optimal shift for each control, the Stata package "Synth" finds the optimal weights w_i^*.

Once these weights are found, the synthetic control unit is produced as the weighted sum of the control groups:

$$\widehat{Y_{1t}^{U}} = \sum_{i=2}^{N} w_i^* Y_{it+d^*}$$

The two-step procedure for time-shifted synthetic controls is a tractable way to implement this methodology by leveraging existing methods. Separating the time shift component from the weighting component is necessary, because allowing weights to be applied to the entire set of leads and lags could produce synthetic control units constructed of multiple instances of the same control at different points in time. However, a better search algorithm over the full space of time shifts and weights could theoretically outperform the two-step procedure by jointly choosing time shifts with weights to better estimate the treated unit in the pre-treatment periods.

In this paper, the outcome variables Y_{it} are all spending amounts. Therefore, the difference between the synthetic control and treated unit is a difference in spending, which can be integrated over the post-treatment periods. I estimate the direct deterrence effect as the discounted difference between the treated and synthetic control over 5 years post-treatment:

$$\widehat{D} = \sum_{t=T}^{T+60} \frac{\widehat{Y_{1t}^{U}} - Y_{1t}}{(1.1^{1/12})^{t-T}} \tag{2.5}$$

where t is the time in months, T is the treatment period, and the denominator $1.1^{1/12}$ provides the monthly rate for a 10% annual discount rate. Deterrence is totaled for 5 years from the treatment date of filing. A positive deterrence value indicates that post-whistleblowing spending Y_{1t} is lower than that of the synthetic control group $\widehat{Y_{1t}^{U}}$.

The analysis conducted here is conservative in two ways. First, the use of time-shifted synthetic controls extends the donor pool of potential controls to anything on a similar pre-whistleblowing trajectory as the treated units at any time. Because a time shift of 0 is available for all control units, time-shifted synthetic controls uses a superset of the controls used for a traditional synthetic control methodology. Second, I compute deterrence using only 5-years of post-treatment effects. In the absence of whistleblowing, fraud or abuse may have continued indefinitely into the future, in which case the total deterrence effect would be like a perpetuity, providing value at all later periods. Rather than assume that the deterrence effects persist indefinitely, the use of 5 years of post-treatment effects is a conservative estimate of the deterrence effects and avoids excess extrapolation.

In order to conduct inference on my results, I employ a permutation test as per Abadie et al. (2010). Each synthetic control is substituted in for the treated unit, and the same two-step procedure detailed above is performed, fitting leads and lags and constructing weights, using all other controls. These weights give a synthetic control unit for the placebo, from which the deterrence measurement can be computed. The deterrence effects corresponding to each control unit form an empirical distribution against which the deterrence effect of the treated unit can be compared.

2.4.2 Case Selection

FCA whistleblowing has impacted many different types of Medicare payment and care delivery. Yet mapping from data on whistleblower lawsuits to the Medicare data is difficult, hindering a complete analysis of all lawsuits. The DOJ case-level data do not provide information on the nature of the cases nor the alleged false claims, beyond naming the defendant and whistleblower. To find details on successful whistleblowing cases, I scraped the universe of press releases from the Department of Justice website that relate to Medicare and whistleblowing, from 1994 (the start of the archives) through 2014. From these press releases, I hand-coded categories of cases that contain similar allegations of fraud in similar types of medical care. For example, one category is the misuse of the outlier payment system, which contains 11 press releases from different settlements with similar allegations. I omit categories for which I do not have data, including enforcement that precedes the start of my data, or allegations related to falsification not visible in the Medicare claims. Appendix A.2 describes this process in detail. Table 2.1 lists the 4 categories of enforcement with the largest total settlement amounts for which I have data, which comprise 29 press releases detailing $1.9 billion in total settlements. In the following sections, I conduct case studies for each of these 4 categories. For each case study, I use court documents including whistleblower complaints and settlement agreements to gather details about the alleged conduct and guide the analysis of claims.

2.4.3 Case Details

Outlier Payment Falsification

The first category concerns the misuse of outlier payments for inpatient hospitalization, for which over $900 million in settlements was recovered by the government between 2004 and 2010. Medicare pays providers of inpatient medical care a fixed reimbursement amount for the diagnosis related group (DRG) under which the patient is coded. By fixing reimbursement for each diagnosis, providers have incentives to keep costs down. However, this raises concerns that providers would be unwilling to treat high-cost patients. In response to these concerns, the Medicare system contains a provision for outlier payments, which are additional reimbursements for very-high-cost patients. Before 2004, to qualify for outlier payments, a patient must have exceeded a cost threshold, computed with a complicated formula based on the provider's labor costs, capital costs, historic charges, and a geographic adjustment factor [6]. This formula provided an opportunity for misreporting: by manipulating charges over time, hospitals were able to change their thresholds and collect more outlier payments.

On November 4, 2002, Tenet Healthcare, a large investor-owned hospital company, was sued under the False Claims Act for manipulating its cost reports in order to illicitly receive additional outlier payments. [7][8] This lawsuit was settled in June, 2006, with Tenet paying $788 million to resolve these allegations without admission of guilt. The DOJ press releases describe 10 other settlements for alleged manipulation of outlier payments. Appendix A.3

[6] Rawlings and Aaron (2005) provide a detailed analysis of this computation.

[7] One lawsuit against HealthSouth, settled in 2004, contains allegations about outlier payment manipulation and an $89 million settlement for those allegations. This settlement related to lawsuits originally filed in 1998, but the original allegations did not concern outlier payments. The HealthSouth settlement does not attribute these allegations to any whistleblower cases (which would make the whistleblowers eligible for compensation for this portion of the recovery); instead, it appears the DOJ added enforcement of outlier overuse to a pending case, following the filing of the Tenet lawsuit. Therefore, I consider the Tenet lawsuit the first outlier lawsuit, and use its filing date as the treatment date.

[8] Around the time of filing, Tenet also received substantial negative press regarding its overuse of outlier payments; the timing of these reports, days before the filing of the lawsuit, may indicate that the whistleblowing case was leaked to investors. Rawlings and Aaron (2005) provide further details about the Tenet case.

contains additional details about the related lawsuits.

Outlier payments constitute their own type of spending by the Medicare system. Therefore, for its controls, I consider all types of payments made by Medicare that are observable in my data. The controls I use include durable medical equipment, home health care, hospice care, nursing care, and disproportionate share payments for hospitals that serve many low-income patients. The two largest categories of Medicare expenditures, inpatient and outpatient claims, are not comparable in scale to outlier payments and are omitted from the donor pool. Spending on drugs via Part D is also not included because these data only start after 2006.

Medically Unnecessary Botox

The second case regards medically unnecessary usage of Botox. Despite popular branding as an "anti-wrinkle" procedure, Botox is FDA-approved for a number of important medical uses, including treatment of crossed eyes (strabismus) and neck spasms (cervical dystonia). Medicare covers medically necessary Botox injections for FDA-approved uses, but not for non-FDA-approved uses. Between 2007 and 2009, Allergan, the sole manufacturer of Botox, was sued by a set of whistleblowers who alleged that Allergan had illegally promoted Botox for non-FDA-approved ("off-label") uses, including headaches. In order to ensure that Medicare would pay for the injections, Allergan allegedly instructed physicians to miscode the injections, using diagnosis codes for approved uses. Additional details about the outpatient coding of Botox and the whistleblower lawsuits are presented in Appendix A.3. On August 31, 2010, Allergan settled with the federal government for $600 million, of which $210 million was for federal civil liability, $375 million was a criminal fine, and $14.85 million was to recompense affected state Medicaid programs.

For the synthetic control design, Botox is compared to other outpatient procedures that saw similar pre-whistleblowing trends in spending. Spending for Botox under the relevant diagnoses codes grew from 2 million dollars in 2003 to more than $5 million in 2006, the year before the lawsuit against Allergan was filed. As controls, I use other outpatient treatments for which spending started between $2 million and $5 million and saw a 2-3x rise over any 3-year period between 2002 and 2011, of which there are 67 control units. Appendix A.3 contains additional details about these control units.

Unnecessary Inpatient Kyphoplasty

Kyphoplasty is a spine procedure to repair vertebral compression fractures that cause pain and deformity of the back, often observed among patients with osteoporosis. Kyphoplasty involves the percutaneous (through the skin) injection of bone cement into an inflatable balloon placed within the affected vertebra. Because the procedure is performed percutaneously, kyphoplasty can be safely conducted as an outpatient procedure. The kyphoplasty procedure was developed, patented, and marketed by the company Kyphon, which sold a spine surgery kit as well as other related medical devices (Kasper, 2010).

In December, 2005, Kyphon was sued by FCA whistleblowers who alleged that Kyphon illegally promoted the procedure as an inpatient procedure, as opposed to outpatient, to receive greater reimbursement. Furthermore, by keeping patients inpatient for a short stay, providers could receive the inpatient reimbursement level for a low amount of inpatient care, very similar to the outpatient procedure. Inpatient stays under the relevant diagnosis-related groups (DRGs) were reimbursed in the $6,000 - $11,000 range, as opposed to outpatient kyphoplasty which was reimbursed between $500 and $2000. In May 2008, Kyphon settled these allegations with the Department of Justice for $75 Million, without admission of guilt. Between 2009 and 2015, the DOJ released another 9 press releases detailing settlements with 140 hospitals having performed unnecessary inpatient kyphoplasty. The sum of the settlements against Kyphon and the defendant hospitals was $214.2 Million. Appendix A.3 provides additional details for these lawsuits. I analyze spending on short stays of 7 or fewer nights under the inpatient DRGs promoted by Kyphon, and outpatient spending on all spine procedures. As controls for short-stay inpatient visits, I use inpatient spending for short stays of 7 or fewer nights under other DRGs. For controls on outpatient spine procedures, I use spending on other outpatient surgical procedures on the musculoskeletal system. Appendix A.3 describes the coding of inpatient and outpatient Kyphoplasty and the control units used.

Unnecessary Inpatient Admission

The fourth case study concerns the unnecessary admission of Medicare beneficiaries for inpatient care at hospitals, instead of receiving observational or outpatient care. Starting in 2004, there were a series of whistleblowing lawsuits that alleged that hospitals around the country unnecessarily admitted patients. Many of these patients presented at the hospital's emergency department and should have been held under observational or outpatient status, which are reimbursed much less than inpatient care. The first successful lawsuit of this type was filed in October 2004, and in total, 7 settlements were reached regarding 135 hospitals for a total of $172.3 million in recovery between 2007 and 2014. The majority of the enforcement comes from the settlement with Community Health Systems, the country's largest operator of acute care hospitals at the time, which settled for $98 million in 2014 for similar conduct in 119 of their hospitals. Appendix A.3 provides additional details about these lawsuits.

Unnecessary admissions were concentrated among certain hospitals, motivating an analysis of the defendants of these lawsuits. The set of potential controls for the defendants are other hospitals not litigated against for unnecessary inpatient procedures. To mitigate spillover effects into the control groups, I restrict the controls to hospitals in the 23 states that contained no defendants. These hospitals treat different patient pools than the defendants and are less likely to have doctors or administrators cross-employed with the defendant hospitals. For each of the defendants, I construct a random sample of 100 control units, using the same number of hospitals as the defendant. For example, two defendants were groups of 6 hospitals each; I create 100 control units of 6 randomly grouped control hospitals, drawn with replacement, from the set of controls. Similarly, to measure substitution by the defendant providers to increased outpatient expenditure, I use randomly grouped outpatient providers from the unaffected states. Appendix A.3 provides further details about this process.

2.4.4 Results

Figure 2-4 shows the main results of the synthetic control method. Spending is used as the outcome variable for each case study. For each case, the first vertical line shows the date of filing of the first lawsuit of that category, which is used as the treatment date, and the second line shows the first settlement. For the unnecessary inpatient admissions case, where I analyze multiple defendants, this graph includes the results from Community Health Systems (CHS), which constituted the bulk of the defendant hospitals. In each case, the control groups provide a good pre-period fit for each case, matching the trends and levels of the treated units. However, there is divergence from the control units in each of the post-whistleblowing periods.

The largest of these effects is in the outlier case (top left): the 5-year discounted deterrence measurement for the outlier payments computed is $17.46 billion, which is roughly 19 times the total settlement value of the outlier whistleblowing lawsuits of $923 million. This is largely driven by the scale of spending on outlier payments, which exceeded $500 million per month in its pre-whistleblowing peak, and then dropped off substantially following the lawsuits. The synthetic control group, constructed of other types of Medicare payments, shows a similar pre-period rise and continues to rise in the absence of whistleblowing. Appendix Table A3 shows the synthetic control weights and time shifts for the control units. Most of the synthetic control weight is placed on disproportionate share payments, with only a 1-month time shift. Disproportionate share payments operate very similarly to outlier payments in that they are additional payments for inpatient stays. The fact that the synthetic control is mostly made up of such a similar type of spending, and with very little time shift necessary to fit the pre-period, reinforces the validity of this control group in estimating outlier payments in the counterfactual without whistleblowing.

Notably, for the Botox case (top right), there is a small negative deterrence effect: Botox spending exceeds the synthetic control group post-lawsuit. The 5-year discounted deterrence effect is -$3.99 million, a little smaller than 1% of the settlement value of $600 million. One potential reason for the negative deterrence effect is that Botox gained FDA approval for migraine coverage about 2 months months after settling with the Department of Justice for illegally promoting botox for headaches (Singer, 2010). Because civil litigation and settlement negotiations can stretch out for indefinite periods of time, it is possible that Allergan timed the settlement to coincide with its expected FDA approval. This case exhibits that deterrence effects are not necessarily positive, and that the future value of misconduct is not necessarily large when compared to the past costs and settlement amount. In this circumstance, the $600 million settlement paid by Allergan to the United States functioned as a penalty for promoting its product for a use that was not yet FDA approved. But given that FDA approval did ultimately arise, the future value of the damages and the direct deterrence effect are small.

For the kyphoplasty case (bottom left) and unnecessary admissions case (bottom right), Figure 4 shows that inpatient spending declined relative to the respective synthetic controls. The short-stay inpatient deterrence total for kyphoplasty is $538.9 million, caused by a post-whistleblowing decline in short kyphoplasty stays and a continued increase in the synthetic control group. Appendix Table A5 displays the synthetic control weight and time shift for each control group. For the unnecessary inpatient admissions case, inpatient deterrence for the defendant Community Health Systems is $693.2 million, and the total inpatient deterrence for all defendants is $1.124 billion. The corresponding plots for the other defendants are included in Appendix Figure A3. Inpatient spending at CHS rose before whistleblowing and fell after whistleblowing, while the synthetic control group shows a continued rise in inpatient spending in the absence of whistleblowing. While the difference between CHS spending and the synthetic control group is visually modest, the large deterrence amount arises because the rate of spending at the time of treatment was more than $150 million per month.

The decreases in inpatient spending in the kyphoplasty case and the unnecessary admissions case must be weighed against an expected increase in outpatient spending in both cases. Figure 2-5 plots the substitute outpatient spending for these cases. Inpatient kyphoplasty spending was substituted for outpatient spending on kyphoplasty, vertebroplasty, and other outpatient spine procedures, as evidenced by a rise in outpatient spending relative to the synthetic control group. The increase in outpatient spending totals to $257.8 million; when compared with an inpatient spending decrease of $538.9 million, this results in a net deterrence effect of $281.1 million. For the unnecessary admissions case, outpatient spending at the defendant CHS did not rise relative to the control providers. The total estimated change in outpatient spending for all defendants combined is a small negative number, indicating if anything that outpatient spending at these hospitals fell post-lawsuit. Appendix Figure A4 displays the similar synthetic control setup for each of the other defendants' outpatient spending, and shows heterogeneity, with some defendants' outpatient spending rising post-lawsuit and others' falling.

Table 2.2 summarizes the deterrence effects for these cases and provides totals, deterrence values, and deterrence ratios. These 4 whistleblower case studies recovered around $1.9 billion for the federal government, but exhibit even greater benefit in deterrence effects, totaling around $18.9 billion. The average deterrence effect for these cases is around 6.7 times the settlement value over 5 years. There is substantial heterogeneity in the deterrence ratios, from a small negative deterrence effect in the botox case to a particularly high positive deterrence ratio for the outliers case. Notably, the deterrence metric used here is computed using a discounted difference between the treated and control units for only 5 years after the filing of the case, and thus may potentially understate the true benefit of whistleblowing. Overall, these results indicate that the direct deterrence benefits of whistleblowing cases often exceed the settlement values many times over, and greatly exceed the retrospective damages used to compute those settlement values. This indicates a large savings to the Medicare program as a result of these whistleblowing cases, exceeding both the direct recoveries to the government from the settlement as well as the whistleblower compensation.

Placebo testing allows for inference on these results. For each control, I conduct a series of placebo studies by applying the time-shifted synthetic control method to a every other control group in the donor pool. The deterrence for each real treated unit is compared to the deterrence values from each of these placebos. I conduct a 1-tailed test, which counts what fraction of placebos exceed the value of the treated unit's deterrence amounts, comparing positive deterrence values to other positive deterrence values and negative to negative. Importantly, the distribution of placebo controls is only one piece of evaluating the synthetic control strategy, and does not account for pre-period fit or the timing of the post-whistleblowing trends. Therefore, while my estimates are statistically significant, these placebo results must also be evaluated in the context of the good pre-period fit of the results displayed in Figures 2-4 and 2-5.

Table 2.3 presents the results of the placebo test. These results indicate that the large positive deterrence effects I find are not due to chance, while the small negative finding for botox is indistinguishable from the value of the placebos. The deterrence figure for outlier payments exceeds 100% of the placebo units. The small negative deterrence effect for Botox does not exceed 10 of the 67 placebos, indicating that this effect may be due to noise, and that the negative point estimate is not distinguishable from 0. For the kyphoplasty case, the reduction in inpatient spending exceeds 26 of the 30 placebos, and the corresponding increase in outpatient spending exceeds 14 of the 15 placebos. For the unnecessary inpatient case, there is strong evidence that the reduction in inpatient spending is not a chance finding; the 5 largest defendants (of 7) exceed between 93 and 99 of the 100 placebo units. However, substitution to outpatient spending shows mixed results, including statistically significant values in both the positive and negative direction. This mix of positive and negative effects indicates heterogeneity in how whistleblowing changed outpatient spending at the defendant hospitals.

2.5 Whistleblower-Induced Changes in Patient Care

In addition to the fiscal effects described in the previous section, whistleblowing under the FCA creates incentives for providers to change the way they conduct healthcare. These changes can could be either positive or negative: if whistleblowing curbs behavior that was profitable to providers at the expense of patient health, then we expect whistleblowing to benefit patients. However, whistleblowing could also induce defensive behavior among physicians, influencing their care decisions away from what is beneficial to patients and instead to what would be justifiable if they were sued.

To examine one instance of these effects, I examine provider care decisions following changes in kyphoplasty care as discussed above. Kyphoplasty is an ideal case study for the discussion of provider care decisions for a few reasons. First, the analysis of claims performed above shows that there was a large reduction in inpatient procedures and a substitution to outpatient procedures, indicating a change in actual care decisions by providers. This is in contrast to the outlier payments case, which seems to be a change in billing procedures, or to the Botox case, where there was little effect on usage. Second, kyphoplasty is a single procedure with previously-studied health effects (discussed below), allowing for a targeted analysis of the effects of the whistleblowing lawsuits on patient care. This is in contrast to the unnecessary inpatient admissions case, which related to a broad set of medical procedures. As such, kyphoplasty is the largest-settlement-value case study for which I can conduct an analysis of provider care decisions.

Kyphoplasty is a spine procedure to repair compressed vertebrae in patients with osteoporosis, and can be very beneficial for patient health. Kyphoplasty is similar to vertebroplasty, a similar, older procedure that does not use a balloon (Denaro et al., 2009). A meta-analysis of vertebral compression fractures (VCFs) shows that patients with VCFs have 2.5 times the mortality rate of patients without them, and that kyphoplasty and vertebroplasty are successful at reducing mortality rates compared with non-operative care (Kurra et al., 2018). Estimates in this meta-analysis range from 35% to 70% mortality reduction over a 3 to 5 year period after receiving kyphoplasty, indicating a substantial health benefit to the procedure. As shown in above, short-stay inpatient treatment for kyphoplasty was drastically reduced following whistleblowing, which could therefore have an impact on patient mortality.

To understand the impact of these changes, I model patient death as a function of receiving inpatient kyphoplasty within a heterogeneous treatment effects framework. The goal of this exercise is to measure how provider care decisions changed due to whistleblowing, based on whether the patient is expected to benefit from the procedure. First, I construct two non-overlapping cohorts of patients from before and after the Kyphoplasty lawsuit. For the purposes of this analysis, I define "treatment" as receiving a 1-night inpatient stay under the DRGs allegedly promoted by Kyphon[9]. My 2005 cohort includes every 70-75 year old treated for the first time in 2005 and the full population of never-before-treated 70-75 year old control Medicare patients. My 2011 sample similarly contains every every 70-75 year

[9] In contrast to this setup, the spending amounts used to compute deterrence in Section 2.4 used all short stays (7 nights or fewer) under these DRGs and all outpatient spine spending. This was useful to compute fiscal effects, but potentially lumped in some non-kyphoplasty treatment, which was unaffected by whistleblowing and was differenced out when computing deterrence values. To compute health effects, I focus on 0 or 1-night stays under these DRGs, which almost completely vanished post-whistleblowing, as I have greater confidence that kyphoplasty was performed during these inpatient visits. Correspondingly, outpatient treatment is restricted to the outpatient codes specifically for kyphoplasty and vertebroplasty.

old treated for the first time in 2011 and the full population of never-before-treated 70-75 year old control patients. These cohorts are non-overlapping, and therefore all 2011 cohort members are not potential controls for the 2005 cohort. The 2005 cohort consists of 8.2 million patients, and the 2011 cohort consists of 9.3 million patients.

For each patient, I analyze extensive data, including treatment, all inpatient medical history, chronic conditions, and demographic data. Patient covariates include age, state, sex, race, original and current reasons for Medicare qualifications (i.e. age or disability), and HMO coverage. Inpatient medical history was taken from the 100% MedPar files for 6 years before the cohort year, i.e. 1999-2004 for the 2005 cohort and 2005-2010 for the 2011 cohort, and includes an indicator for any inpatient stay, the number of stays, the patient's total inpatient stay length, a count of the number of stays under each DRG, and the total Medicare payment amount for that patient's inpatient treatment. Furthermore, I include chronic condition indicators for each patient in the cohort year, which detail a patient's history of chronic conditions such as Alzheimer's, hip fractures, or osteoporosis. Finally, for each patient I collect death dates, and produce an indicator of whether the patient died within 5 calendar years post-surgery, i.e. 2005-2010 for the 2005 cohort and 2011-2016 for the 2011 cohort. The length of the medical history and death data used are due to the availability of data, which span from 1999-2016, and the requirement that the cohorts not overlap.

Using these data, I directly estimate the heterogeneous treatment effects on mortality of receiving a short-stay inpatient treatment among the 2005 cohort, with the following logistic regression:

$$Death_i = \alpha + \beta T_i + \gamma' C_i + \delta' T_i C_i + \eta' M_i + \theta' T_i M_i + \varepsilon_i \tag{2.6}$$

The outcome variable $Death_i$ is an indicator if patient i died within 5 years. This limited dependent variable motivates a logit framework for the regression. T_i is the treatment indicator, C_i is the matrix of patient covariates and M_i is the matrix of patient medical history and chronic conditions. This specification models death in terms of treatment interacted fully with these controls. As such, the fitted model captures the effect of each covariate and each aspect of medical history on death, with or without short-stay inpatient kyphoplasty treatment. Appendix Table A1 details the results of this regression.

Using this model fitted to the 2005 sample, I can then predict the effects of kyphoplasty among both 2005 and 2011 patients. I construct: $\widehat{Y_{i1}} = P(Death|M_i,C_i,T_i = 1)$ and $\widehat{Y_{i0}} = P(Death|M_i,C_i,T_i = 0)$ for each patient, using the regression coefficients fit to the 2005 sample. I then produce a predicted treatment effect for each patient:

$$\widehat{TE_i} = \widehat{Y_{i1}} - \widehat{Y_{i0}} = P(Death|M_i,C_i,T_i = 1) - P(Death|M_i,C_i,T_i = 0) \tag{2.7}$$

It is important to note that treatment T_i is not assigned randomly. This model makes the standard conditional independence assumption: that conditional on a rich set of controls, here medical history, chronic conditions, and patient covariates, that potential outcomes Y_{1i} and Y_{0i} under treatment or non-treatment are independent of actual treatment status. That is, by controlling for the factors that influence probability of treatment, one can construct both potential outcomes for each patient, despite only ever observing Y_1 or Y_0 for any given patient. Appendix Figure A7 shows the histogram of estimated treatment effects for patients in 2005 and 2011. These histograms exhibit a similar shape between the cohorts, which means the comparison between these cohorts is between like populations.

Figure 2-6 plots the probability of short-stay inpatient treatment by estimated treatment effect in each cohort [10]. The estimated treatment effect on the horizontal axis is the change in mortality from receiving treatment versus not receiving treatment. Units to the left of 0

[10] To satisfy Medicare data cell-size suppression rules against reporting any result with $n < 11$, the bins in Figure 2-6 have been top coded at .4 and -.2 to achieve a minimum number of treated units within the extreme bins.

are estimated to have reduced mortality if treated, while units to the right of 0 have increased mortality if treated. The ideal targeting of treatment to patients who benefit would place all of the mass to the left of 0. In both 2005 and 2011, patients who stood to benefit from the procedure were about twice as likely to receive treatment. The comparison between 2005 and 2011 shows that there was a reduction in inpatient probability treatment across the spectrum of treatment effects, both for those harmed and helped by the treatment.

Total inpatient treatment volume for kyphoplasty was counteracted by substitution to outpatient treatments. Figure 2-7 plots the probability of receiving an outpatient kyphoplasty or vertebroplasty within each group. Because these procedures are similar whether performed inpatient or outpatient, differing mostly in terms of billing, the estimated inpatient treatment effect on the horizontal axis is a reasonable way of understanding the effect of having had the procedure in either location. In both cohorts, the probability of receiving treatment is higher for those helped by the treatment, to the left of the distribution. Between 2005 and 2011, outpatient treatment probability grew for all types of patients, but substantially more for patients for whom kyphoplasty is expected to reduce mortality [11].

To examine the net effect of the substitution from inpatient to outpatient procedures, I examine the probability of either inpatient or outpatient treatment by heterogeneous treatment effect. Figure 2-8 breaks the population into two categories: those with reduced mortality if they receive the procedure (negative value treatment effect) and those with increased mortality (positive valued treatment effect). The results show an overall decrease in treatment probability to those harmed by the procedure between 2005 and 2011, and an increase in treatment probability to those who benefit from the treatment across the same time period. Those helped by the procedure saw an increase from 0.144% to 0.155% probability of treatment, a 7.6% increase. The small raw percentages reflect the fact that

[11] Similar to Figure 2-6 , bins in Figure 2-7 have been top coded to achieve a minimum number of treated units within the extreme bins, in compliance with Medicare cell-size suppression rules.

the analysis sample is the entire never-before-Medicare treated 70-75 year old population in these years, and that kyphoplasty is relatively rare. Correspondingly, patients who were expected to by harmed by the procedure saw a decrease in probability of treatment from 0.0547% to 0.0506%, a decrease of 7.4%. Appendix Figure A9 breaks this same analysis into finer groups, and plots the probability of receiving either inpatient kyphoplasty or outpatient kyphoplasty or vertebroplasty by heterogeneous treatment effects. There was, in general, an increase in the probability of receiving treatment between 2005 and 2011 for those who benefit from the procedure, and a decrease in treatment for those harmed by the procedure. As a caveat, this analysis uses a coarse measure of patient health, mortality, which does not capture changes to patient quality of life from receiving treatment.

These results are consistent with the better targeting of kyphoplasty following the whistleblowing cases that reduced the volume of care. Patients who benefit from kyphoplasty were more likely to receive the procedure following the lawsuit, and those harmed were less likely to receive the procedure. One potential explanation is the change in incentives for providers, who before the lawsuit were more profit-motivated in their treatment, picking low-cost patients to receive procedures that could be heavily reimbursed, with less focus on the patient's expected health outcomes. Under this explanation, whistleblowing refocused provider attention on expected patient health outcomes, creating better targeting toward individuals who benefit the most.

Overall, in the case of kyphoplasty, whistleblowing seems to have had positive effects on patients by inducing better targeting of the procedure to those who benefit from it. This evidence indicates that kyphoplasty was overused in 2005, before the lawsuit, as evidenced by treatments performed on those expected to be harmed by the procedure. Whistleblowing enforcement was successful at reducing treatment to those individuals as well as increasing treatment to individuals likely to be helped by the procedure. In this case, the positive effects of whistleblowing went beyond financial benefits to the Medicare program, and indeed had substantial positive effects for patient care.

2.6 Conclusion

Whistleblowing under the False Claims Act is an effective way to both penalize and deter overbilling and fraud to the Medicare program. I undertake a set of case studies of the top categories of whistleblowing enforcement and measure direct deterrence effects. Settlements for the 4 categories of whistleblowing analyzed here recovered $1.9 billion in federal funds; I estimate that they generated more than $18 billion in direct deterrence effects. For comparison, the entire federal judiciary system had a budget of around $8 billion in 2018, and the Department of Justice spent $3.4 billion in 2018 on the entirety of its litigation and attorney costs. To first order, while unsuccessful cases put a burden on the federal courts and US attorneys, the recovery and deterrence effects of these few cases far exceed even the entire annual costs of running these departments. Furthermore, the deterrence effects for just these cases exceeds the total payment of $4.29 billion to whistleblowers for all healthcare-related cases in my data.

Changes in medical care induced by whistleblowers can have a variety of effects on patient care. In the example of kyphoplasty, I find beneficial changes to patient care, with better targeting to patients who are expected to benefit from this procedure. This case study motivates further analysis of the effects of whistleblowing on patient care. Whistleblowing generates changes to the care of patients that are potentially unrelated to the quality of the provider or the procedure, and this may provide experimental variation that other researchers find useful in the analysis of medical outcomes.

Whistleblowing has other potential costs and benefits not analyzed in this paper. The risk of litigation may cause providers to forgo misreporting in the first place, particularly when whistleblowers are empowered to directly sue for their own profit. These ex-ante deterrence effects are hard to measure without knowing the types of potential fraud that could have been committed. Conversely, increased compliance requirements impose costs on providers that are not measured here. FCA whistleblowing also incurs attorney expenses for both plaintiffs and defendants, for which little data are available. I do not impose the assumptions necessary here to compute net welfare. However, this paper estimates some of the parameters required for a broader efficiency computation, and motivates future work measuring other costs and benefits of the False Claims Act.

The results of this paper estimate the fiscal benefits of privatized enforcement as compared to the absence of such enforcement. However, a different counterfactual would be better public enforcement. Paying whistleblowers 15-30% of recovered funds is expensive if the government could produce similar recoveries without whistleblowing. Given the vast amount of data collected by the Medicare program, some of the effects of whistleblowing could likely be accomplished through machine learning, pattern detection, and automated audits. The fact that these programs are not yet in place may point to the limited capacity of the federal bureaucratic institutions. From this perspective, a major benefit of the False Claims Act is not just the information provided by the whistleblower, but also the profit motivation to conduct private enforcement.

Bibliography

31 U.S. Code Section 3731, "False Claims Procedure," 1986.

AAPC Coder, "Surgical Procedures on the Musculoskeletal System CPT Code range 20100-29999," https://coder.aapc.com/cpt-codes-range/230 2019. Accessed: September 13, 2019.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, 2010, *105* (490), 493–505.

_ **and Javier Gardeazabal**, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, March 2003, *93* (1), 113–132.

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 1963, *53* (5), 941–973.

Becker, Gary S. and George J. Stigler, "Law Enforcement, Malfeasance, and Compensation of Enforcers," *The Journal of Legal Studies*, 1974, *3* (1), 1–18.

Center for Medicare and Medicaid Services, "List of Diagnosis Related Groups (DRGS), FY 2005," https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/downloads/DRGdesc05.pdf 2005. Accessed: August 20, 2019.

Chamber of Commerce of The United States of America, Amicus Brief, State Farm Fire and Casualty Company v United States ex rel Cori Rigsby; Kerri Rigsby, 15-513 (Supreme Court 2015) 2015.

Civil Division, U.S. Department of Justice, "Fraud Statistics - Health And Human Services," https://www.justice.gov/civil/page/file/1080696/download 2018. Accessed: August 28, 2019.

Congressional Budget Office, "Health Care," https://www.cbo.gov/topics/health-care 2019. Accessed: August 19, 2019.

Dafny, Leemore S., "How Do Hospitals Respond to Price Changes?," *American Economic Review*, December 2005, *95* (5), 1525–1547.

Denaro, V., U. G. Longo, N. Maffulli, and L. Denaro, "Vertebroplasty and kyphoplasty," *Clinical Cases in Mineral and Bone Metabolism*, May 2009, *6* (2), 125–130.

Depoorter, Ben and Jef De Mot, "Supreme Court Economic Review," *Supreme Court Economic Review*, 2006, *14*, 135.

Dyck, Alexander, Adair Morse, and Luigi Zingales, "Who Blows the Whistle on Corporate Fraud?," *The Journal of Finance*, 2010, *65* (6), 2213–2253.

Engstrom, David Freeman, "Harnessing The Private Attorney General: Evidence From Qui Tam Litigation," *Columbia Law Review*, 2012, *112* (6), 1244–1325.

_ , "Public Regulation Of Private Enforcement: Empirical Analysis Of DOJ Oversight Of Qui Tam Litigation Under The False Claims Act," *Northwestern University Law Review*, 2013, *107* (4), 1689–1756.

Heese, Jonas, "The Role of Overbilling in Hospitals' Earnings Management Decisions," *European Accounting Review*, 2018, *27*.

_ **and Gerardo Perez Cavazos**, "Fraud Allegations and Government Contracting," *Journal of Accounting Research*, 2019, *Forthcoming.*

Kasper, D. M., "Kyphoplasty," *Seminars in Interventional Radiology*, Jun 2010, *27* (2), 172–184.

Kurra, S., U. Metkar, I. H. Lieberman, and W. F. Lavelle, "The Effect of Kyphoplasty on Mortality in Symptomatic Vertebral Compression Fractures: A Review," *International Journal of Spine Surgery*, Oct 2018, *12* (5), 543–548.

Kwok, David, "Evidence From The False Claims Act: Does Private Enforcement Attract Excessive Litigation?," *Public Contract Law Journal*, 2013, *42* (2), 225–249.

Nicholas, Lauren Hersch, Caroline Hanson, Jodi B. Segal, and Matthew D. Eisenberg, "Association Between Treatment by Fraud and Abuse Perpetrators and Health Outcomes Among Medicare Beneficiaries," *JAMA Internal Medicine*, 10 2019.

Polinsky, A. Mitchell, "Private versus Public Enforcement of Fines," *The Journal of Legal Studies*, 1980, *9* (1), 105–127.

Rawlings, R. Brent and Hugh E. Aaron, "The Effect of Hospital Charges on Outlier Payments under Medicare's Inpatient Prospective Payment System: Prudent Financial Management or Illegal Conduct?," *Annals of Health Law*, 2005, *14* (2), 267–328.

Silverman, Elaine and Jonathan Robert. Skinner, "Medicare upcoding and hospital ownership.," *Journal of health economics*, 2004, *23 2*, 369–89.

Singer, Natasha, "Botox Shots Approved for Migraine," *The New York Times*, Oct 2010.

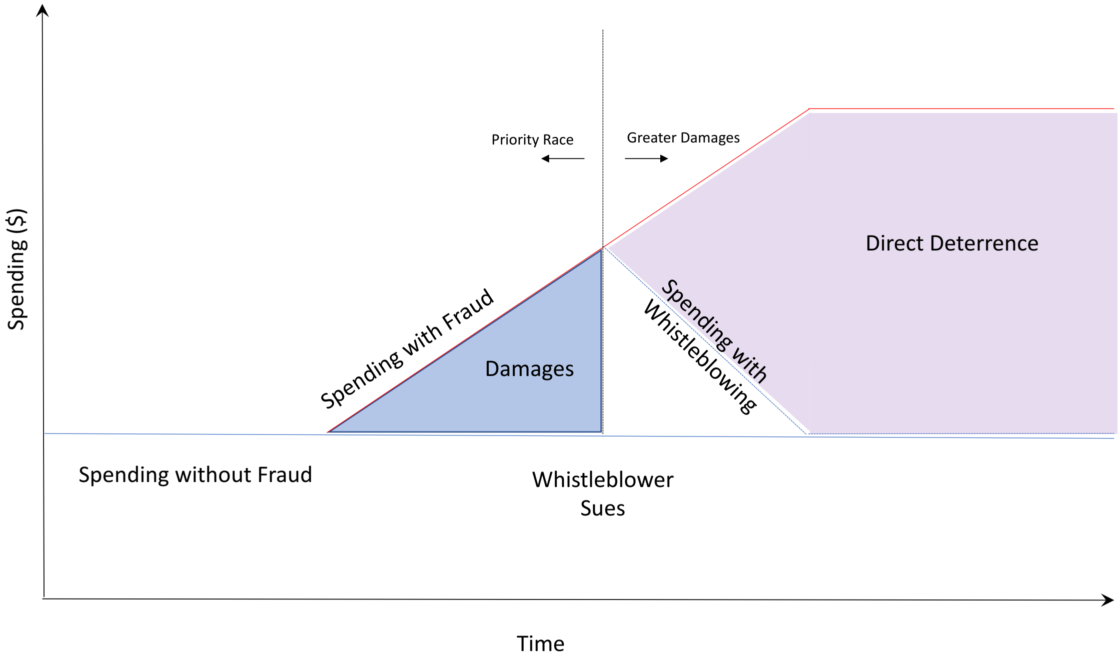
United States Department of Justice, "Justice Department Celebrates 25th Anniversary of False Claims Act Amendments of 1986," https://www.justice.gov/opa/pr/justice-department-celebrates-25th-anniversary-false-claims-act-amendments-1986 January 2012. Accessed: August 19, 2019.

United States District Court, D.C., US ex rel. Rockefeller v. Westinghouse Electric Co, 274 F.Supp.2d 10 (D.DC 2003) 2003.

United States Government Accountability Office, "Report to Congressional Committees: High Risk Series. Substantial Efforts Needed to Achieve Greater Progress on High-Risk Areas," March 2019, *GAO-19-157SP.*

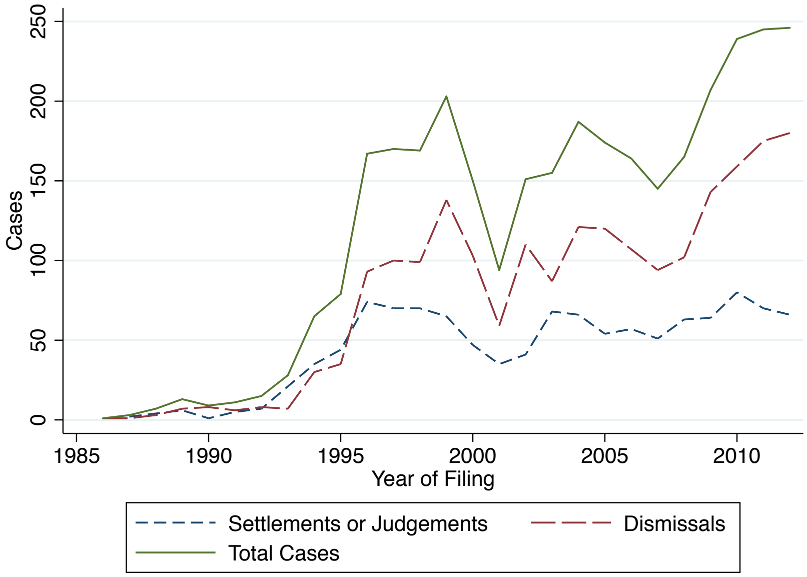
Figures and Tables

Figure 2-1: The Theoretical Effects of Whistleblowing



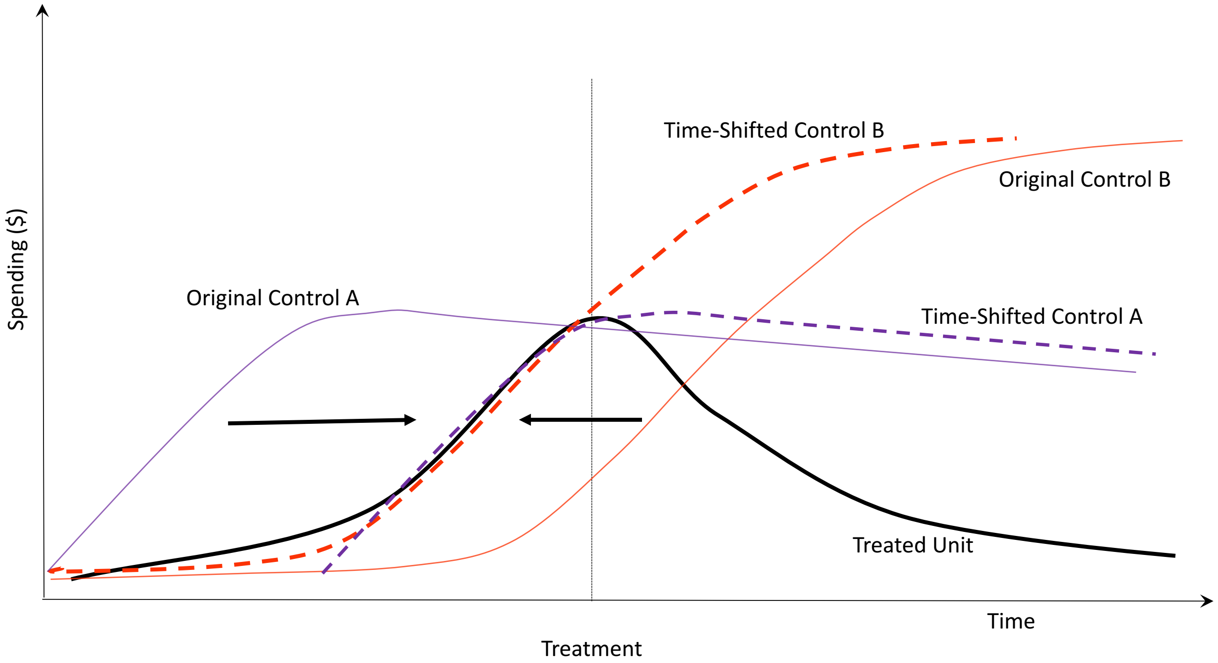
Notes: This figure describes the theoretical effects of a successful whistleblowing case on federal spending. When fraud is committed, the government has damages that are the difference between spending with fraud and the counterfactual spending without fraud. After the whistleblower sues, spending decreases back to its pre-fraud levels. The direct deterrence effect is the difference between how much would have been spent without whistleblowing and how much is spent after whistleblowing occurs. Because whistleblowers are paid proportionally to the damages, they have incentives to blow the whistle later and allow the damages to accumulate; however, because the first whistleblower to come forward receives greater compensation, they have countervailing incentives to file as soon as possible.

Figure 2-2: Trends in Healthcare Whistleblowing Cases



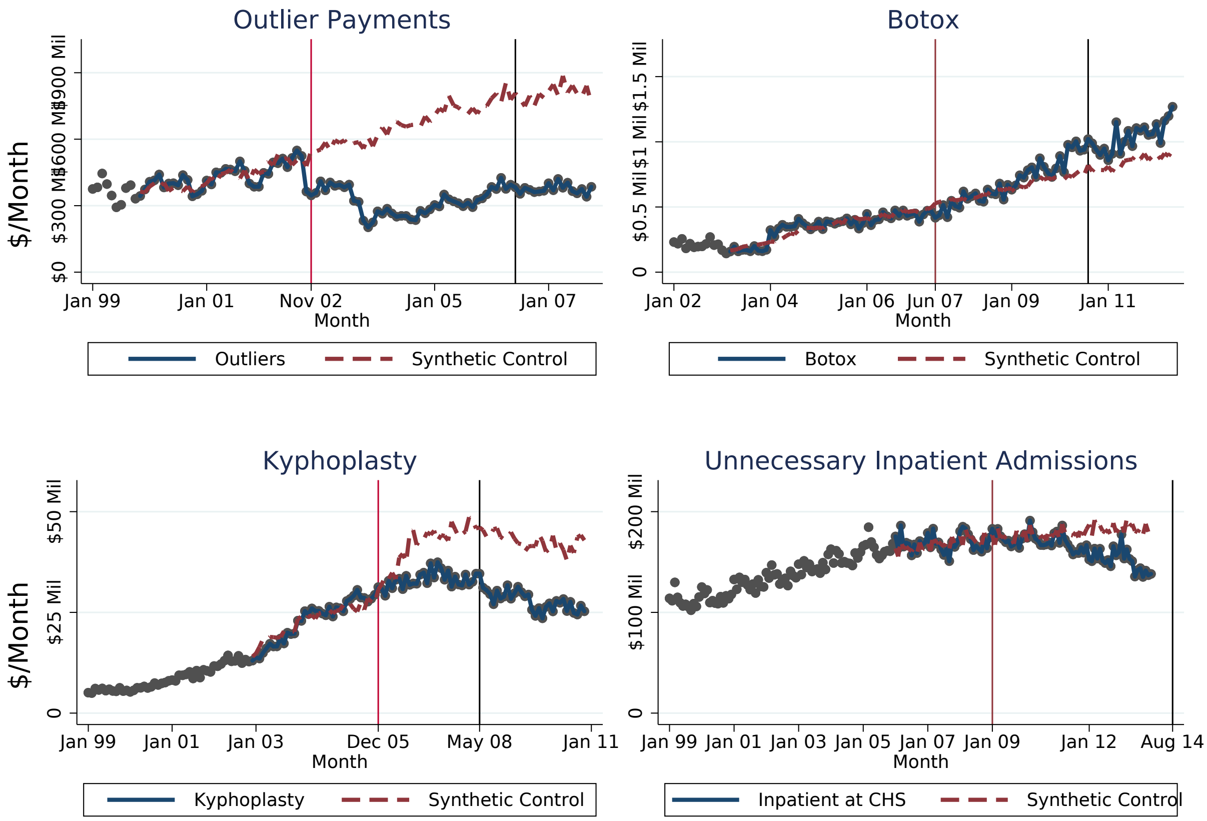
Notes: This figure plots the number of healthcare-related whistleblower lawsuits by year and splits the data by the outcome of the lawsuit. Data begin in 1986, when Congress amended the False Claims Act to allow for whistleblower lawsuits, and go through 2012, the last available year of data. Settlements rose to around 50 per year in 1995 and have stayed relatively constant, while total cases and dismissed cases have both continued to rise.

Figure 2-3: Example of Time-Shifts for Synthetic Controls



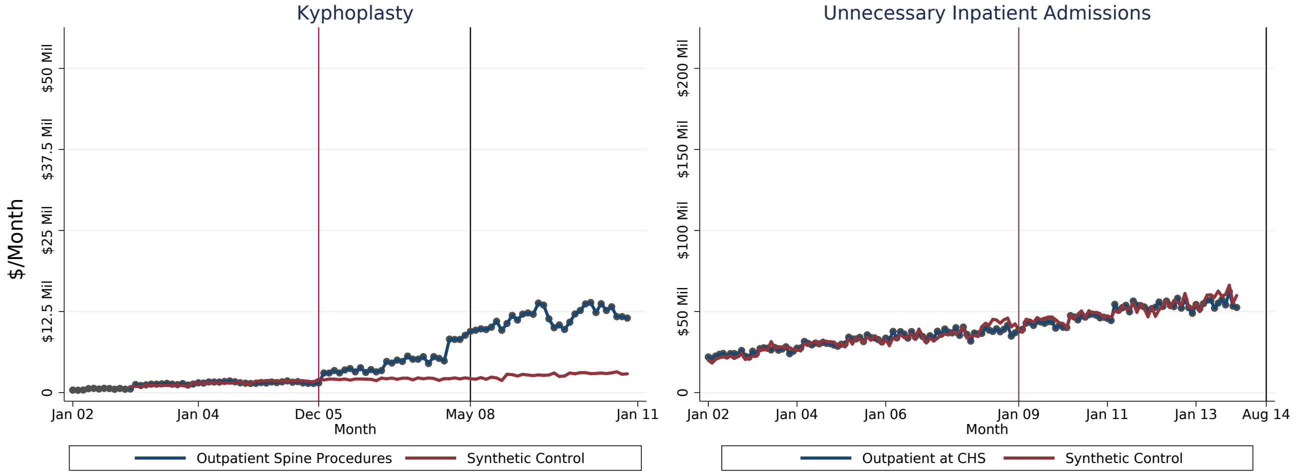
Notes: This figure exemplifies the time fitting process for time-shifted synthetic controls. Spending on the treated unit is a solid black line that increases pre-treatment and decreases post-treatment. Control A exhibits a similar rise to the pre-period, but at an earlier time, and is shifted forward. Control B exhibits a comparable rise at a later period, and is shifted backward. The shifts are picked to best approximate the pre-treatment period in both shape and levels. These fits are agnostic to how the controls develop in the post-treatment period; Control A falls while Control B continues to rise. Following these fits, a synthetic control unit can be constructed from Time-Shifted Control A and Time-Shifted Control B.

Figure 2-4: Effects of Whistleblowing on Medicare Expenditure



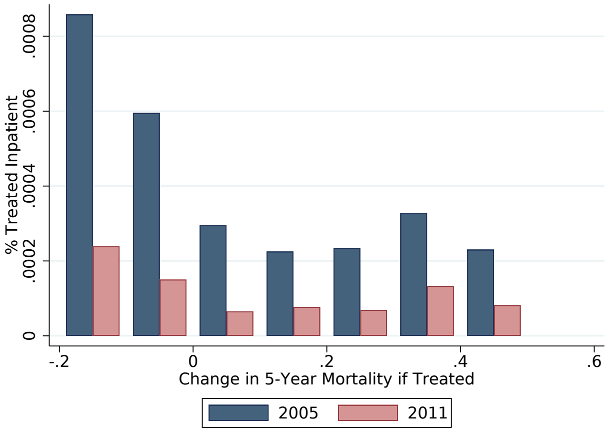
Notes: This figure plots the main effects of the 4 case studies: Outlier payments (top left), Botox (top right), kyphoplasty (bottom left), and unnecessary inpatient admissions (bottom right). For each case, the spending affected by whistleblowing is plotted in solid blue, while the synthetic control series is plotted in dashed red. The synthetic controls are produced using time-shifted control groups, and hence the period with which the synthetic control overlaps with the treated unit differs for each case. The grey dots represent the spending on the treated unit in the period before it overlaps with the synthetic control group. The first vertical line of each case represents the filing of the first related whistleblower lawsuit, which is used as the treatment date, and the second vertical line reflects the first settlement. Post-treatment effects are analyzed for 5 years after the treatment date. For the unnecessary inpatient admissions case (bottom right), multiple defendant providers were analyzed, and the series included here reflects Community Health Systems, the largest

Figure 2-5: Synthetic Controls for Substitute Outpatient Spending



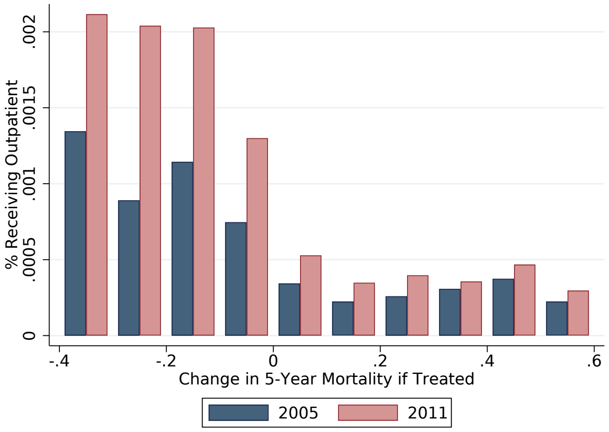
Notes: This figure plots the substitution effect to outpatient spending for the kyphoplasty (left) and unnecessary inpatient admission (right) case studies. These graphs correspond to the bottom half of Figure 2-4 and are scaled identically to those panels for comparison. In both lawsuits, whistleblowers alleged that patients should have been treated outpatient instead of inpatient. Outpatient spine procedure spending (left) rose following the kyphoplasty case as compared to the synthetic controls. However, there is no increase in outpatient spending at defendant hospitals (right) following the unnecessary admissions case. For the unnecessary inpatient admissions case, multiple defendant providers were analyzed, and the series included here reflects Community Health Systems, the largest defendant provider. Appendix Figure A4 plots the same figure for the other defendants in that case and shows heterogeneity in the outpatient effects, with both increases and decreases at different defendants.

Figure 2-6: Kyphoplasty Short-Stay Inpatient Treatment Probability by Estimated Treatment Effect



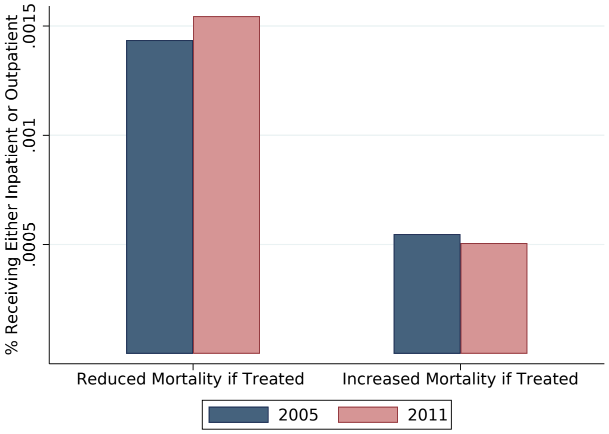
Notes: This figure plots the probability of receiving short stay inpatient kyphoplasty among the 2005 and 2011 cohorts, by the estimated treatment effect. Treatment effect is scaled as the difference in the probability of death in the next 6 years, and negative values correspond to a lower probability of dying. Absolute treatment probabilities are low, reflecting the inclusion of the full population in this analysis and the relative rarity of kyphoplasty. In both cohorts, patients with higher expected benefits are more likely to receive the treatment. The reduction in treatment probability occurs evenly across the treatment effect distribution.

Figure 2-7: Outpatient Treatment by Heterogeneous Treatment Effects



Notes: This figure plots the probability of receiving outpatient kyphoplasty or vertebroplasty by expected treatment effect among the 2005 and 2011 cohorts. The treatment effect is scaled as the change in probability of death in the next 6 years if one receives inpatient treatment; negative values indicate a lower probability of dying if treated. The whistleblower lawsuit settled in 2008 alleged that patients should have been treated outpatient instead of inpatient, and correspondingly, patients in 2011 were much more likely to receive outpatient treatment. These gains are greatest among patients to the left of the treatment effect distribution, which corresponds to the greatest benefits from the procedure.

Figure 2-8: Inpatient + Outpatient Treatment Probability by Health Benefit



Notes: This figure plots the probability of receiving either inpatient or substitute outpatient treatment by the estimated treatment effect, before and after the 2008 whistleblower settlement concerning unnecessary inpatient kyphoplasty. Patients in 2011 who are expected to benefit from the procedure were 7% more likely to be treated in 2011 than in 2005, and patients who are expected to be harmed were 7% less likely to be treated.

Table 2.1: Categories of Medicare Whistleblowing Enforcement

Type of Care	Type of Fraud	First Case Filed	First Settlement	# DOJ Press Releases	Settlement Total
Inpatient	Manipulation of Outlier Payments	Nov, 2002	Dec, 2004	11	$923,033,623
Botox	Off-Label Promotion	June, 2007	Aug, 2010	1	$600,000,000
Kyphoplasty	Inpatient Procedure Should be Outpatient	Dec, 2005	May, 2008	10	$214,238,775
Inpatient	Unnecessary Hospital Admissions	Nov, 2004	Dec, 2007	7	$172,296,460

Notes: This table shows the 4 highest settlement value categories of Medicare whistleblowing enforcement for which I have data. Categories are constructed using Department of Justice press releases to link cases with similar allegations. Cases filed before the start of the data are omitted, as are cases that concern allegations not potentially observable in the Medicare claims data. For each category, I conduct a case study of the effects of whistleblowing on spending on the related conduct. Appendix A.2 contains more details about the construction of categories and on the smaller categories of enforcement not studied here.

Table 2.2: Deterrence Effects of Major Whistleblowing Categories

Type of Care	Type of Fraud	Settlement Total	Direct Deterrence	Deterrence Ratio
Inpatient	Manipulation of Outlier Payments	$923 Million	$17.46 Billion	18.92
Botox	Off-Label Promotion	$600 Million	-$3.99 Million	- .006
Kyphoplasty	Inpatient Procedure Should be Outpatient	$214.2 Million	$281.1 Million	1.31
Inpatient	Unnecessary Hospital Admission	$172.3 Million	$1.124 Billion	6.52
	Total	**$1.91 Billion**	**$ 18.86 Billion**	
			Average Ratio	**6.69**

Notes: This table summarizes the results of case studies on the 4 largest categories of Medicare whistleblowing enforcement. Direct deterrence values are computed using a time-shifted synthetic control strategy to compare treated units to their counterfactual in the absence of whistleblowing. The direct deterrence is computed over 5 years post-treatment with a 10% annual discount rate compounded monthly. The deterrence ratio is computed as the ratio of the deterrence value to the settlement total.

Table 2.3: Placebo Tests for Synthetic Controls

Case	Deterrence Value	1-Tail Placebo Test
Outlier Payments	+$17.46 Billion	0.0 ($n = 5$)
Botox	-$-3.99 Million	0.149 ($n = 67$)

Case	Inpatient Deterrence	1-Tail Placebo Test	Outpatient Deterrence	1-Tail Placebo Test
Kyphoplasty	+ $538.9 Mil	0.133 ($n = 30$)	-$257.8 Mil	0.067 ($n = 15$)
Unnecessary Inpatient Admission:				
Defendant: St Joseph's Atlanta	+$44.8 Mil	0.01 ($n = 100$)	-$27.4 Mil	0.00 ($n = 100$)
Defendant: Wheaton Hospital	+$5.8 Mil	0.13 ($n = 100$)	-$83.8k	0.31 ($n = 100$)
Defendant: El Centro Medical Center	+$5.3 Mil	0.35 ($n = 100$)	-$4.0 Mil	0.02 ($n = 100$)
Defendant: Overlook Hospital	-$16.0 mil	0.02 ($n = 100$)	+$10.7 mil	0.08 ($n = 100$)
Defendant: Morton Plant Hospitals	+$266.6 Mil	0.01 ($n = 100$)	+$12.7 Mil	0.07 ($n = 100$)
Defendant: Shands Hospitals	+ $124.2 Mil	0.02 ($n = 100$)	+$50.7 Mi	0.00 ($n = 100$)
Defendant: Community Health Systems	+$693.2 Mil	0.07 ($n = 100$)	+$54.5 Mi	0.20 ($n = 100$)

Notes: This table summarizes the placebo test for the synthetic control strategy. For each control group, I compute the placebo deterrence effect, using the time-shifted synthetic control method with all other controls. The 1-tail test counts how many placebo groups exceed the deterrence value of the treated unit. For the kyphoplasty and unnecessary admissions cases, this test is conducted separately for the inpatient and outpatient spending. Deterrence effects are positive if spending on the treated unit is less than the control unit, and negative if spending on the treated unit is greater than the control unit.

Appendix

A.1 Cleaning of the FOIA Data on Qui Tam Whistleblower Suits

Data on the full set of whistleblowing lawsuits were gathered from a Freedom of Information Act (FOIA) request I conducted on the Department of Justice. For each lawsuit, the available data include: the docket number, district of filing, and case caption; the date the Attorney General was served notice of the suit; the primary federal agency to which the lawsuit related; whether or not the government intervened, and what date that election was made; the date of the settlement, judgement, or dismissal; the settlement amount if any, and the whistleblower's share. Each line of the FOIA dataset contains information about a suit that was dismissed, or in the event of a settlement, a settlement related to that suit. Lawsuits against multiple defendants can have more than one settlement, and therefore appear in more than one line of the data. To correct this issue, I collapse the data by docket, filing state, and year: if two lawsuits contain identical docket numbers and were filed within the same state and year, I assume they are a single suit, and create a total of their settlement values. For the descriptive statistics on medical-related lawsuits provided in Subsection 2.3.1, I restrict to suits for which the primary federal agency is either Health and Human Services, the Center for Medicare and Medicaid Services, or the Food and Drug Administration.

A.2 Constructing Categories of Enforcement from DOJ Press Releases

The FOIA data described in Appendix A.1 present a complete set of court-related information, but do not give information about the alleged behaviors for which the whistleblower sued, which are necessary for the case studies conducted here. To find details about the nature of these lawsuits, I scraped the Department of Justice press release archives for all press releases that contain the words "false claims," "Medicare," and either "qui tam" or "whistleblower." The DOJ makes an effort to publicize all of its successful cases, in particular because this strengthens later cases against providers who claim ignorance about what conduct constitutes an FCA violation.

From this universe of press releases, I created categories of enforcement against different types of improper conduct. First, I read and hand-coded all press releases through 2014, which contained 325 press-releases. The majority of press releases describe settlements; however, press releases occasionally describe government intervention in a case, or provide year-end totals of successful recoveries, and were discarded. Then, each settlement press release was coded for the type of medical care and the type of fraud it pertains to.

Certain types of care and certain types of fraud are not analyzable with my data and were omitted from the pool of potential cases to study. For example, cases regarding hospital cost reports, cases against Medicare claims processors, or cases that primarily concerned Medicare Advantage plans were discarded due to the lack of data. Similarly, some of the alleged frauds involve illegal kickbacks or improper financial relationships between providers. However, my available Medicare data do not contain financial information of providers, and therefore these types of cases were excluded from this study.

Following these restrictions, there are 170 remaining press releases that I group into categories. Press releases are grouped by the type of fraud and the type of care they describe, and within each case I create a total settlement amounts. For 3 of the largest settlements, all against groups of hospital providers, the settlement press releases describe multiple types of allegations relating to different categories of conduct reflected in other press releases. In these cases, the settlements were apportioned to the different categories of conduct, as described in the settlement agreement or press release. For example, the June 2006 Tenet Healthcare settlement (described in Appendix A.3) was a $900 million settlement, and the press release states that $788 million was for outlier payments and $46 million was for DRG upcoding. The outlier payment category therefore is apportioned $788 million from this press release and the DRG upcoding category gets $46 million.

The categorization process results in 54 distinct categories, most of which are very small. There are 23 categories with total settlements of less than $10 million and each contain 1 or 2 press releases. The top 11 categories detail more than $100 million in settlements each; these categories are described in Appendix Table A2. Within these categories, one final restriction was imposed, due to availability of data. If a lawsuit began before the data are available, I am unable to observe a pre-whistleblowing period, and therefore the case is omitted. In one category, hospice care, there is insufficient data in the court filings or within the public records to identify the defendant providers, and this case is omitted. Appendix Table A2 details the exclusion reasons for each of the top categories that were omitted, including the timing of the first lawsuit in circumstances where that drove the omission. Researchers with access to earlier data may be able to conduct similar analyses on these categories of whistleblowing.

The press release data do not contain sufficient detail to conduct analyses in the Medicare data, only to generally compare allegations. To augment the details of the press releases, I collect whistleblower complaints and settlement agreements from the lawsuits detailed by the press releases. The identification of these cases is done either by docket number, which the press release sometimes specifies, or by defendant name. The FOIA data described in Appendix Section A.1 were also used for mapping from press releases to court case docket numbers, which allowed for the retrieval of court documents. Whistleblower allegations and settlement documents contain specifics on the allegations of fraud or misconduct, including information on the medical coding of related procedures.

A.3 Lawsuit Details for Case Studies

Outlier Payment Case Study Details

Medicare reimburses most inpatient stays under a prospective payment system, with each stay classified under a Diagnosis Related Group (DRG). Hospitals are paid a fixed reimbursement for each DRG based on the average costs of treating patients under that DRG. This incentivizes providers to keep costs down, as they can recover profits by spending less per patient than the DRG pays. However, this contains the potential incentive to avoid treating high-cost patients. To correct this issue, Medicare has a system by which hospitals treating exceptionally high cost patients receive additional reimbursements called outlier payments. The gravamen of the accusations in the outlier payment lawsuits were that the defendants manipulated the reimbursement process for outlier payments to classify more patients as outliers and receive additional payments.

Between December, 2004 and March, 2010 the Department of Justice published 11 press releases detailing settlements related to outlier payment falsification. The outlier-related conduct from these press releases totals to $923 million in settlements. The first press release in this category was in December 2004, for the case U.S. ex rel James Devage et al. v. HealthSouth et al. This case was originally filed in 1998; however, looking at the court documents from this case, whistleblowing was only a portion of this settlement, and the allegation of outlier falsification was not alleged by the whistleblower. Rather, it appears the Department of Justice included a provision for outlier falsification in this settlement at a later date. The first whistleblower complaint alleging outlier falsification comes from U.S. ex rel. [Under Seal] v. Tenet Healthcare Corporation. et al., Case No. 02-8309, (E.D. Pa.). The filing of the Tenet Case, November 4, 2002, is used as the treatment date for this case. This case settled in June 2006, and was followed immediately by a Department of Justice press release. The Tenet settlement contains $788 million of recovery for outlier falsification, the bulk of the settlement total for this category.

Outlier data were gathered from the 100% Medpar files, which detail each inpatient stay paid for by Medicare, from 1999-2016. There are more than 5 million total stays classified as cost outliers in this period, and at its peak usage in 2002 (pre-whistleblowing), outlier payments exceeded $500 million per month. The outlier payment system also theoretically contained a provision for outpatient outlier payments. However, in practice there are almost no outlier payments listed in the outpatient claims files, even at the height of inpatient outlier spending. This analysis is therefore restricted to inpatient cost outliers.

The control groups for the Outlier payment case are other categories of expenditure that are of similar size and nature to outlier payments. Medicare pays for durable medical equipment (DME), home health aide services (HHA), hospice care (HOS), and skilled nursing facilities (SNF) as part of its broader package of benefits for older Americans. Spending on each of these types of care are included in the pool of potential controls. Furthermore, Medicare has a system for compensating hospitals who provide services to a disproportionate share of low income patients, called disproportionate share hospital (DSH) adjustments. Much like outlier payments, DSH payments are an adjustment above regular inpatient DRG pricing.

Table A3 details the time shifts (in months) and synthetic control weights for these control groups in constructing a synthetic control unit. The synthetic control method places greatest weight to DSH payments, which are the most similar in nature to outlier payments.

Botox Case Study Details

The whistleblower lawsuits against Botox alleged that Botox was prescribed for non-FDA approved, non-Medicare-reimbursable uses. The whistleblowers further allege that Allergan, the maker of Botox, explicitly promoted the product for these “off-label” uses, giving Allergan civil liability for the False Claims made to the Medicare and Medicaid programs. In September

2010, Allergan settled with the Department of Justice to resolve 3 pending whistleblower lawsuits of the same accusations: these cases have federal court docket numbers 1:07-cv-1288, 1:08-cv-1883, and 1:09-cv-3434, all conducted in the Northern District of Georgia. The first case was filed on June 5, 2007, which is used as the treatment date for this case. As part of this settlement, Botox agreed to pay $600 Million to the federal government, which includes both a civil settlement and a criminal penalty, for which whistleblowers received $37.8 million. This settlement was described in a Department of Justice press release in September, 2010.

Botox injections are outpatient procedures. Outpatient treatments are given a Current Procedural Terminology (CPT) or Healthcare Common Procedure Coding System (HCPCS) code that determines the reimbursement for the procedure, and an ICD-9 diagnosis code for the condition being treated. Documents from the whistleblower lawsuits provide details on the coding of outpatient Botox procedures. Medicare allowed reimbursement for Botox injections coded under CPT/HCPCS codes 64612, 64613, 64614, 64640, 64650, 67345, or J0585. The settlement agreement specifies that it resolves liability for false claims under ICD-9 diagnosis codes for spasm of muscle (728.85), other facial nerve disorders (351.8), spasmodic torticollis (333.83), unspecified torticollis (723.5), and bladder conditions (788.30 through 788.34, and 599.82).

Botox spending data were compiled from 100% samples of outpatient claims from January 2002-September 2015, using the CPT codes listed above and filtered for claims where the principal diagnosis matched the ICD-9 codes specified in the settlement. Data start at 2002 due to the availability of cleaned outpatient files, and data are truncated from October 2015 onwards due to the change from ICD-9 to ICD-10 diagnosis codes. Spending for Botox under the relevant diagnoses codes grew from 2 million dollars in 2003 to more than $5 million in 2006, the year before the lawsuit against Allergan was filed.

There are thousands of CPT/HCPCS codes, motivating a restriction of these groups to better potential controls. The candidate groups for this study are all other outpatient CPT/HCPCS codes for which spending started between \$2 million and \$5 million and saw a 2-3x rise over any 3-year period between 2002 and 2011, of which there are 67 control units. Table A4 shows the weights and time shifts for the 10 control groups given the highest weights by the synthetic control method.

Kyphoplasty Case Study Details

The main allegation of the kyphoplasty lawsuits was that hospital providers, at the urging of the product manufacturer Kyphon, were conducting kyphoplasty as an inpatient procedure, rather than outpatient. Under Medicare, inpatient stays are paid a fixed amount for the Diagnosis Related Groups (DRG) under which a patient is coded. Therefore, for short stays, providers can receive the full reimbursement and incur relatively low costs. Kyphon allegedly instructed its sales representatives and marketers to push usage of the DRGs 233, 234, and 216, which are various non-specific inpatient spine surgery codes, not designed for kyphoplasty. The specific descriptors of these DRGs were, in 2005, the year the lawsuit was filed: DRG 234: “Other musculoskeletal system & connective tissue O.R. procedure without comorbidities and complications”; DRG 233, ibid, “... with comorbidities and complications”; and DRG 216: “biopsies of musculoskeletal system & connective tissue.” (Center for Medicare and Medicaid Services, 2005)

Tracing spending on DRGs across time requires cross-walking when new versions of the DRG coding are released. This occurred twice in the relevant time period, in October 2007 and in October 2015. The October 2007 change was a complete overhaul of the DRG system, and changed from DRGs to a severity-based system (called MS-DRGs). Under this change, sets of 1 to 2 DRGs before October 1, 2007 usually correspond to 3 DRGs after that date. No 1 to 1 crosswalk exists, and so I collapse the DRGs into groups which can be cross-walked through this change. The DRGs allegedly promoted by Kyphon exhibit this pattern: DRG

216 became MS-DRGs 477, 478, and 479, and DRGs 233-234 became MS-DRGs 515, 516, and 517. I create groups for the DRGs that map across this change, and these DRG groups provide the control units. I omit DRGs that were entirely eliminated or newly generated during this switchover, as they cannot be analyze across the relevant time period. To handle the second DRG coding change in 2015, data from the 2016 fiscal year (starting October 1, 2015) are dropped and no crosswalk is implemented. This second change is close to the end of the available data and happened many years after the relevant lawsuit, so these are not necessary for analysis.

The treated unit for this analysis is the total payment for stays of 7 nights or fewer under the groups corresponding to DRGs 233, 234, and 216, the DRGs allegedly promoted by Kyphon. The set of controls are payments for stays of 7 nights or fewer under other DRG groups. I include DRG groups which experienced a more than double growth in annual spending over any 3-year period before 2011. The restriction to growing groups picks DRG groups on similar trajectories to the treated unit, which experienced a 2.5 times increase between 2002 and 2004, the year before the lawsuit was filed. The cutoff for growing controls is placed at 2011 to ensure that the data can be shifted back to match the Kyphoplasty series and still allow for 5 years of post-treatment comparison, as my data end in 2015. I exclude DRGs which saw discontinuous jumps (a 500%+ increase in any single month), or which were not in use for at least 12 months of the pre-whistleblowing period. There are 30 DRG groups included as controls. Appendix Table A5 details the time shift and synthetic control weights for these DRGs.

The kyphoplasty lawsuits alleged that kyphoplasty should have been coded as an outpatient procedure rather than inpatient. Outpatient procedures are billed to Medicare under HCPCS codes. Kyphoplasty was a new technology during this period, and coding for it changed over the course of the relevant period. Kyphoplasty was often billed under the catch-all unlisted spine procedure code 22899, but also was coded under the HCPCS codes 22523, 22524, 22525, 22513, 22514, 22515, C9718, or C9719 at various times, the latter two very infrequently. Furthermore, to measure substitution effects to outpatient procedures, I need to consider spending on vertebroplasty, a close substitute procedure, which was coded under HCPCS codes 22520, 22521, 22522, 22510, 22511, or 22512.

For the purposes of the health analysis in Section 2.5, the codes listed in the previous paragraph are used to identify outpatient kyphoplasty and vertebroplasty, as almost everything under these codes were in fact those procedures. However, whistleblowers also alleged that Kyphon, the maker of the kyphoplasty kit, also pushed providers to miscode the procedure under HCPCS codes 22327, 22325, 22328 for open reduction of thoracic or lumbar vertebrae. Kyphoplasty is not an open procedure, but is rather percutaneous. To analyze the sum of the fiscal effects, and to construct appropriate control groups, the outpatient deterrence analysis considers spending on all outpatient spine procedures, in the CPT code range 22010-22899. Some of these procedures were unaffected by whistleblowing, and therefore will difference out on average before and after the treatment period and will not bias the deterrence measurement. As controls, I consider other categories of surgical outpatient procedures on the musculoskeletal system, all of which are in the 20000-29999 range, of which the treated unit is a subset. These categories are constructed from the AAPC Coder code ranges (AAPC Coder, 2019) and include procedures like shoulder surgeries, hip surgeries, etc. and are not substitutes for the treated procedure. Two other codes in this range, CPT Codes 20000 and 20005, which correspond to surgical drainage procedures, were also included; these codes were deprecated in 2019. Table A5 gives the time shifts and weights for these control units.

Unnecessary Inpatient Admission Case Study Details

When a patient visits a hospital, particularly for emergency services, physicians at that hospital make a decision on whether to admit the patient for an inpatient stay, which generally results in an overnight stay of at least one night. Besides admitting patients, doctors have the ability to treat a patient outpatient, or to hold them for observation without admission. Inpatient admission receives greater reimbursement than outpatient or observational care. Under Medicare rules, inpatient stays are reserved for acute illnesses, and hospitals are expected to conduct utilization reviews to ensure that patients are admitted appropriately. The allegations in this category of whistleblowing are that the defendant hospitals improperly admitted Medicare patients because of the greater reimbursement provided.

Between 2007 and 2014, the Department of Justice issued press releases detailing 7 settlements with different providers and provider chains regarding this conduct. Four of the settlements concerned a single hospital: St Joseph's Atlanta; Wheaton Hospital in Wheaton, Minnesota; El Centro Medical Center in Southern California; and Overlook Medical Center in Summit, NJ. Two of the settlements concerned groups of 6 hospitals: Shands Hospitals and Morton Plant Hospitals, both in Florida. The final settlement was against Community Health Systems (CHS), described the Department of Justice in this press release as the "nation's largest operator of acute care hospitals." CHS settled for $98 million for conduct in 119 hospitals in 28 states. The total recovery from these 7 settlements was $172.29 million.

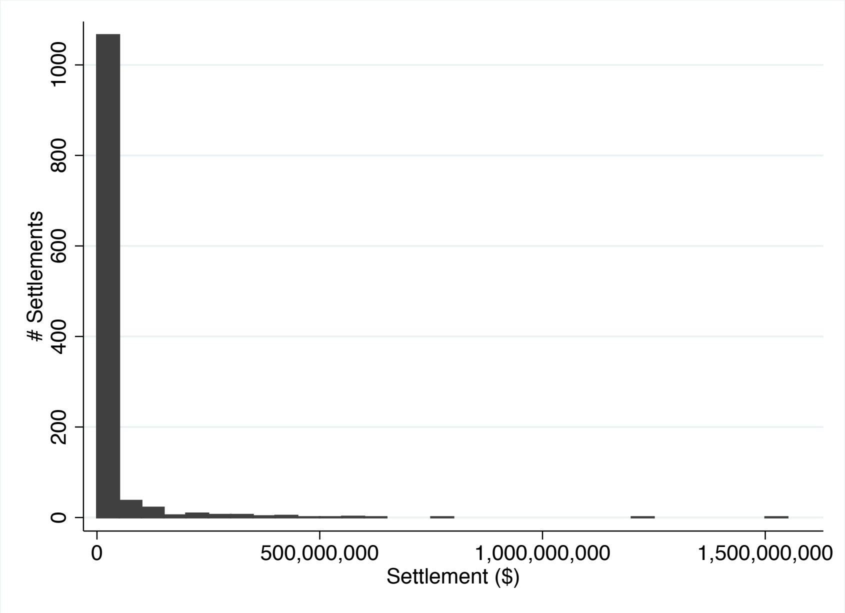
The evidence suggests that the conduct described in these cases was localized among the defendants. Appendix Figure A5 plots the total inpatient spending from all providers in the US and shows no changes with the filing of the first lawsuit in October 2004. This is unsurprising, as total Medicare inpatient spending was around $10 billion per month at the time of filing, and the entirety of these settlements was less than $200 million. Therefore, the computation of direct deterrence conducted here focuses only on the defendants. This may undercount spillover affects to other hospitals who were also deterred from unnecessary inpatient admissions as a result of these settlements.

The goal of this analysis is to measure the direct deterrence effects of these lawsuits on spending at the defendant providers. Because the lawsuits indicate that patients were unnecessarily admitted to the hospital rather than being seen outpatient, I expect a decrease in inpatient spending and an increase in outpatient spending. To measure this change, I construct control units using a set of untreated hospitals. Because some of the untreated hospitals may have been affected by spillovers, I restrict my control sample to hospitals in the 23 states (including the District of Columbia) with no defendant providers. These control units see different patient populations than the defendants and are less likely to be influenced by their behavior. This ensures the control units are isolated from the treated units, at least geographically, to mitigate spillover effects. Next, I construct a random sample of 100 control units for each defendant. For the four defendants that were 1 hospital, the control units are 100 randomly selected hospitals. For the two defendants which were 6 hospitals, the control units are 100 units of 6 randomly-grouped hospitals, drawn with replacement from the set of control hospitals. For CHS, which had 119 hospitals settle, I construct 100 control units of 119 randomly grouped hospitals, drawn with replacement from the set of control hospitals. These control units serve as the controls for the inpatient spending. For outpatient spending, I repeat the same process, drawing from the set of outpatient providers in states with no defendants.

Each of the 7 defendants here is conducted as its own case study. Each has its own controls, and the treatment date for each defendant is the earliest filing date of the lawsuit(s) settled in the settlement agreement with that hospital. Because CHS constitutes 119 of the 135 hospitals in this study, plots from CHS are included in the main results. Inpatient and outpatient plots from the other defendants are presented in Appendix Figures A3 and A4 respectively.

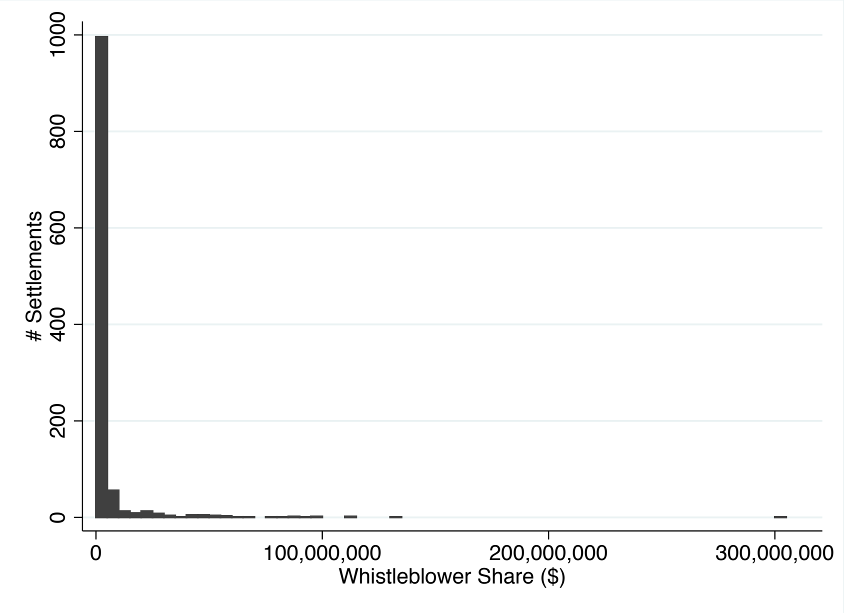
Appendix Figures

Figure A1: Histogram of Healthcare-Related False Claims Act Settlement Values



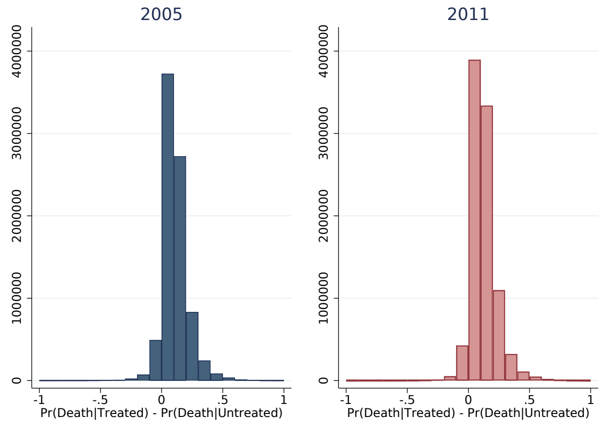
Notes: This figure plots the histogram of settlement values for settled False Claims Act whistleblower lawsuits related to healthcare from 1986-2012. Each bin is of width $50 million, and all cases included here have nonzero settlement values. Appendix A.1 describes the cleaning process for these data. The median settlement value is $1.5 million, and the maximum is $1.52 billion. The total of all settlements for healthcare-related cases is $26.4 billion.

Figure A2: Histogram of Healthcare-Related False Claims Act Whistleblower Shares



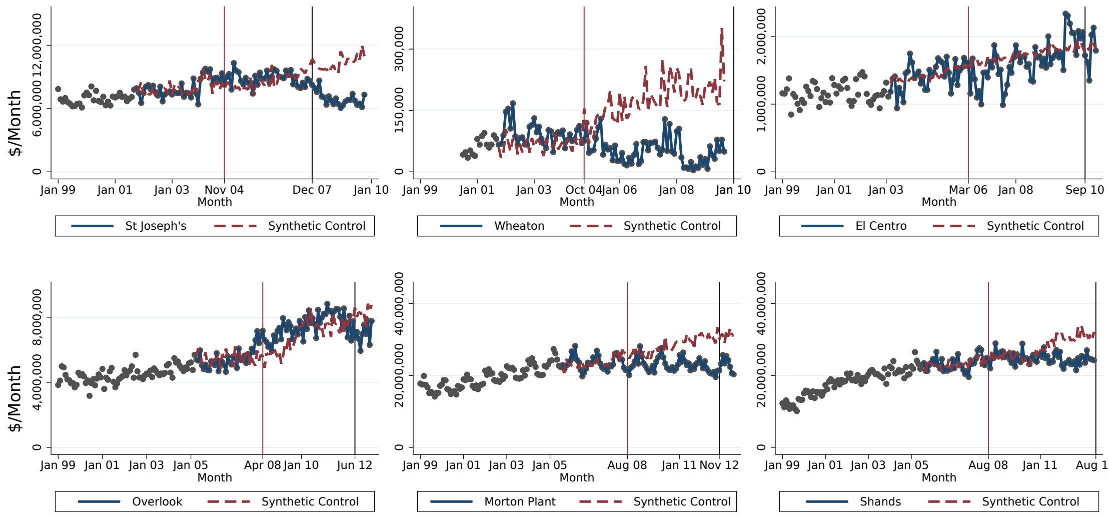
Notes: This figure plots the histogram of whistleblower shares from settled False Claims Act lawsuits related to healthcare from 1986-2012. Each bin is of width $5 million, and all cases included here have nonzero whistleblower share. Appendix A.1 describes the cleaning process for these data. The median whistleblower share is $250,307, and the maximum is $300.7 million. The total of all whistleblower payments for healthcare-related cases is $4.29 billion.

Figure A7: Treatment Effect Histogram by Cohort



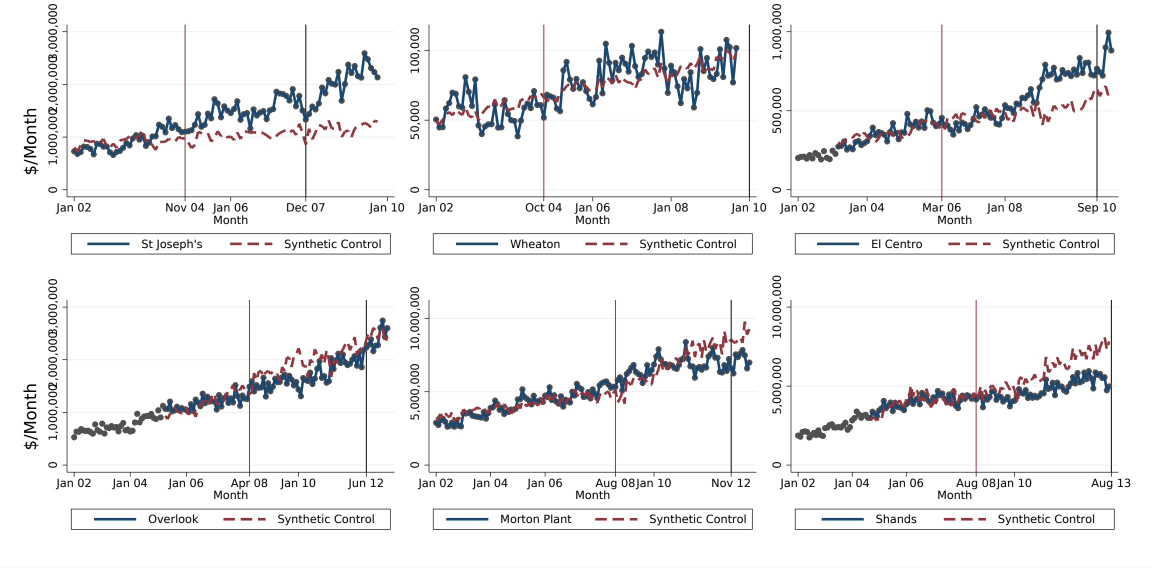
Notes: This figure plots the histogram of expected patient health effects from receiving a short-stay inpatient kyphoplasty treatment among the population of never-before-treated 70-75 year olds in 2005 and in 2011, which correspond to pre- and post-whistleblowing in the kyphoplasty case. Each cohort contains roughly 8 million patients. The horizontal axis is the difference in probability of death in the next 6 years if one receives treatment; values greater than 0 indicate a greater probability of death if treated, and negative values indicate a lower

Figure A3: Inpatient Spending at Other Defendants in the Unnecessary Admissions Case Study



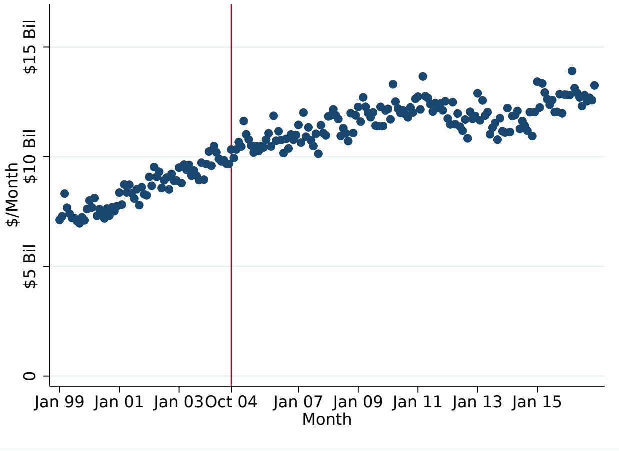
Notes: This figure plots the synthetic control strategy for inpatient spending at the other defendant providers in the unnecessary inpatient admissions case. The largest defendant, CHS, appears in the bottom-right panel of Figure 2-4. On average, inpatient spending at these providers fell relative to the synthetic control group.

Figure A4: Substitute Outpatient Spending at Other Defendants in the Unnecessary Admissions Case Study



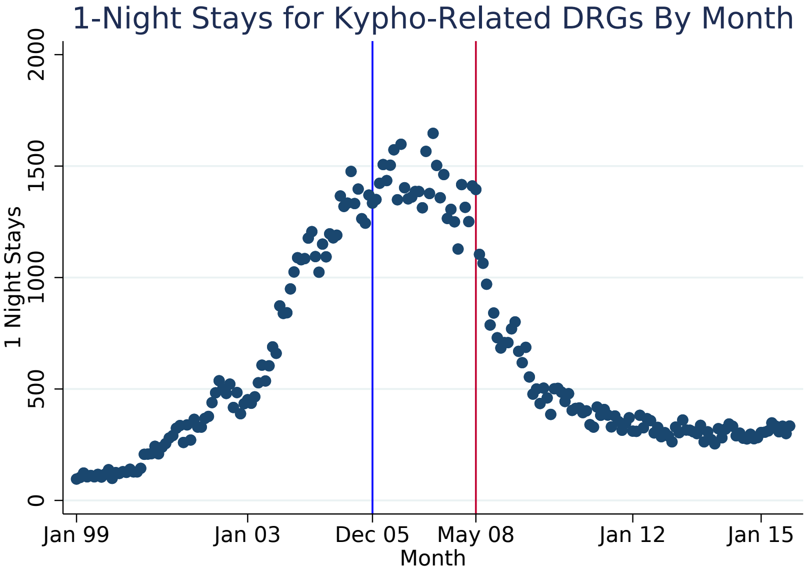
Notes: This figure plots the synthetic control strategy for outpatient spending at the other defendant providers in the unnecessary inpatient admissions case. The largest defendant, CHS, appears in the right panel of Figure 2-5. On average, outpatient spending at these providers did not increase, even when inpatient spending fell. However, there is heterogeneity among the defendants, with some experiencing increases in outpatient spending and others experiencing decreases.

Figure A5: Total Inpatient Spending Over Time



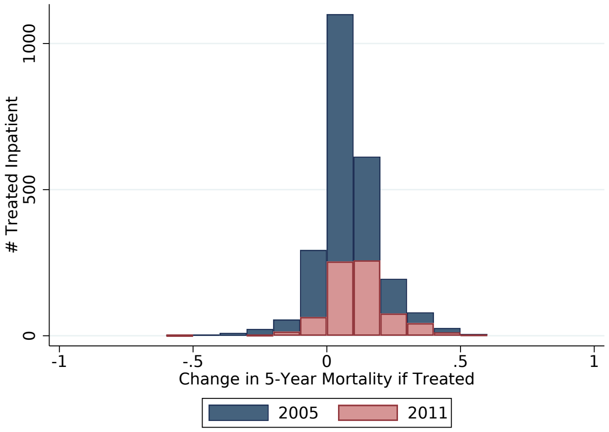
Notes: This figure plots total inpatient spending against the timing of the first unnecessary inpatient admissions lawsuit. There is no visible change in overall inpatient spending, which motivates an analysis focused on the defendants in these lawsuits.

Figure A6: One-Night Inpatient Kyphoplasty Claims



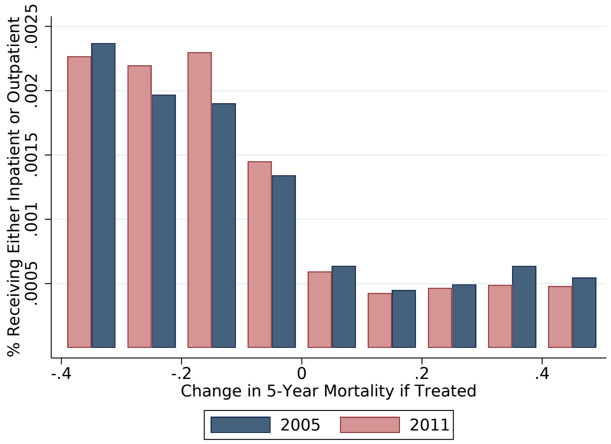
Notes: This figure plots inpatient stays for the DRGs promoted by Kyphon for inpatient kyphoplasty that lasted 1 night or less. The first vertical line shows the filing of the lawsuit, and the second line shows the settlement of the lawsuit.

Figure A8: Kyphoplasty Short-Stay Inpatient Treatment Count by Heterogeneous Treatment Effect



Notes: This figure plots the number of patients receiving short-stay inpatient kyphoplasty among the 2005 and 2011 cohorts of never-before-treated 70-75 year olds. Inpatient treatment counts were vastly reduced following the lawsuits against Kyphon and hospitals providing this treatment, which first settled in 2008. The treatment effect is identical to the horizontal axis in Figure A7, and is scaled as the change in probability of death when receiving treatment. The reduction in treatment volume occurs across the treatment effect distribution. The shape of these distributions is mostly driven by the number of units in each bin, as shown in Appendix Figure A7, motivating an analysis by probability of treatment as shown in Figure 2-6.

Figure A9: Inpatient or Outpatient Treatment Probability by Treatment Effect by Year



Notes: This figure plots the change in total (inpatient or outpatient) treatment probability as a function of the estimated treatment effect. It presents the same result as Figure 2-8, broken out by the treatment effect bin. Treatment effects are scaled as the change in probability of dying when receiving inpatient kyphoplasty. To satisfy Medicare cell-size-suppression rules, patients with treatment effects in the tails of the distribution are recoded to ±0.4. Patients with beneficial treatment effects, i.e. less than 0, are on average more likely to receive treatment after whistleblowing, while patients that are expected to be harmed are less likely to receive treatment after whistleblowing.

Table A1: Selected Logit Regression Coefficients for Heterogeneous Treatment Effects of Kyphoplasty

	Coef	SE	95% CI
Treated	2.876	2.863	[-2.736, 8.488]
Age	.0928	.000716	[.0914, .0942]
Treated × Age	-0.0117	0.0373	[-0.0848, 0.0614]
Female	-0.419	0.00222	[-0.423, -0.415]
Treated × Female	-0.193	0.139	[-0.465, 0.0782]
Race White	0.00245	0.0354	[-0.0670, 0.0719]
Treated × Race White	-1.35	0.679	[-2.68, -0.0220]
Race Black	0.0706	0.0356	[0.000847, 0.140]
Treated × Race Black	-1.32	0.811	[-2.91, 0.273]
OREC = DIB	0.526	0.00310	[0.520, 0.532]
Treated × OREC = DIB	-0.447	0.170	[-0.780, -0.116]
OREC = ESRD	0.558	0.0505	[0.459, 0.657]
Treated × OREC= ESRD	1.28	1.99	[-2.62, 5.19]
Previous Inpatient Stay	0.253	0.00283	[0.248, 0.259]
Treated × Previous Stay	-0.0688	0.134	[-0.331, 0.194]
Constant	-8.43	.0631	[-8.55, -8.31]

Notes: This table presents selected coefficients from the heterogeneous treatment effects regression described in Equation 2.6. The full model contains hundreds of coefficients due to the inclusion of state fixed effects and counts for stays under each inpatient DRG as well as full interaction with the treatment indicator. The coefficients presented here are given as examples. OREC indicates the original reason for Medicare qualification. ESRD denotes End Stage Renal Disease and DIB denotes disability insurance benefits.

Table A2: More Categories of Medicare Whistleblowing Enforcement

Type of Care	Type of Fraud	First Settlement Year	Settlement Total	Included or Omitted	Reason for Omission
Pharmaceuticals	Off-Label Promotion	2004	14,359,380,000	Omitted	Part D Data Start 2006
Inpatient	Outlier Payment Falsification	2004	923,033,623	Included	
Botox	Off-Label Promotion	2010	600,000,000	Included	
Inpatient	DRG Upcoding	2000	458,260,000	Omitted	Case Filed 1995
Home Health	Medically Unnecessary Care	2000	424,700,000	Omitted	Case Filed 1995
Nursing Home	Inadequate Care	2001	219,000,000	Omitted	Case Filed 1996
Kyphoplasty	Inpatient Should be Outpatient	2008	214,238,775	Included	
Physical Therapy	Unlicensed providers; Group Therapy Billed as One-on-One	2004	185,600,000	Omitted	Case Filed 1998
Hospital	Unnecessary Admissions	2007	172,296,460	Included	
Nursing Home Therapy	Falsified Hours Spent	2000	132,700,000	Omitted	Case Filed 1996
Hospice	Ineligible Patients	2006	114,886,000	Omitted	Defendants Not Identifiable from Court Data
Laboratory Tests	Medically Unnecessary; Unbundling Tests	1997	111,161,000	Omitted	Case Settled Before Data Start

Notes: This table describes the top categories of whistleblowing enforcement, as constructed from the Department of Justice press release data using the method described in Appendix A.2. Cases not related to Medicare claims data are excluded. These are all of the categories for which enforcement totaled to more than $100 million. Four of the top categories are included as case studies. Seven cases are omitted from the case study analysis of this paper because the first lawsuit was filed before the data are available. My available data start in 1999 for all categories except outpatient care and pharmaceuticals, which start in 2002 and 2006 respectively. One category, ineligible hospice patients, is omitted because the lawsuit documents do not identify the defendant providers.

Table A3: Synthetic Control Weights and Time Shifts for Outlier Payments Case

Control	Time Shift (Months)	Synthetic Control Weight
DME	+9	.049
DSH	+1	.837
HHA	+10	.024
HOS	-23	.083
SNF	+10	.007

Notes: This table details the synthetic control time shifts and weights used for the Kyphoplasty case. The control units are other types of Medicare spending, described in detail in Appendix A.3. The time shift describes the number of months the control unit must be shifted to align with the treated unit in the pre-whistleblowing period. Positive values mean the control unit is shifted forward in time, and negative months mean the control is shifted back in time. For example, a time shift of +9 means that the control unit in March, 2005 serves as a control for the treated unit in December, 2005.

Table A4: Synthetic Control Weights and Time Shifts for Botox Case

CPT Code	Descriptor	Time Shift (Months)	Synthetic Control Weight
85379	Pathology: Quantitative D-Dimer	-7	0.368
38510	Biopsy or excision of lymph node(s)	14	0.096
67028	Intravitreal injection of a pharmacologic agent	-24	0.06
82570	Pathology; Measurement of Creatinine	0	0.028
58563	Surgical hysteroscopy with electrosurgical ablation of endometrium	-24	0.022
63047	Laminectomy	2	0.019
37607	Ligation or banding of angioaccess arteriovenous fistula	2	0.015
73718	Magnetic resonance imaging, lower extremity other than joint; without contrast	-12	0.014
75635	Computed tomographic angiography with contrast	-10	0.014
43259	Esophagogastroduodenoscopy, flexible, transoral	-10	0.013

Notes: This table details the synthetic control time shifts and weights used for the Botox case. The control units are other types of outpatient care, described in detail in Appendix A.3. The time shift describes the number of months the control unit must be shifted to align with the treated unit in the pre-whistleblowing period. Positive values mean the control unit is shifted forward in time, and negative months mean the control is shifted back in time. For example, a time shift of +2 means that the control unit in October, 2005 serves as a control for the treated unit in December, 2005.

Table A5: Synthetic Control Weights and Time Shifts for Kyphoplasty Case

Inpatient

DRG V-24	MS-DRG V-25	Descriptor	Time Shift (Months)	Synthetic Control Weight
462	945, 946	Rehabilitation	47	0.431
533, 534	037, 038, 039	Extracranial Procedures	3	0.049
524	69	Transient Ischemia	5	0.045
518	250, 251	Percutaneous cardio procedures w/o coronary artery stent	17	0.045
535	222, 223	Cardiac defibrilator implant with cardiac catheterization	-5	0.037
519, 520	471, 472, 473	Cervical spinal fusion	-12	0.035
155, 156, 567, 568	326, 327, 328	Stomach, esophagealm and duodenal procedures	-58	0.03
515	226, 227	Cardiac defibrillator implant w/o cardiac catheterization	21	0.029
523	896, 897	Alcohol/drug abuse or dependence w/o rehabilitation therapy	-58	0.026
496	453, 454, 455	Combined anterior/posterior spinal fusion	-58	0.025

Outpatient

CPT Code Range	Surgical Category	Time Shift(Months)	Weight
20000, 20005	Incision and Drainage	0	0.158
22900-22999	Abdomen	0	0.115
21920-21936	Back or Flank	0	0.096
21501-21899	Neck or Thorax	0	0.079
26990-27299	Pelvis or Hip	0	0.077
21010-21499	Head	0	0.076
27301-27599	Femur or Knee	0	0.061
27600-27899	Leg or Ankle	0	0.058
29000-29799	Casts	11	0.051
25000-25999	Forearm or Wrist	11	0.048

Notes: This table details the synthetic control time shifts and weights used for the Kyphoplasty case. The top panel describes the controls for inpatient spending, which are groups of other inpatient DRGs. The bottom panel describes the controls for outpatient spending, which are other CPT code ranges of surgery on the musculoskeletal system. These controls are described in detail in Appendix A.3. The time shift describes the number of months the control unit must be shifted to align with the treated unit in the pre-whistleblowing period. Positive values mean the control unit is shifted forward in time, and negative months mean the control is shifted back in time. For example, a time shift of +3 means that the control unit in September, 2005 serves as a control for the treated unit in December, 2005.

Chapter 3

Maimonides Rule Redux

by Joshua D. Angrist, Victor Lavy, Jetson Leder-Luis and Adi Shany

Publication Note

This paper was published as: Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides' Rule Redux." American Economic Review: Insights, 1 (3): 309-24.

Appendix Details

The online appendix and original data are available at: DOI: 10.1257/aeri.20180120

Abstract

We use Maimonides Rule as an instrument for class size in large Israeli samples from 2002-2011. In contrast with Angrist and Lavy (1999), newer estimates show no evidence of class size effects. The new data also reveal enrollment manipulation near Maimonides cutoffs. A modified rule that uses birthdays to impute enrollment circumvents manipulation while still generating precisely estimated zeros. In both old and new data, Maimonides Rule is unrelated to socioeconomic characteristics conditional on a few controls. Enrollment manipulation therefore appears to be innocuous. We briefly discuss possible explanations for the change in class size effects since the early 1990s.

3.1 Introduction

The Maimonides Rule research design for estimation of class size effects exploits statutory limits on class size as a source of quasi-experimental variation. As first noted by Angrist and Lavy (1999), Israeli schools face a maximum class size of 40, so that, in principle, grade cohorts of 41 are split into two classes, while slightly smaller cohorts of 39 may be taught in one large class. This produces a distinctive sawtooth pattern in average class size as a function of grade-level enrollment, a pattern seen in Israeli data on enrollment and class size as well in data from school districts around the world.

Analyzing data on class average scores for the population of Israeli 4th and 5th graders tested in June 1991, Angrist and Lavy (1999) reported a substantial return to class size reductions – on the order of that found in a randomized evaluation of class size for US elementary grades (discussed by Krueger 1999). Many applications of the Maimonides Rule research design in other settings also report statistically significant learning gains in smaller classes (see, e.g, the Urquiola 2006 results for Bolivia). Other studies exploiting Maimonides Rule, however, find little evidence of achievement gains from rule-induced class size reductions (as in the Angrist et al. 2017 study of Italian schools).

This paper revisits the class size question for Israel with more recent data and a larger sample than that used in Angrist and Lavy (1999). Specifically, we look at a large sample of Israeli 5th graders tested between the school years ending spring 2002 and spring 2011. This update uncovers two findings. First, an econometric analysis paralleling that in Angrist and Lavy (1999) generates robust, precisely estimated zeros. Second, the new data reveal enrollment manipulation at Maimonides cutoffs: there are too many schools with enrollment values that produce an additional class.

Our investigation of enrollment patterns suggests a simple explanation for enrollment manipulation, and allows a straightforward remedy. A memo from Israeli Ministry of

Education (MOE) officials to school leaders cautions headmasters against attempts to increase staffing ratios through enrollment manipulation. In particular, schools are warned not to move students between grades or to enroll those overseas so as to produce an additional class. This reflects MOE concerns that school staff adjust enrollment (or enrollment statistics) close to cutoffs so as to produce smaller classes (e.g., by driving enrollment from 40 to 41, and thereby opening a second class). School leaders might care to do this because educators and parents prefer smaller classes. MOE rules that set school budgets as an increasing function of the number of classes also reward manipulation.

We address this problem by constructing an alternative version of Maimonides Rule that is largely unaffected by manipulation. The alternative rule pools students in 4th-6th grade and uses information on their birthdays to impute enrollment by applying the official birthday cutoff for 5th grade enrollment to a sample that includes all students in 4th-6th grade with birth dates that make them eligible for 5th grade. Imputed enrollment also generates a strong first stage for class size, but shows no evidence of sorting around birthday-based Maimonides cutoffs. Moreover, class size effects estimated using the statutory rule are also small, precisely estimated, and not significantly different from zero. Consistent with the absence of manipulation, Maimonides Rule constructed from imputed enrollment is unrelated to socioeconomic status.

Finally, we return to the 1991 data analyzed by Angrist and Lavy (1999). As first noted by Otsu et al. (2013), these data show evidence of sorting around the first Maimonides cutoff.[1] As in the more recent data, however, enrollment sorting in the original Maimonides sample does not appear to be highly consequential for class size effects. In particular, we show that the original formulation of the rule (constructed using November enrollment) is unrelated to students' socioeconomic status. More

[1] Figure 2 in Otsu et al. (2013) appears to exaggerate this; we discuss corrected estimates of the 1991 sorting pattern below.

recent data show small correlations between Maimonides Rule and socioeconomic status, but these disappear when estimated with a few school-level controls.

The birthday-based imputation used to eliminate enrollment sorting in recent data cannot be applied in the older data because birthdays and individual test scores are unavailable for the earlier period. But other simple corrections, such as a "donut" estimation strategy that discards observations near the first cutoff, leave the original results substantively unchanged.[2] The discrepancy between the old and new class size effects therefore seems more likely to be due to a change in the Israeli education production function rather than a sorting artifact. As we discuss in the conclusion, in light of the 2002-2011 results, the evidence for a large, externally valid class size effect in Angrist and Lavy (1999) also seems weaker in hindsight. It now seems especially noteworthy that estimates for a 1992 sample of 3rd graders reported in Angrist and Lavy (1999) show no evidence of achievement gains in smaller classes. Use of a more modern cluster adjustment in place of the parametric Moulton correction used in the original Maimonides Rule study also increases the uncertainty associated with the original estimates.

The next section reviews institutional background on the Israeli school system. We then document the Maimonides first stage in our more recent sample, explain how our birthday-based Maimonides instrument is constructed, and show that birthday-based imputed enrollment generates no evidence of running variable manipulation. Section 4 reports two-stage least squares (2SLS) estimates constructed using the two alternative Maimonides' instruments, and Section 5 looks again at the 1991 and 1992 samples. The conclusion considers possible explanations for changing class size effects.

[2] Barreca et al. (2011) appears to be the first to use the donut strategy to examine the consequences of sorting near regression discontinuity cutoffs.

3.2 Background and Context

3.2.1 Israeli Schools

Schooling in Israel is compulsory beginning in first grade, starting around age 6. Israeli students attend neighborhood schools, which serve catchment areas determined by a student's home address. Our analysis focuses on secular and religious students in Jewish public schools, the group that constitutes the bulk of public school enrollment. Public schools are administered by local authorities, but funded centrally by the MOE. Maimonides Rule, which caps class sizes at 40, has guided class assignment and school budgeting since 1969. The rule is well-known among school administrators and teachers. Most parents have few options by way of school choice other than to move. We therefore expect any manipulation of enrollment to reflect the behavior of teachers and school administrators rather than parents.

3.2.2 Related Work

Maimonides-style empirical strategies have been used to identify class size effects in many countries, including the US (Hoxby 2000), France (Piketty 2004 and Gary-Bobo and Mahjoub 2013), Norway (Bonesronning 2003 and Leuven et al. 2008), Bolivia (Urquiola 2006), and the Netherlands (Dobbelsteen et al. 2002). On balance, these results point to modest returns to class size reductions, though mostly smaller than those reported by Angrist and Lavy (1999) for Israel. A natural explanation for this difference in findings is the large size of Israeli elementary school classes. In line with this view, Woessmann (2005) finds a weak association between class size and achievement in a cross-country panel covering Western European school systems in which classes tend to be small. Recently published regression estimates for Israeli using 2006 and 2009 data show no evidence of a class size effect; this study also documents the vigorous debate over class size in Israel (Shafrir et al. 2016).[3]

A number of studies look at data manipulation and how this might compromise attempts to estimate causal class size effects. Urquiola and Verhoogen (2009) uncover evidence of sorting around Maimonides cutoffs in a sample from Chilean private schools. Angrist et al. (2017) show that estimates from Maimonides style experiments in southern Italy probably reflect increased manipulation of test scores by teachers in small classes. As noted above, Otsu, Xu, and Matsushita (2013) report evidence of sorting around the first Maimonides cutoff in the Angrist and Lavy (1999) sample; we return to this finding below. In related work, Jacob and Levitt (2003) document manipulation of test scores in Chicago public schools.

Methodological investigations of sorting in regression discontinuity (RD) running variables originate with McCrary (2008), who introduced the statistical test for running variable manipulation used here. Barreca et al. (2016) show that manipulation and nonrandom heaping of a running variable can bias RD estimates. Barreca et al. (2011) explore manipulation of the birthweight data used by Almond et al. (2010) to identify the causal effects of neonatal health care. Gerard et al. (2018) derive bounds on causal effects estimated using RD designs that are built on running variables compromised by sorting. Arai et al. (2018) introduce a test for validity of fuzzy regression discontinuity designs based on the joint distributions of treatment status and observed outcomes at cutoffs, applying this to the Angrist and Lavy (1999) sample. This test suggests manipulation of 1991 enrollment is not a source of bias in the original Angrist and Lavy estimates.

[3] Results in Sims (2008) suggest class size reductions obtained through combination classes have a negative effect on students' achievement.

3.3 Data and First Stage

3.3.1 Data and Descriptive Statistics

The test scores used in this study come from a national testing program known as Growth and Effectiveness Measures for Schools, or GEMS. Starting in 2002, fifth graders in half of Israeli schools were sampled for participation in GEMS (which also tests 8th graders). Tests are given in math, native language skills (Hebrew or Arabic), science and English. GEMS test scores are reported on a 0-100 scale, similar to the scale used in Angrist and Lavy 1999. Math scores average around 68, with standard deviation of about 11 for class average scores; language scores average around 72, with a standard deviation around 8 for class averages. Student-level standard deviations are roughly double the standard deviations of class means. These statistics appear in Appendix Table A1. The appendix also describes the GEMS data further.

Data on test scores were matched to administrative information describing schools, classes, and students. The unit of observation for most of our statistical analyses is the student. School records include information on the enrollment figures reported by headmasters to the MOE each November. This enrollment variable, henceforth called "November enrollment", is used by the MOE to determine school budgets. We also have data on class size collected at the end of the school year, in June. We refer to this variable as "June class size". Individual student characteristics in the file include gender, parents' education, number of siblings, and ethnicity. Schools in the GEMS samples are identified as secular or religious. Each school is also associated with an index of socioeconomic status (SES index).[4]

Our statistical analysis looks at fifth grade pupils in the Jewish public school

[4] The school SES index is an average of the index for its students. Student SES is a weighted average of values assigned to parents' schooling and income, economic status, immigrant status and former nationality, and the school's location (urban or peripheral). The index ranges from 1-10, with 1 representing the highest socioeconomic level. Schools with more disadvantaged students (high SES index) receive more funding per student. We observe only the school average SES.

system, including both secular and religious schools. The analysis excludes students in the special education system, who do not take GEMS tests. Our analysis covers data from 2002 through 2011 (2002 was the first year of the GEMS tests). In 2012, the MOE began implementing a national plan to reduce class size, rendering Maimonides' Rule less relevant (Vurgan 2011). We focus here on math and (Hebrew) language exam results.

The matched analysis file includes 240,310 fifth grade students from 8,823 classes. The data structure is a repeated cross-section; the sample of GEMS schools changes from year to year. Table A1, which reports descriptive statistics for classes, students, and schools in the estimation sample, shows that the mean and median elementary school class has about 28 pupils, and there are roughly 58 pupils and 2 classes per grade. Ten percent of classes have more than 35 pupils, and 10 percent have fewer than 21 pupils. Demographic data show that 90 percent of students are Israeli-born. Many in the sample are the children of immigrants; 16 percent are the children of immigrants from the former Soviet Union, for example.

3.3.2 The Maimonides First Stage

Maimonides' Rule reflects MOE regulations requiring that classes be split when they reach the statutory maximum of 40. Strict application of the rule produces class sizes that are a non-linear and discontinuous function of enrollment. Writing f_{jt} for the predicted 5th grade class size at school j in year t, we can write rule-based enrollment as

$$f_{jt} = \frac{r_{jt}}{\left[int\left(\left(r_{jt}-1\right)/40\right)+1\right]}, \tag{3.1}$$

where r_{jt} is the November enrollment of 5th graders at school j in year t, and $int(x)$ is the largest integer less than or equal to x.

Appendix Figure A1 plots actual average June class size and rule-based predictions,

f_{jt}, against November enrollment. Plotted points show the average June class size at each level of enrollment. The fit here looks similar to that reported using 1991 data in Angrist and Lavy (1999). Predicted discontinuities in the class size/enrollment relationship are also diminished by the fact that many classes are split before reaching the theoretical maximum of 40.

The first-stage effect of f_{jt} on class size is estimated by fitting

$$s_{ijt} = \pi f_{jt} + \rho_1 r_{jt} + \delta_1' X_{ijt} + \gamma_t + \varepsilon_{ijt} \tag{3.2}$$

where s_{ijt} is the June class size experienced by student i enrolled in school j and year t; X_{ijt} is a time-varying vector of student and school characteristics, f_{jt} is as defined above, and ε_{ijt} is a regression error term. The student characteristics in this model include a gender dummy, both parents' years of schooling, number of siblings, a born-in-Israel indicator and ethnic-origin indicators. School characteristics include an indicator for religious schools, the school SES index, and interactions of the SES index with dummies for the 2002-3 period and the 2008-11 period.[5] We also include year fixed effects (γ_t) and control for alternative functions of the running variable, r_{jt}.

Estimates of π in Equation (3.2) are remarkably stable at around 0.62. This can be seen in Appendix Table A2, which reports first stage estimates using a variety of running variable controls, including linear and quadratic functions of enrollment and the piecewise linear trend used by Angrist and Lavy (1999). This trend function picks up the slope on the linear segments of the rule. Specifically, the trend is defined on

[5] Interactions of the SES index with dummies for these two periods control for changes in the weights and the components of the index implemented in 2004 and 2008.

the interval $[0, 200]$ as follows:

$$\begin{array}{rl} r_{jt} & r_{jt} \in [0, 40] \\ 20 + r_{jt}/2 & r_{jt} \in [41, 80] \\ 100/3 + r_{jt}/3 & r_{jt} \in [81, 120] \\ 130/3 + r_{jt}/4 & r_{jt} \in [121, 160] \\ 154/3 + r_{jt}/5 & r_{jt} \in [161, 200] \end{array}$$

The constants here join the Maimonides linear segments at the cutoffs.

3.3.3 Sorting Out Enrollment Sorting

The budget for Israeli primary schools comes from local municipal authorities and the national MOE. The local authority funds administrative costs, while the MOE funds teaching and other educational activities. The MOE's budget for instruction time is based on the predicted number of classes determined by the November enrollment figures reported to the MOE (Ministry of Education 2015a). This generates an incentive to manipulate enrollment, either directly by moving students between grades, or through false reporting.[6]

As first noted by McCrary (2008), manipulation of a running variable may be revealed by discontinuities in the running variable distribution. Figure 3-1 plots the histogram of November enrollment in our 2002-11 sample. Vertical lines indicate Maimonides cutoffs. The figure shows a clear spike in enrollment just to the right of the cutoffs at 40 and 80, with apparent holes in the distribution to the left.

The forces producing these spikes are hinted at in MOE memoranda on enrollment reporting distributed at the end of the school year. These memoranda remind headmasters

[6] Funding rules for 2004-7 were revised so as to make total enrollment the major funding determinant rather than the number of classes but this reform was never fully implemented. In 2007, the MOE returned to the class-based funding rule (Lavy 2012; Vurgan 2007).

of the need for accurate enrollment reporting to determine funding. The 2015 circular also cautioned headmasters against enrollment manipulation. In particular, schools were warned not to move students between grades, to enroll a student in more than one school, or to enroll students residing overseas so as to produce an additional class. In 2016, the MOE began auditing enrollment data in an effort to prevent this type of manipulation, though sanctions are as yet undetermined (Ministry of Education, 2015b). Interestingly, Figure 3-1 offers further evidence of financially-motivated enrollment manipulation in the spike at a class size of 20. While budgetary rules set funding as a function of the number of classes, classes with enrollments below 20 are generally allotted half the regular funding.

Although the incentive for headmasters to push enrollment across Maimonides cutoffs seems clear, the question of whether this produces only misreporting or actual movement between grades is less easily addressed. Real enrollment changes can be accomplished by skipping students a grade ahead or through grade retention. A further likely channel is flexible age at entry for first graders. Although the official start age policy specifies a Chanukah-based birthday cutoff (detailed below), in practice, school headmasters have some discretion as to when children may start school.

Appendix Figure A2 suggests that at least some of the enrollment changes resulting from manipulation are real and persistent, rather than misreported. This figure plots the histogram of the number of 5th graders present for the GEMS tests in our sample. The evidence here is strongest for bunching around the first Maimonides cutoff, with somewhat weaker evidence of missing mass to the left of 80. Missing data for values below the second cutoff might be explained by the fact that roughly 10 percent of students enrolled miss the test.

Our primary concern is the possibility of selection bias resulting from enrollment manipulation. We might expect, for example, that more sophisticated school leaders understand the budgetary value of moving enrollment from just below to just beyond Maimonides cutoffs. And schools led by sophisticated leaders may also enroll higher-SES students, on average, producing a spurious achievement increase at the point where rule-based predicted class size drops.

We mitigate selection bias from enrollment manipulation by constructing a version of Maimonides Rule that uses birthday-based imputed enrollment in place of reported November enrollment. Israel's compulsory attendance laws specify rules for student enrollment in first grade according to whether a child's 6th Hebrew birthday falls before or after the last day of Chanukah. Students born after the last day of Chanukah are too young for first grade and must wait an additional year to start school. Most manipulation appears to result from single-grade retention or advancement relative to birthday-based enrollment, either as a result of delayed or accelerated school entry or advancement since first grade. Data on a sample of 4th, 5th, and 6th graders therefore includes almost all students who should be in 5th grade and can therefore be used to reconstruct the enrollment values that would be observed in a world where school officials follow official rules.

We apply the Chanukah-based birthday rule to June enrollment data for the sample of all 4th-6th graders in each school in the same year we see that school's 5th graders taking GEMS tests. This produces an imputed enrollment variable for 5th graders that is unlikely to reflect manipulation by school officials. Figure 3-2, which plots the imputed enrollment histogram, suggests that enrollment imputed in this manner is indeed manipulation-free. The figure shows a reasonably smooth distribution, with no evidence of spikes to the right of Maimonides cutoffs or at 20.

The McCrary (2008)-style density plots in Appendix Figure A3 are also consistent with the view that imputed birthday-based rule eliminates sorting in the November enrollment data. The upper panel of the figure plots empirical and fitted densities for November enrollment, allowing for a discontinuity at the first and the second

Maimonides cutoffs. Here, the jumps at 41 and 81 seem clear enough. By contrast, Panel B, which shows the same sort of plot for imputed enrollment, suggests the imputed enrollment distribution is smooth through these cutoffs.[7]

Appendix Table A3 reports estimates of the first stage regression of class size on Maimonides rule when the latter is computed using imputed birthday-based enrollment. These estimates are about half the size of those constructed using November enrollment. As when estimating November data, however, key first stage parameters are estimated precisely and largely insensitive to the nature of the running variable control.[8]

3.4 Class Size Effects: 2002-2011

Class size effects are estimated using a two-stage least squares (2SLS) setup that models y_{ijt}, the GEMS score of student i enrolled in 5th grade at school j in year t, as a function of 5th grade class size, running variable controls, year effects (μ_t), and additional controls, X_{ijt}. Second-stage models with a linear running variable control can be written:

$$y_{ijt} = \beta s_{ijt} + \rho_2 r_{jt} + \delta_2 X_{ijt} + \mu_t + \eta_{ijt}, \tag{3.3}$$

where β is the causal effect of interest and η_{ijt} is the random part of potential achievement. The first stage for 2SLS estimation of equation (3.3) is equation (3.2).

2SLS estimates of β in equation (3.3) suggest class size has no causal effect on achievement. Estimates of effects on language and math scores, reported in columns 2-4 and 6-8 of Table 3.1, range from -0.03 to 0.03 with standard errors around 0.03 to 0.04, and are not statistically different from 0. These reasonably precise zeros

[7] These plots use DCdensity (http://eml.berkeley.edu//~jmccrary/DCdensity/), which generates a graph of estimated densities with standard error bands, allowing for a single discontinuity, as described in McCrary (2008). Dots in the figure are histograms in an one-unit bin width.

[8] Appendix Figure A4 plots actual average June class size against birthday-based predicted enrollment, comparing the birthday-based first stage to the first stage constructed using actual November enrollment. Consistent with the smaller birthday-based first stage, actual class size follows birthday-based predictions less closely, with smoother size changes at Maimonides cutoffs. The non-linear and non-monotonic relationship between enrollment and class size remains.

contrast with the Angrist and Lavy (1999) estimates around -0.25. Interestingly, OLS estimates of a version of equation (3.3), reported in columns 1 and 5 of the table, are also small, though positive (indicating bigger classes improve test scores) and significant for math scores. The large precisely estimated negative SES effect reported in Table 3.1 implies that a 1 standard deviation increase in the school-wide SES index (lower SES) is associated with about 0.1 standard deviation lower language and math score. In estimates not reported in the table, we also see large ethnicity and parental school coefficients. This suggests that our dependent variables are informative measures of student achievement and bolsters the case for interpreting small insignificant class size effects as true zeroes.

It seems fair to say that the education production function identified by Maimonides Rule in more recent data differs markedly from that estimated using similar specifications for 1991 data. The earlier Angrist and Lavy (1999) results are replicated in Appendix Table A4, with the modification that the replication reports "Stata clustered" standard errors (clustered on school) rather than standard errors clustered using the Moulton formula as in Angrist and Lavy (1999). In contrast with the small effects found for 2002-2011, Maimonides Rule instruments in the 1991 sample, with linear running variable controls, generates an estimated effect of -0.277 for 5th grade language (with a standard error of 0.076) and an estimated effect of -0.231 for 5th grade math (with a standard error of 0.099).[9] Estimates for 4th graders are smaller; only that estimated for language with linear enrollment controls is (marginally) significantly different from zero.

Perhaps the new findings showing zero class size effects in recent data are an artifact of running variable manipulation. This possibility is explored in Table 3.2, which reports a set of 2SLS estimates paralleling those in Table 3.1, but computed in this case using version of Maimonides rule derived from birthday-based imputed

[9] As in Angrist and Lavy (1999), 1991 test scores are measured as a composite percentile, ranging from 0-100, with means around 70 and standard deviations of 8-10.

enrollment. Like the estimates in Table 3.1, the results in Table 3.2 show little evidence of achievement gains in smaller classes. In the 2002-2011 data, therefore, the lack of a class size effect appears unrelated to school leaders' efforts to open an additional class by pushing enrollment across Maimonides Rule cutoffs.

We also estimated models where the effect of class size on test scores is interacted with the SES index, thereby allowing for the possibility that class size matters most for disadvantaged students. The instruments in this case are f_{jt}, and $f_{jt} * SES_{jt}$, where SES_{jt} is the SES index for school j at year t. These results likewise show no evidence of class size effects or SES interactions. As can be seen in Appendix Figure A5, estimation of class size effects separately for each year also generates small, mixed positive and negative, and (with one exception), insignificant effects. This weighs against the hypothesis that the absence of a class size effect reflects extensive test preparation in more recent data, since Israeli media reports suggest test preparation efforts have intensified over time.

Gerard et al. (2018) note that sorting around RD cutoffs is innocuous when manipulated units are similar to those unaffected by sorting. To check for possible discontinuities in school characteristics induced by sorting, we regressed the school-by-year SES index (increasing from 1 to 10 as SES declines) on Maimonides rule in a version of equation (3.2) fit to school-year averages. Panel A of Appendix Table A5, reports these results when Maimonides Rule is constructed from November enrollment data, showing schools with larger predicted class size have somewhat higher SES. For example, the estimates in column 2 suggest that a 10 student increase in predicted class size is associated with a reduced disadvantaged index (that is, higher SES) of about 0.2. This seems like a modest change, amounting to less than one-tenth of a standard deviation of the index. The estimates in columns 4-6 of Table A5 show that this relationship disappears when Maimonides Rule is constructed using birthday-based imputed enrollment.

Although encouraging for the thesis that imputed enrollment data are uncompromised by systematic sorting, the results in Panel A of Table A5 suggest we might worry about non-random enrollment manipulation when working with November enrollment. But Panel B of the table shows that the association between November-based Maimonides Rule and SES disappears in models that control for a pair of school average covariates (fathers' schooling and family size), while these zeros are still precisely estimated. Moreover, Maimonides Rule computed using imputed enrollment is unrelated to SES with or without additional covariate controls. Since our findings on class size are consistent using both enrollment variables and when estimated with and without covariates, it seems unlikely that non-random sorting across Maimonides cutoffs in the November enrollment data is an important source of bias.[10]

3.5 Earlier Estimates Explored

The evidence of running variable manipulation in 2002-2011 data naturally raises questions about manipulation artifacts in the results reported in Angrist and Lavy (1999). Appendix Figure A6 plots estimated enrollment histograms and densities for the Angrist and Lavy samples of 4th and 5th graders tested in 1991. This figure shows evidence of a gap in the enrollment distribution below the first Maimonides cutoff of 41. The figure also reports estimates of the associated densities, allowing for a discontinuity at 41. Here too, we see evidence of a jump.[11] Appendix Figure A7 presents the enrollment histogram for the sample of 3rd graders tested in 1992; this figure shows a somewhat more modest enrollment jump to the right of the first cutoff.[12]

[10] 2SLS estimates of class size effects from models without covariates other than running variable controls are small and positive, marginally or not significantly different from zero.

[11] The discontinuity at 81 (the split from 2 to 3 classes) in the 1991 data is not statistically significant

[12] The discontinuity at 41 in the 1992 data is statistically significant; the discontinuity at 81 is not.

Otsu et al. (2013) includes figures similar to our Figure A6. These earlier plots, however, appear to count the 1991 enrollment distribution in terms of classes rather than schools. Because many grade cohorts are indeed split into additional classes at or near 40, the number of classes in schools with enrollments just above 40 jumps with or without sorting. The Otsu et al. (2013) discontinuity check therefore confounds the density discontinuity induced by sorting with the causal effect of Maimonides Rule on the number of classes. This concern notwithstanding, however, Figure A6 indeed shows evidence of sorting around the first Maimonides cutoff in 1991.

Additional analyses of the older data (not reported here) suggest sorting was less pervasive in 1991 and 1992, with little evidence of manipulation beyond the first Maimonides cutoff. Even so, in view of the discontinuity in the 1991 enrollment distribution seen in Figure A6, it's worth asking whether enrollment manipulation is likely to be a source of omitted variables bias in the older estimates. Table 3.3 therefore reports estimates from a regression of school-level SES on Maimonides Rule using 1991 data, similar to the estimates reported in Table A5. As in the more recent data (with covariates), we see little evidence of a relationship between Maimonides Rule and school-level SES. The negative associations estimated for 5th graders are not significantly different from zero, while the sign flips to (insignificant) positive for 4th and 3rd graders.[13]

The individual student data required for a birthday-based imputation of 1991 enrollment are unavailable. We turn therefore to an alternative check on the replicated results that omits observations near the first Maimonides cutoff.[14] The results of this further exploration of the consequences of sorting in 1991 are reported in Appendix Table A6. For example, the estimated class size effect of -0.234 in column 1 of Table A6 was computed using a sample omitting schools with 5th grade enrollments

[13] The 1991 SES index is scaled as "percent disadvantaged."

[14] Barecca et al. (2011) appear to be the first to propose this simple adjustment for sorting, sometimes referred to as an RD "donut".

between 39 and 41. This can be compared with the full-sample estimate of -0.277. Although somewhat less precise, the donut estimates in Table A6 differ little from those for the full sample estimates reported in Table A4.

3.6 Summary and Conclusion

The Maimonides Rule identification strategy for class size effects generates precisely estimated zeros in large Israeli samples for 2002-2011. These samples also show clear evidence of enrollment manipulation around Maimonides class size cutoffs, likely reflecting school leaders' desire to open an additional class when enrollment is close to a cutoff. Enrollment imputed using information on grade-eligible birthdates appears unaffected by manipulation, however, and 2SLS estimates derived from imputed enrollment instruments show similarly small class size effects. Maimonides Rule constructed using birthday-based imputed enrollment is also unrelated to a school-level measure of SES.

We find only weak evidence of *systematic* enrollment sorting: more recent data generate small estimated effects of the original Maimonides Rule on socioeconomic status, but these effects disappear after conditioning on a few covariates. The fact that estimated class size effects are similar whether Maimonides Rule is constructed using November or birthday-based enrollment further reinforces our conclusion that the finding of a null class size effect in recent data is not a manipulation artifact. The estimates of zero class size effect in more recent data contrast with the substantial negative class size effects reported by Angrist and Lavy (1999). We also see some evidence of manipulation around the first Maimonides cutoff in the older data analyzed by Angrist and Lavy (1999). But the absence of a relationship between Maimonides' Rule and school average SES, and results from a donut strategy that omits data near the cutoff, suggest these estimates too are unaffected by manipulation near cutoffs.

This conclusion is likewise supported by specification test results reported in Arai et al. (2018).

The disappearance of Israeli class size effects may reflect changes in the Israeli education production function. The fact that Israeli class size has fallen from a median of 31 in 1991 to 28 in more recent samples may be relevant. Yet Figure A5, which plots 2SLS estimates by year, shows no evidence of declining effects over the period 2002-2011. It may also be relevant that, since the early 2000's, some schools have hired additional teaching staff, a staffing increase funded mostly by parents in high SES schools (Vurgan, 2014). Weighing against the importance of these changes for class size estimates, our analysis fails to show significant class size/SES interactions or significant effects in earlier years.

We briefly explored changes in other inputs that might explain the absence of class size effects in recent data (these data are from an analysis reported in Blass et al. 2012). Regressions of total hours of instruction provided by school staff and others on predicted class size show small, marginally significant increases on the order of 0.5 percent for each additional student. We also see small, marginally significant increases in the share of class time going to small group instruction. Per-pupil spending however, falls about 2 percent for each additional student. In future work, we hope to be able to identify causal effects of these additional inputs, and better gauge their interaction with class size in education production.

It seems noteworthy that the 1991 estimates reported in Angrist and Lavy (1999) are strongest for 5th graders, but less impressive for 4th graders, for whom only estimates for language are significantly different from zero, and in only one specification. The original Angrist and Lavy study also reported zero class effects in a 1992 sample of 3rd graders, a result attributed in the original write-up to extensive test preparation and changes in testing protocols. These forces may be at work in the more recent GEMS data analyzed here as well. Some analysts have suggested schools are increasingly and effectively teaching to GEMS tests (e.g. Kliger, 2009). Here too, however, there's no smoking gun for mediating interactions: our analysis uncovers no changes in class size effects over time that might be linked to changes in test preparation.[15] On balance, it seems fair to say that the 1991 results are unusual in showing strong class size effects, while the null effects reported for 1992 have emerged as more representative of the causal relationship between class size and test scores in Israel.

[15] In view of the unusually early administration of tests in 2004-6, Appendix Table A7 reports estimates analogous to those in Tables 1 and 2, computed in a sample omitting data from 2004-6. This change in sample leaves the results unchanged.

Table 3.1: Class Size Effects Estimated Using November Enrollment Instruments (2002-2011)

	Language				Math			
	OLS	2SLS	2SLS	2SLS	OLS	2SLS	2SLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class size	0.0091	-0.0220	-0.0299	-0.0288	0.0492	0.0283	0.0158	0.0137
	(0.0173)	(0.0314)	(0.0329)	(0.0322)	(0.0229)	(0.0422)	(0.0436)	(0.0429)
SES index	-0.4268	-0.4312	-0.4281	-0.4286	-0.3660	-0.3688	-0.3640	-0.3635
	(0.0602)	(0.0602)	(0.0603)	(0.0603)	(0.0800)	(0.0801)	(0.0802)	(0.0801)
November enrollment	0.0025	0.0052	0.0226		-0.0008	0.0010	0.0285	
	(0.0037)	(0.0043)	(0.0144)		(0.0048)	(0.0058)	(0.0183)	
Enrollment squared/100			-0.0103				-0.0163	
			(0.0081)				(0.0101)	
Piecewise linear trend				0.0147				0.0100
				(0.0094)				(0.0128)
N	225,108				226,832			

Notes: This table reports OLS and 2SLS estimates of equation (3) in the text. The endogenous variable is June class size; Maimonides Rule is constructed using November enrollment. Standard errors reported in parentheses are clustered at the school and year level. The dependent variable is a math or language test score. Additional covariates include student characteristics (a gender dummy, parents' years of schooling, number of siblings, a born-in-Israel indicator, and ethnic-origin indicators), year fixed effects, an indicator for religious schools, socioeconomic index and interactions of the socioeconomic index with dummies for 2002-3 period and 2008-11 period. The reported SES coefficient is for 2004-7.

Table 3.2: Class Size Effects Estimated Using Birthday-based Imputed Enrollment (2002-2011)

	Language				Math			
	OLS	2SLS	2SLS	2SLS	OLS	2SLS	2SLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class size	0.0070 (0.0172)	0.0012 (0.0608)	-0.0089 (0.0666)	-0.0061 (0.0657)	0.0386 (0.0231)	-0.0366 (0.0814)	-0.0623 (0.0889)	-0.0616 (0.0878)
SES index	-0.4254 (0.0602)	-0.4263 (0.0610)	-0.4246 (0.0609)	-0.4252 (0.0609)	-0.3570 (0.0799)	-0.3680 (0.0809)	-0.3636 (0.0810)	-0.3644 (0.0809)
Birthday-based enrollment	0.0033 (0.0035)	0.0038 (0.0060)	0.0163 (0.0184)		0.0037 (0.0046)	0.0099 (0.0081)	0.0418 (0.0239)	
Enrollment squared/100			-0.0068 (0.0086)				-0.0173 (0.0109)	
Piecewise linear trend				0.0113 (0.0151)				0.0318 (0.0203)
N	225,108				226,832			

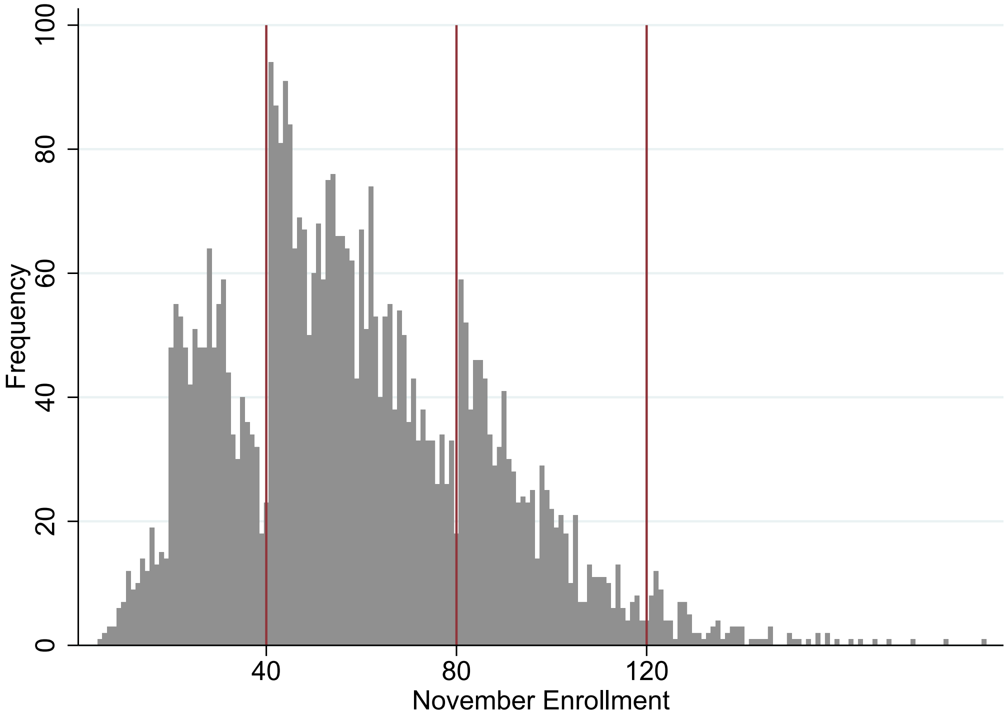
Notes: This table reports OLS and 2SLS estimates of equation (3). The endogenous variable is June class size; Maimonides Rule is constructed using birthday-based enrollment. Standard errors reported in parentheses are clustered at the school and year level. The dependent variable is a math or language test score. Additional covariates include student characteristics (a gender dummy, parents' years of schooling, number of siblings, a born-in-Israel indicator, and ethnic-origin indicators), year fixed effects, an indicator for religious schools, socioeconomic index and interactions of the socioeconomic index with dummies for 2002-3 period and 2008-11 period. The reported SES coefficient is for 2004-7.

Table 3.3: Maimonides Rule Effects on Socioeconomic Status in 1991 Data

	Fifth Grade			Fourth Grade			Third Grade		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
f_{jt}	-0.0592	-0.0686	-0.0709	0.0532	0.0636	0.0718	0.0400	0.0444	0.0581
	(0.0784)	(0.0848)	(0.0865)	(0.0810)	(0.0878)	(0.0894)	(0.0845)	(0.0906)	(0.0910)
November Enrollment	-0.0598	-0.0475		-0.0691	-0.0824		-0.0940	-0.101	
	(0.0152)	(0.0448)		(0.0155)	(0.0461)		(0.0165)	(0.0534)	
Enrollment Squared/100		-0.0068			0.0076			0.0041	
		(0.0235)			(0.0248)			(0.0307)	
Piecewise Linear Trend			-0.1010			-0.1332			-0.1796
			(0.0350)			(0.0349)			(0.0358)
N	1,002	1,002	990	1,013	1,013	1,003	989	989	982

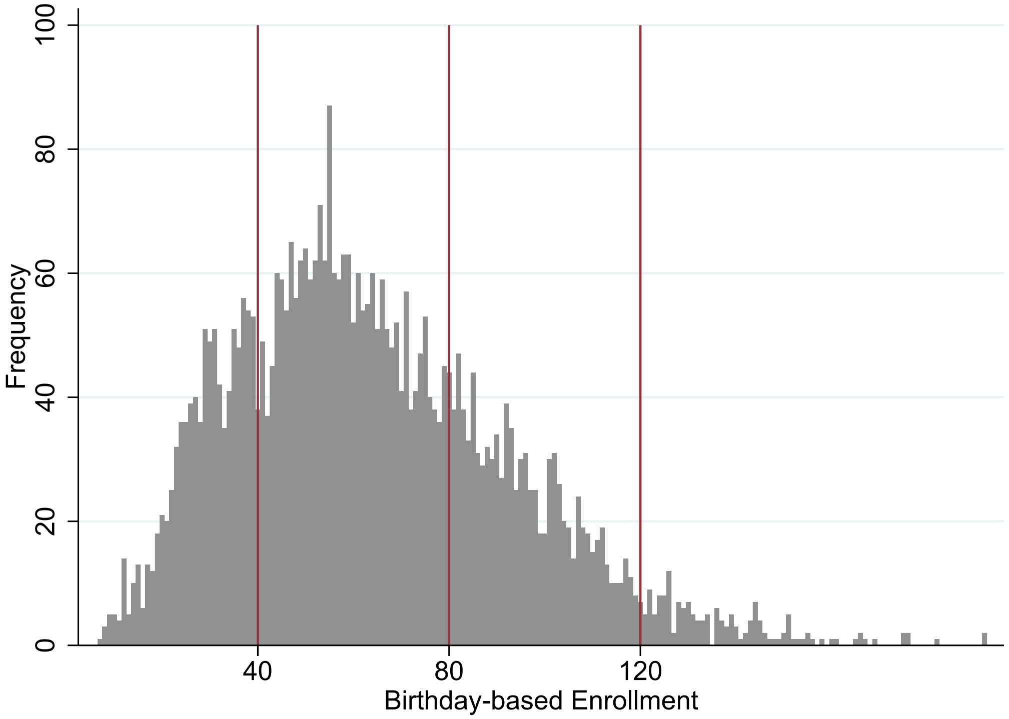
Notes: This table reports OLS estimates of the effect of Maimonides Rule on a school-level index of socioeconomic status. The unit of analysis is the school. The third grade sample is limited to schools appearing in the fourth and fifth grade sample. The piecewise linear control in columns (3), (6), and (9) omits enrollments above 160.

Figure 3-1: The 5th grade Enrollment Distribution Reported in November (2002-2011)



Notes: This figure plots the distribution of 5th grade enrollment as reported by school headmasters in November. Reference lines indicate Maimonides Rule cutoffs at which an additional class is added.

Figure 3-2: The 5th Grade Birthday-based Imputed Enrollment Distribution (2002-2011)



Notes: This figure plots the distribution of birthday-based imputed enrollment for 5th graders by school. Birthday-based imputed enrollment is computed from the birthday distribution of students enrolled in 4th-6th grade in June of each year. The birthday rule counts 4th-6th graders born between Chanukah 11 years before and Chanukah 10 years before the current school year. Reference lines indicate Maimonides Rule cutoffs at which an additional class is added.

Bibliography

Almond, Douglas, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams, "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns," *The Quarterly Journal of Economics*, 2010, *125* (2), 591–634.

Angrist, Joshua D. and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 1999, *114(2)*, 533–575.

_ **, Erich Battistin, and Daniela Vuri**, "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno," *American Economics Journal: Applied Economics*, 2017, *9* (4), 216–249.

Arai, Yoichi, Yu-Chin Hsu, Toru Kitagawa, Ismael Mourifie, and Yuanyuan Wan, "Testing Identifying Assumptions in Fuzzy Regression Discontinuity Design," 2018. Unpublished mimeo.

Barreca, Alan I., Jason M. Lindo, and Glen R. Waddell, "Heaping-Induced Bias in Regression Discontinuity Designs," *Economic Inquiry*, 2016, *54* (1), 268–293.

_ **, Melanie Guldi, Jason M. Lindo, and Glen R. Waddell**, "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *The Quarterly Journal of Economics*, 2011, *126* (4), 2117–2123.

Blass, Nachum, Shay Tsur, and Noam Zussman, "What Did You Learn in School Today, Dear Little Boy of Mine? The Use of Teacher Work Hours in Primary Schools," 2012. Bank of Israel, Research Department, Working Paper 2012.3, February 2012, Hebrew. http://www.boi.org.il/deptdata/mehkar/papers/dp1203h.pdf.

Bonesronning, Hans, "Class Size Effects on Student Achievement in Norway: Patterns and Explanations," *Southern Economic Journal*, 2003, *69* (4), 952–965.

Dobbelsteen, Simone, Jesse Levin, and Hessel Oosterbeek, "The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition," *Oxford Bulletin of Economics and Statistics*, 2002, *64(1)*, 17–38.

Gary-Bobo, Robert J. and Mohamed-Badrane Mahjoub, "Estimation of Class-Size Effects, Using "Maimonides' Rule" and Other Instruments: the Case of French Junior High Schools," *Annals of Economics and Statistics*, 2013, *111-112*, 193–225.

Gerard, Francois, Miikka Rokkanen, and Christoph Rothe, "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable," 2018. NBER Working Paper No. 22892.

Hoxby, Caroline, "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 2000, *115 (4)*, 1239–1285.

Jacob, Brian A. and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 2003, *118(3)*, 843–77.

Kliger, Aviva, "Between GEMS and Reality: Teacher and Principal Understanding of the GEMS Testing Program," *Dapim ('Pages')*, 2009, *47*, 142–184. Hebrew.

Krueger, Alan B., "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 1999, *114*, 497–532.

Lavy, Victor, "Expanding School Resources and Increasing Time on Task: Effects of a Policy Experiment in Israel on Student Academic Achievement and Behavior," 2012. NBER Working paper, No. 18369.

Leuven, Edwin, Hessel Oosterbeek, and Marte Ronning, "Quasi-experimental Estimates of the Effect of Class Size Achievement in Norway," *The Scandinavian Journal of Economics*, 2008, *110(4)*, 663–693.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

Ministry of Education, "Budgeted Instruction Hours in Elementary Schools in the Formal Regular Education System for the School Year 2016," 2015. Senior Vice President and Director of the Pedagogic Administration, Israeli Ministry of Education Memo, released August 2015, Hebrew.

_ , "Guidelines for Students Enrollment Reporting - Primary schools, Academic Year 2016," 2015. ICT Administration and Information Systems, Pedagogical Administration, Data collection center, Israeli Ministry of Education Memo, released June 2015, Hebrew.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita, "Estimation and Inference of Discontinuity in Density," *Journal of Business & Economic Statistics*, 2013, *31* (4), 507–524.

Piketty, Thomas, “Should We Reduce Class Size or School Segregation? Theory and Evidence from France,” 2004. presentation at the Roy Seminars, Association pour le dévelopement de la recherche en économie et en statistique (ADRES), 22 November, available at: http://www.adres.polytechnique.fr/SEMINAIRE/221104b.pdf.

Shafrir, Reut, Yossi Shavit, and Carmel Blank, “Is Less Really More? On the Relationship between Class Size and Educational Achievement in Israel,” in Avi Weiss, ed., *Taub Center State of the Nation Report: Society, Economy, and Policy in Israel*, 2016.

Sims, David, “A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California,” *Journal of Policy Analysis and Management*, 2008, *27(3)*, 457–478.

Urquiola, Miquel, “Identifying class Size Effectsin Developing Countries: Evidence from Rural Bolivia,” *Review of Economics and Statistics*, 2006, *88(1)*, 171–177.

_ **and Eric Verhoogen**, “Class Size Caps, sorting, and the Regression Discontinuity Design,” *American Economic Review*, 2009, *99(1)*, 179–215.

Vurgan, Yuval, “The Change in the Allocation of Instructional Hours to Elementary Schools,” 2007. Knesset of Israel, Research and Information Center, memorandum dated July 11, 2007, Hebrew. http://www.knesset.gov.il/mmm/data/pdf/m01847.pdf.

_ , “The Number of Students in Class In the Israeli Education System - A Snapshot,” 2011. Knesset of Israel, Research and Information Center, memorandum dated August 30, 2011, Hebrew. http://www.knesset.gov.il/mmm/data/pdf/m02912.pdf.

_ , “Teachers’ Outsourcing,” 2014. Knesset of Israel, Research and Information Center, memorandum dated June 30, 2014, Hebrew. http://www.knesset.gov.il/mmm/data/pdf/m03413.pdf.

Woessmann, Ludger, "Educational production in Europe," *Economic Policy*, 2005, *43*, 445–493.

Chapter 4

Measuring Strategic Data Manipulation: Evidence from a World Bank Project

By Jean Ensminger and Jetson Leder-Luis

Abstract

We develop new statistical tests to uncover strategic data manipulation, and we apply these methods to a World Bank project in Kenya. Data produced by humans follow different patterns than naturally occurring data, which motivates an analysis of digit distributions. These new tests unmask profitable data fabrication and suggest efforts to subvert detection. We find evidence consistent with higher levels of fraud in poorly monitored sectors and in a Kenyan election year when graft also had political value. These methods are validated with results from a forensic audit of the same project, which found extensive levels of suspected fraudulent transactions.

4.1 Introduction

Corruption is difficult to measure because those who commit illicit acts have an incentive to conceal their behavior. In this paper, we develop new digit analysis methods that exploit irregular digit patterns produced during strategic human data fabrication. These tests derive from two principles: that fraudulent data manipulation responds to economic and political incentives, and that humans are poor generators of random numbers. We apply our new tests to data from a Kenyan World Bank development project. This analysis is specifically tailored to detect patterns consistent with strategic cheating, and these methods also outperform the statistical power of existing digit analysis methods.

Our digit results point to serious levels of data fabrication. The scale of the problem is noteworthy, as is the finding that it intensified during the national election cycle when politics added a further incentive for graft. We are able to statistically validate our method by comparing variation across districts from our digit analysis (showing a failure rate of 30% to 70% , with 60% overall for all districts) to findings for the same districts from the forensic audit (showing 44% to 75% suspected fraudulent or questionable transactions, with an average of 66% across all districts).[1]

Our method involves 10 distinct digit analysis tests that exploit different dimensions of the data. These tests include comparisons to the appropriate theoretical distributions of digits, Benford's Law, but take the novel approach of simultaneously analyzing multiple digit places to provide increased statistical power. We augment this result with a comparison between geographic regions in our dataset, which can be expected to have similar results.

[1] The World Bank flagged 66% of the district transactions as suspicious; of these, 49% were classified as suspected fraudulent and 17% as questionable.

Furthermore, we compare distributions by characteristics of the spending, such as expenditure type, year, and value of the manipulated digits. The consistency of these results with the forensic audit suggests that this method can be used to mine data for patterns consistent with fraudulent behavior. Such applications may contribute to real time project monitoring and the targeting of costly forensic audits.

This work relates to a large literature on the monitoring of corruption, including the literature on audits. Ferraz and Finan (2008) show that randomized audits affect the electability of corrupt officials in Brazil; Avis et al. (2016) show that these same audits deter subsequent corrupt behavior. Olken (2007) demonstrates that a rise in Indonesian government audits from 4 percent to 100 percent decreases missing expenditures in road projects by 8.5 percent. Despite their effectiveness, field verified audits are expensive and difficult to implement, as they are labor intensive, use highly trained personnel, and require cooperation from individuals who may be implicated in the corruption, including auditors themselves (Duflo et al., 2013). Similarly, monitoring efforts that track expenditures (e.g. Reinikka and Svensson, 2006) suffer from issues of scale and the need for cooperation from those who may be complicit in the fraud.

A number of creative studies have pioneered methods for identifying and measuring corruption and fraud that require no cooperation from the subjects being monitored. Jacob and Levitt (2003) use aberrant patterns in test scores to uncover cheating in Chicago public schools. Satellite data have been used to track illegal logging (Burgess et al., 2012); stock market fluctuations have been used to measure financial returns to political influence in Indonesia (Fisman et al., 2006, Fisman, 2001). Like these studies, digit analysis does not require the cooperation of potential subjects. It has the added advantages that it scales well and holds potential for application across many domains.

Digit analysis as a method is not unique to this paper, nor are its applications restricted to auditing, though it has had considerable impact there (Amiram et al., 2015, Durtschi et al., 2004, Nigrini, 2012, Nigrini and Mittermaier, 1997). Digit analysis has also been employed to detect and monitor election fraud (Alvarez et al., 2008, Beber and Scacco, 2012), IMF data manipulation (Michalski and Stoltz, 2013), campaign finance fraud (Cho and Gaines, 2012), scientific data fabrication (Diekmann, 2007), and enumerator integrity during survey research (Bredl et al., 2012, Judge and Schechter, 2009, Schräpler, 2011). Our methods build upon this existing work and sharpen the method, providing both new tests and improved statistical power.

The remainder of this paper is organized as follows. Section II describes our dataset and the context of the World Bank project. Section III motivates our digit analysis tests with a discussion of the economics of data manipulation and an overview of the mathematical principles that govern digit distributions. Section IV presents our statistical tests and results. Section V validates our results by comparison to the World Bank forensic audit of the same project, and section VI concludes.

4.2 Data and Institutional Context

We analyze data from the Kenyan Arid Lands Resource Management Project (World Bank, 2003). This World Bank project ran from 1993 to 2010, eventually serving 28 arid and semi-arid districts, encompassing over 75 percent of Kenya's land area. The project spent $ 224 million USD, targeting the most impoverished people in the heavily drought-prone regions of Kenya. It funded small infrastructure (schools, dispensaries, and water systems), income generating activities, drought and natural resource initiatives, and training exercises for villagers.

The data used in these analyses are from the original 11 arid districts that received funds from the project; they cover the years 2003 to 2009. These districts share many similar characteristics. Their economies depend primarily upon livestock and are among the poorest in Kenya—remote from centers of power, poorly educated, and sparsely supplied with infrastructure (roads, schools, health services, access to clean water, and electricity).

The expenditure and participant data used in these analyses were culled from electronic project reports produced in each of the 11 districts. These reports break out the expenditures and numbers of male and female participants associated with each activity undertaken by the project that year. Ensminger's interview data with project staff indicate that usually only 1 or 2 individuals were involved in entering the data for a project component (natural resources and drought management, community driven development, and support for local development). Officers in the districts had considerable latitude over the magnitude of expenditures within budget categories.[2]

The district staff was subject to oversight both from project headquarters in Nairobi and to a lesser extent from the World Bank. Government auditors also routinely audited the project at both the district and the national levels, but Ensminger's interviewees consistently reported that the auditors at both the district level and headquarters were bought off. The World Bank's supervisory missions and financial management oversight were also

[2] We exclude community driven development projects from our analysis. These expenditures were grants that were subject to caps, and as such are not appropriate for digit analysis. However, we do use data (expenditures and numbers of participants) from the training exercises and transport costs that supported these activities, as they were not subject to fixed caps.

largely ineffective in their monitoring. Numerous missions rated project financial management "satisfactory" across many years of operations right up to the point of the forensic audit, and the project was labeled "exemplary" in the project renewal proposal (World Bank, 2003: 84). Further, its financial management system was lauded and used as a model for another project, indicating that the World Bank's standard monitoring did not pick up problems (World Bank, 2007).

In 2009 the World Bank's Integrity Vice Presidency (INT) began a forensic audit of the project that lasted 2 years and culminated in a public report (World Bank, 2011). Auditors sampled 2 years' worth of receipts for 7 districts, 5 of which were arid districts examined in this analysis. They examined 28,000 transactions. The auditors worked from actual project receipts and supporting documents, such as cashbooks and bank statements. They also travelled to the districts to conduct interviews with suppliers to verify the legitimacy of suspicious transactions. We conduct digit analysis on the reported total expenditures for each of these transactions, such as the total cost of a training exercise, while the forensic auditors investigated the underlying individual receipts for the same transactions.

To understand the decision-making of project staff contemplating embezzlement, it is important to establish their perceptions of the probability of getting caught, and the likely consequences should that happen. Despite the fact that this project was eventually the subject of an intensive World Bank audit, there is reason to believe that staffers would not have considered that to be a likely outcome. To the best of our knowledge, no other field-verified, transaction-based, forensic audit of this scope had taken place at the World Bank before this one, nor has one occurred since (Stefanovic, 2018).[3] More likely, staff engaging in

[3] For example, this is the only such audit on the World Bank INT website (World Bank Integrity Vice

embezzlement feared that their superiors or the Kenyan auditors would expect kickbacks if their activities were exposed.[4] Given this environment, the costs and consequences of being caught embezzling consisted mainly of paying a portion of one's takings, rather than risks of career consequences or prosecution.

We make use of 2 characteristics in the structure of our dataset. First, we have data from 11 arid districts with similar demographics, livelihoods, and ecological conditions, reporting on similar activities, and operating under the same project rules. We proceed from the null hypotheses that digit distributions will be similar across districts. Second, we have data on the number of participants in hundreds of training exercises, which is a count of people who responded to an open invitation for a training exercise. When the same pattern of deviations from theoretical distributions appears in both the expenditure and the participant datasets, it is strongly indicative of human tampering.

4.3 Theory

We begin by examining the problem of a bureaucrat's decision to accurately report or to fabricate data when tasked with producing expenditure reports. Using a set of receipts dedicated to a single transaction, such as the construction of a classroom, an honest bureaucrat calculates the sum of all the construction related receipts and enters the total in the report. These data follow the digit patterns of natural data, described later, as they

Presidency, 2018).

[4] The World Bank referred the Arid Lands case to the Kenyan Anti-Corruption Commission after completing a joint review together with the Kenya National Audit Office, which confirmed the findings and resulted in the Kenyan government's agreement to repay the World Bank $4 million USD for disallowed charges (Integrity Vice Presidency of the World Bank and Internal Audit Department, 2011). It is noteworthy, therefore, that no one from this project was taken to court, and this speaks to the probability of consequences in the current Kenya context.

accurately reflect the data without human interference. Across different geographic regions of this World Bank project, we would expect similar patterns in the financial data when reporting is conducted honestly.

Bureaucrats have an incentive to falsify expenditure data and embezzle both for personal gain as well as to satisfy kickback demands from superiors. Embezzlers weigh the costs and benefits of such behavior, including the probability of getting caught and the size of the penalty, in line with rational crime theory (Becker, 1968). Other costs may include payoffs to auditors or others who detect their fraud, as we discuss in the previous section.

When a bureaucrat determines that the benefits of data manipulation outweigh the costs, we can expect that they will manipulate the data to maximize payout and minimize the probability of detection. This can consist of a number of behaviors. A manipulator may change digits to maximize payout, or may invent new line items to increase the total reported expenditure. In line with a rational decision to commit fraud, we can expect that reporters would increase data tampering in response to greater incentives to steal, and attempt to produce data that appear random to subvert detection. We would furthermore expect that the bureaucrat would expend lower effort in subverting detection for data that are less likely to be monitored.

We analyze each dimension of the data and provide a set of non-overlapping tests that capture different ways in which data can be manipulated. Our tests fall into 3 categories: tests of digit conformance to expected distributions, tests of covariate characteristics of the data (e.g. district, year, and sector), and tests of strategic intent to deceive.

Benford's Law describes the distribution of digits in many naturally occurring circumstances, including financial data. Benford's Law is given mathematically by (Hill, 1995):

$$P(D_1 = d_1, \ldots, D_k = d_k) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \times 10^{k-i}}\right)$$

We have, for example, the probability that the first three digits are "452":

$$P(D_1 = 4, D_2 = 5, D_3 = 2) = \log_{10}\left(1 + \frac{1}{452}\right)$$

In the first digit place, Benford's Law produces an expected frequency of 30.1 percent of digit 1 and 4.6 percent of digit 9. In later digit places, this curve flattens, and by the 4th digit place the distribution is nearly identical to the uniform distribution, with expected frequency 10.01 percent of digit 1 and 9.98 percent frequency of digit 9 (Hill, 1995, Nigrini and Mittermaier, 1997). Table 1 shows the full digit-by-digit place table of expected frequencies under Benford's Law. Datasets known to follow Benford's Law include financial data and population data, but also everything from scientific coefficients to baseball statistics (Amiram, Bozanic and Rouen, 2015, Diekmann, 2007, Hill, 1995, Nigrini and Mittermaier, 1997).

The intuition behind Benford's Law is revealed if one imagines it as a piling-up effect: increasing a first digit from 1 to 2 requires a 100 percent increase, while increase from a first digit of 8 to 9 requires a 12 percent increase (Nigrini and Mittermaier, 1997). Furthermore, Benford's Law arises from data drawn as random samples from random distributions (Hill, 1995). Because numbers repeatedly multiplied or divided will limit to the Benford distribution (Boyle, 1994), financial data can be expected to follow this natural phenomenon (Hill, 1995, Nigrini and Mittermaier, 1997).

The appropriateness of Benford's Law for analysis of our data set is confirmed by the conformance of the first digits to the Benford distribution, as we show later. The nature of our expenditure data, which are based upon sums of numerous receipts that in turn include sums and multiplication of price times quantity, provides a theoretical basis for why we can expect Benford's Law to be the appropriate distribution. In our analysis, we consistently performed robustness checks by comparing our observed distributions to both the Benford and the uniform distributions. The statistical significance under the uniform distribution is even greater than those reported here. Finally, regardless of Benford's Law, tests of later digit places, particularly last digits, should be uniformly distributed under most conditions.[5]

When experimental subjects are asked to produce random numbers, studies consistently show divergent patterns of human digit preferences. When students were asked to make up strings of 25 digits, their results followed neither the Benford distribution nor the uniform distribution (Boland and Hutchinson, 2000). The patterns produced by the subjects varied greatly, with individuals exhibiting different preferences for certain digits. Other experiments have shown similar results of individual digit preferences, confirming the inability of humans to produce random digits (Chapanis, 1995, Rath, 1966).

It is possible that specific digit preferences are culturally influenced, in which case it is instructive to have a culturally representative baseline for comparison. Evidence of specific digit preferences from Africa comes from an overview of African census data, where statis-

[5] In the study of elections, the use of Benford's Law has been contested based on concerns over the distributions of data that produce voting counts (Beber and Scacco, 2012, Deckert et al., 2011, Walter R. Mebane, 2011). However, these criticisms do not extend to our financial dataset or individual participant counts, both of which come from distributions that can be expected to conform to Benford's Law. Specific auditing guidelines over which types of data conform to Benford's Law includes these types of data (Durtschi, Hillison and Pacini, 2004).

ticians discuss a phenomenon known as age heaping, wherein self-reported demographic records show a preference for certain ages. Many Africans of older generations do not know their exact age, and their responses to census takers represent their best approximation. This is an example of humanly generated data that shows specific digit preferences. Among the African censuses, we see a strong preference for the digits 0 and 5, with secondary strong preferences for 2 and 8, and disuse of 1 and 9 (Nagi et al., 1973, UN Economic and Social Council Economic Comission for Africa, 1986). Throughout our analysis, we omit 0 and 5, which are heavily overrepresented, and analyze digits 1-4 and 6-9; we report rounding levels as measured by 0 and 5 separately.

4.3.1 Digit Tests And Results

A. Tests of Digit Conformance to Expected Distributions: All Digit Places Beyond the First

Our first test is a simultaneous analysis of all digit places beyond the first digit for conformance to Benford's Law. We do not include the first digit because individuals tampering with data may not have complete control over the leading digit, or may avoid changing it to subvert detection. Compared with single digit place tests, which are common in the existing literature, a simultaneous analysis of multiple digit places increases sample size for statistical testing and therefore vastly increases statistical power.[6] The increase in sample size afforded by simultaneous digit place analysis is especially helpful when analysis can benefit from data disaggregation, resulting in low n.

[6] Individual digit place analyses beyond the first include second and last digit analysis. (Beber and Scacco, 2012, Diekmann, 2007, Nigrini and Mittermaier, 1997).

We use a two-way chi square test to compare the contingency table of all digit places beyond the first against the Benford distribution. As discussed before, we omit 0 and 5 from this analysis, which are handled separately in a discussion of rounding, below. For each digit place (1st digit, 2nd digit, etc), the frequency of each digit (1, 2, 3, 4, 6, 7, 8, 9) is compared with the expected frequencies given in Table 1. This is in contrast to existing studies, which analyze a single digit place with a single chi square test. Because the Benford distribution gives different frequencies by digit place, the two-way chi square test is the appropriate test rather than testing individual digit places. Furthermore, it corrects for multiple hypothesis testing issues that arise from individual digit place analysis.

Figures 1 and 2 present the data of all digit places beyond the first for expenditure (1) and participant data (2). The data are projected onto one axis for visualization. Among the expenditure data for all districts in Figure 1, we see a strong preference for digits 2 and 8, underreporting of 1 and 9, and overall non-conformance to the expected Benford distribution ($p = 3.9 \times 10^{-15}$). Strikingly, these same digit patterns appear in the participant data (Figure 2), and the result for all district data combined is again highly significant ($p = 5.7 \times 10^{-51}$). This pattern is also consistent with the humanly generated African census pattern described earlier. To account for multiple non-overlapping tests, we use a Bonferroni correction: we divide our desired significance level (.05) by the number of tests (10) and set a significant level of $p = .005$, used throughout our analyses. In 8 of our 11 districts we reject the null hypothesis that all digit places conform to Benford's Law for both the expenditure data and the participant data at the $p < 0.005$ level. The lack of conformance to the expected distribution, consistency with known humanly generated data from African census studies, and similar patterns across both expenditure and participant data, are strong indicators that these data have been tampered with.

We do not include a test of the last digit place among our 10 tests because it is technically subsumed under this test, and we wish to avoid non-independence across our tests. Benford's Law predicts a uniform distribution in digit places beyond the fourth; that is, there is no reason that more data should end with a 4 instead of a 3. For comparison to other studies, we include the results of last digit analysis in Appendix A. Both the expenditure and the participant data diverge significantly from the predicted distributions, and both are consistent with our other tests, though we do not include them in the final tally of tests.

B. Tests of Digit Conformance to Expected Distributions: First Digits

Next, we test conformance to the Benford distribution in the first digit place of the expenditure data, where we expect digits to follow (Hill, 1995):

$$P(\text{First Digit} = d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

Figure 3A plots this distribution as a solid line and shows the conformance of the first digits to Benford's Law. Data from the full sample of districts are not statistically significantly different from the expected distribution ($p = 0.089$) under a chi-square test. This supports the hypothesis that Benford's Law is the appropriate theoretical distribution for our dataset. Importantly, this does not indicate that the data are legitimate, as pooled data may cancel out different individual signatures of manipulation and replicate Benford's Law (Diekmann, 2007). This becomes evident when we look at the data from individual districts where the reports were constructed. Figure 3B shows the first digits from Ijara district, with

$p = 2.3 \times 10^{-13}$. Ijara District uses the digit 2 in the first digit place almost twice as often as predicted. Seven of our 11 districts are significantly different from Benford's Law at the $p < 0.005$ level.

C. Tests of Digit Conformance to Expected Distributions: Digit Pairs

Underuse of digit pairs, e.g. 11, 22... 99, is a common feature of humanly produced data (Boland and Hutchinson, 2000, Chapanis, 1995). Other applications of digit analysis examine the last two digits (Nigrini, 2012), or explicitly test for digit pairs (Beber and Scacco, 2012).

Among the participant data, we expect a uniform distribution of terminal pairs, 9 of 99 pairs. We omit the pair 00 in case of rounding. We compare the observed number of digit pairs against the expected proportion using a binomial test, where the number of trials is the total combination of terminal digits observed. These data most typically record the number of women and men (listed separately) who showed up in response to an open invitation to appear for a training exercise in their village. To avoid use of first digits, we use participant data only if it has 3 or more digit places. This test is performed on the sum of male and female participants. A digit pair analysis of participant data is shown in Figure 4. Six of the 11 districts significantly underuse final digits pairs in the participant data at $p < 0.005$ significance, as does the combined sample of all districts ($p = 1.4 \times 10^{-9}$). However, Isiolo District significantly overuses repeated pairs, with $p = 5.6 \times 10^{-5}$ in the binomial test.

Due to the low value of the Kenyan shilling, rounding at the 1 shilling level may be legitimate among expenditure data. Therefore, an equivalent analysis of expenditure data is not justified, as an underuse of digit pairs (e.g. 22) is confounded by a legitimate use of

1-shilling rounding (e.g. 20). For this reason, we confine our analysis to the beneficiary data, where there is no legitimate reason for rounding in the ones place, as participant data are reported as exact counts.

Our next 4 tests exploit the attribute data available in our dataset: districts, years, and sectors (civil works, goods and equipment, training, and transport). These tests do not rely upon Benford's Law.

D. Tests of Digit Covariate Characteristics: Comparisons of District Patterns in Rounding and Repeating

Our next two tests uncover patterns consistent with human tampering, as evidenced by substantial variation across districts without a plausible naturally occurring explanation. It is common for auditors to look for both high levels of rounded and repeated data, and these are often viewed as potential evidence of human tampering (Nigrini, 2012, Nigrini and Mittermaier, 1997). In the absence of theoretically acceptable levels of rounding and repeating, we compare districts to each other, as there is no reason to expect differences among them.

The Kenyan shilling was 66 to $ 1 USD in 2008. Its value was low enough that many receipt data would legitimately show high levels of 0s and 5s in the terminal digit places. However, one must bear in mind that these expenditure data represent sums of many receipts; it takes only one receipt ending in a non-0 or 5 to create a different terminal digit for the entire transaction, and it is these transaction totals that we are examining.

We count rounded digits rather than rounded line items, tallying the number of trailing 0s (0, 00, 000, etc.), or digits in terminal strings of 5, 50, or 500, as a fraction of the total digits in the district dataset. For example: the number 30,000 has 4 rounded digits; the number 12,350 has 2 rounded digits; and the number 11,371 has 0 rounded digits. Rather than indicating individual line items, counting rounded digits is a more sensitive indicator because it penalizes use of numbers such as 10,000 (4 rounded digits) more than the use of a number such as 10,600 (2 rounded digits). We compute the percentage of rounded digits for each district.

Figure 5 shows the percentage of rounded digits by district, with the crosshatched districts in the top quartile of rounding. While we don't know the empirically correct level of rounding that one should observe in the dataset, there is good reason to expect that the same type of retailers, servicing the same type of contracts for similar districts, practiced the same rates of rounding. In the absence of an expected level of rounding, we flag those three districts, roughly the top quartile, that round most heavily, which is more than twice that of the lowest rounding district.

Exactly repeated numbers are also a red flag for auditors (Nigrini and Mittermaier, 1997). Our hypothesis is that embezzlers expended less effort in data fabrication when there was less reason to expect scrutiny. Repeated values are consistent with low-effort data fabrication. One such example is remote training exercises, which are particularly hard to verify.

A specific example from the Tana District Report of 2003-6 illustrates the problem of repeated data (Republic of Kenya, 2006). On page 49 we find 8 training exercises listed that took place in different villages for three weeks, each from March 5-27. The district had neither enough vehicles, nor enough training staff to run 8 simultaneous trainings. Among the 8 expenditures listed, we find the identical cost (245,392 Kenyan Shillings) listed for 3 different trainings, and another number (249,447) exactly repeated twice. Trainings are the summed costs of the per diems for 4-5 trainers and 1 driver (at different rates), the cost of fuel to the destination, stationary for the seminar, and 100 Kenyan Shillings per day, per trainee, for food costs. The number of trainees for each of these seminars is listed, and they range from 51 to 172. The expenses reported do not track the estimated food costs, as one would expect; indeed, the cost of training for 172 trainees should have exceeded all of the amounts listed.

Note that duplicate entries for the same project were removed from the dataset. In our calculations, repeating numbers refer to the use of identical expenditure amounts for completely different activities. We define an exact repeat to be an expenditure matching year, district, sector, and expenditure value. There is no correction for rounding in the repeating data, as we wish to maintain the independence of our tests for rounding and repeating.

Figure 6 shows the results for the percentage of line items that repeat exactly. As we did with rounding, we indicate the top three districts that most heavily repeat numbers; for example, Baringo approaches 50 percent, while Turkana has about 5 percent. Although the empirically appropriate level of repeating is unknown, we rely on the fact that there is no reason for patterns across districts to differ. Figures 5 and 6 flag different districts in rounding and repeating behavior, indicating that these two tests pick up different signals.

E. Tests of Digit Covariate Characteristics: Year Effects and the 2007 Kenyan Election

We take advantage of the extra power afforded by use of our new test for analyzing multiple digit places simultaneously to partition our data by project year. This test is designed to detect potential fraud in a presidential election year (2007), which increased incentives to embezzle money for political campaigns. We look for padding of high digit numbers by project year by analyzing the proportion of high to low digits (6, 7, 8, and 9 versus 1, 2, 3, and 4) in all digit places beyond the first. We conduct a chi-square test on the contingency table of high versus low digits. We expect that the probabilities of high and low digits should follow the total probability of those digits from Benford's Law in each digit place. As before, we project this contingency table onto one axis for visualization.

As we see in Figure 7, while all other years slightly underused high digits on average, in 2007 (the only election year) there was a statistically significant overuse of high digits ($p = 6.5 \times 10^{-6}$). This is consistent with a greater incentive to embezzle during a presidential election year to support political campaigns. This phenomenon is supported by Ensminger's interview data and the well-known general pattern of large corruption scandals just prior to national elections in Kenya.

F. Tests of Digit Covariate Characteristics: Sector Effects

Economic theory (Becker, 1968) and empirical work (e.g. Olken, 2007) indicate that individuals are more likely to cheat when there is a lower risk of detection. Training and transport (travel, fuel, and vehicle maintenance) provide greater opportunities for individuals to conduct fraud when compared to civil works projects or the purchase of goods and equipment, because the latter leave physical evidence of spending, while the former do not. For example, tracking down nomads who were reported as present for a training exercise in a remote village two years prior to an audit is all but impossible. Similarly, fuel can be diverted to private vehicles while leaving no trace. Therefore, we predict that individuals fabricating data for these sectors may do so with less effort expended on deception. We look for evidence of a greater incidence of repeated numbers in these sectors. We plot the percentage of repeated line items that match year, district, sector, and amount, for each of the districts by sector. Figure 8 shows this result.

We crosshatch those districts that have three times the number of repeats in training and transport as compared to the average number of repeats in civil works and goods and equipment. Six of 11 districts and the all district test fail, but Turkana District provides evidence that there is no structural reason for there to be more repeated data in training and transport.

G. Tests of Strategic Intent: Unpacking Rounded Numbers

Much of what auditors catch in their routine work falls into the category of sloppy bookkeeping. While there may be a strong correlation between firms and individuals whose paperwork is sometimes incomplete or missing, and actual embezzlement, it is not necessarily the case that sloppy bookkeepers are misappropriating funds. For this reason, evidence that points to consistently profitable deviations from expected digit distributions, or evidence of strategic efforts to avoid detection, bring us a step closer to deducing intent to defraud. We turn now to the first of two new tests that reveal strategic data manipulation.

Project staff had an incentive to inflate the number of participants in training activities because they claimed food expenses for each participant at 100 Kenyan Shillings (about $ 1.50 USD) per person, per day. The authors of the annual district reports also had reason to expect that participant data would not be as carefully scrutinized as expenditure data. First, the impact of participants on expenditures was obscured because it was only one component of the full costs of a single training exercise, and second, training exercises in remote villages are notoriously difficult to verify. With the threat of oversight reduced, we speculate that less effort was devoted to covering up data fabrication.

We further surmise that officers fabricating participant data may have begun with an embezzlement target in mind, which they converted to a round number of participants. This total number of participants was then split into males and females, as was required for reporting. Therefore, we expected greater indicators of data fabrication when the total number of participants was a round number (e.g. 300).

To test this, we analyze the distribution of all but first digits of numbers of total participants (males and females) when their sum ends in a 0 versus a non-0 digit. We perform a chi-square test on the contingency table of digits in digit places beyond the first, versus Benford's Law. Theoretically, the breakout of participant data by gender should show statistically identical digit distributions between these conditions. However, we see a much higher instance of 2s and 8s and low incidence of 1s and 9s when the gender specific data come from a pooled number that ends in 0 (Figure 9A). This pattern is consistent with humanly generated data and not with naturally occurring data. There is still evidence of human generation in the data when the gender total is not round, Figure 9B ($p = 1.9 \times 10^{-6}$), but the statistical significance is considerably higher in the rounded data, Figure 9A ($p = 2.6 \times 10^{-64}$ in the sample of all districts). For 8 out of 11 districts, we reject the null hypothesis that the male and female participant data, when totaling to a round number, are Benford conforming ($p < 0.005$).

H. Tests of Strategic Intent: Value of Digit Place with Monte Carlo Simulation

Our final new test reveals patterns consistent with data manipulation that is both profitable and consistent with attempts to conceal such manipulation. We identify padding of expenditures by measuring overuse of high digits based on the monetary value of the digit place. We hypothesize that individuals fabricating data do so strategically, and therefore place additional high digits in the more valuable digit places. Furthermore, we detect signs of behavior consistent with an attempt to make it more difficult to detect the padding by overusing low digits in less valuable digit places.

Benford's Law governs the distribution of digits by position from the left (1st digit, 2nd digit), but not by value, which depends on digit place from the right (e.g. 1s, 10s, 100s place). To overcome this limitation, we compute the expected mean under Benford's Law by digit place from the right (10s, 100s), using the length of the numbers in our dataset to match left-aligned digit places and right-aligned digit places. We compare the observed mean of our data to the expected mean under Benford's Law. This is the difference of means statistic, for which a positive value indicates a mean greater than the expected mean under Benford's Law. We then perform a Monte Carlo simulation of 100,000 Benford-distributed datasets, and compare the difference-of-means statistic of the project data to the simulated data, and find the probability of observing our results under the Benford distribution. The Appendix contains technical details of this process.

Figure 10 shows the project data by sector against the Benford expected distribution. The 0 line indicates the Benford mean; anything above the line represents an overuse of high digits, and anything below the line represents an underuse. The project data in the 10,000s place exceeded 100 percent of the 100,000 simulated Benford-conforming datasets ($p = 1.0 \times 10^{-5}$). We also see a significantly high mean ($p = 2.3 \times 10^{-4}$) in the thousands place. At the district level there is statistically significant evidence of padding in the 10,000's place for 8 of 11 districts. Ten thousand Kenyan shillings was worth approximately $ 150 USD in 2007.

An interesting finding in Figure 10, which corroborates the strategic placement of digits, is the decline in the use of high digits as one goes from the 10,000s to the 1,000s, 100s, 10s, and 1s places among the pooled sector data, represented by the black bars. This is consistent with a strategy of padding extra high digits in the high value places and compensating by *underutilizing* high numbers in the low digit places. The human data generators may have been trying to avoid detection from an auditor or supervisor, who might otherwise have noticed the overuse of high numbers in any given table in the report.

I. Summary: Application of Digit Tests

These tests, taken together, comprise a set of non-overlapping analyses along different dimensions of potential data manipulation. Importantly, some of the tests are not a turnkey system for digit analysis under other circumstances. Some characteristics of this dataset, such as the comparison of expenditure to beneficiary tests, are particular to these data, but are likely to have analogies in many real world situations.

The exact battery of tests that can be performed on other datasets depends on both the incentives for manipulation in that dataset, as well as the specifics of the attribute data that are available. What we show is that analysis along all available dimensions of *our* data can be used to uncover suspicious patterns in an efficient and effective way. We expect that our new tests, especially the powerful test using all digit places, and the last test, which takes account of the value of the digit place, should prove useful in many contexts.

By facilitating the full use of our attribute data, this battery of tests helps reveal the magnitude of potential fraud in this project, as well as the important finding that aid funds were very likely being diverted to campaign coffers during an important presidential election year.

4.4 Comparing Digit Analysis to The World Bank Forensic Audit

Table 2 compiles the results of 10 tests for each district. To correct for type 1 error due to the number of tests we ran, we perform a Bonferroni correction. We divide our desired significance level (0.05) by the number of tests (10), and therefore choose a significance level of 0.005. For the rounding and repeating tests where districts are compared to each other in the absence of a theoretical measure (Figures 5 and 6), the top quartile of districts, 3 of 11, are flagged. For the sector effects (Figure 8), we mark those districts for which training and transport exhibit more than triple the level of repeating compared to the other sectors. These 10 tests avoid overlap and pinpoint different aspects of data tampering. In the bottom row, we sum the number of failed tests by district, which ranges from 3 to 8 out of 10.

The existence of an extensive forensic audit for this project provides us with a measure of external validity for our digit analysis. In Table 3 we compare the results of our digit analyses by district to the results of the World Bank auditors (World Bank, 2011). The World Bank audit found that 4 of the 5 districts for which we have both digit and audit results had 62-75 percent suspected fraudulent or questionable expenditures. In our digit analysis, we rejected the null hypotheses for those same 4 districts in 6 to 8 of our 10 digit tests. The remaining district, Tana, had considerably lower levels of suspected fraud than the other districts (44 percent), and we rejected the null on 3 of our 10 digit tests. A Pearson's correlation test of the 5 districts for which we have both digit tests and the World Bank audit shows a correlation of .939, and a 95% confidence interval of [.338, .996]. We reject the null hypothesis of no correlation at the 5% significance level, with $p = 0.018$. The World Bank's forensic audit confirms the findings from our digit analysis tests.

We also found significant digit violations in all of the unaudited districts, which is consistent with the conclusions of the auditors that these problems were systemic throughout all sectors and all districts of the project. Of the remaining 6 districts that were not audited by the World Bank, we see that half (Mandera, Baringo, Ijara) have some of the highest number of digit analysis violations (8, 6, and 6) in our sample. This underscores the potential gains of using digit analysis as a diagnostic for targeting costly auditing techniques.

4.5 Conclusion

We present new methods to detect data tampering and demonstrate their use on data from a World Bank dataset in Kenya. Our tests reveal patterns consistent with strategic and profitable data fabrication. Notably, the presence of an independent forensic audit of the same project strongly correlates with our digit analysis, lending external validity to the method and the substantive findings.

One of our new tests, employing the generalized Benford's Law to analyze multiple digit places, provides a statistically powerful test applicable to even relatively small datasets. The ability to work on smaller sample sizes allows more multi-dimensional analyses, such as our comparisons across districts, years, and sectors. By partitioning by project years, we are able to demonstrate that more suspicious patterns emerge in a presidential election year, consistent with allegations that World Bank funds were illegally diverted to fund political campaigns.

Our new test of overuse of high digits in valuable digit places uncovers patterns consistent with profitable deviations as well as attempts to evade detection. This is the first test we know of that relates aberrant digit patterns to the monetary value of the digit place. This is consistent with intentional behavior, rather than sloppy bookkeeping, and to the best of our knowledge is something heretofore not demonstrated in Benford analyses.

The substantive findings of this project attest to the need for, and importance of, better measures and identification of corruption. The forensic auditors determined that 66% of the district transactions they examined were suspected fraudulent or questionable. On average, the districts we examined failed 60% of the all district digit tests.

Our new tests provide a particularly powerful toolkit for monitoring budget expenditures and uncovering suspected fraud. This method works even when field monitoring is challenging, as is often the case in remote and insecure parts of the developing world. In addition, it requires minimal cooperation from those inside the organization or government, who may have an incentive to impede an investigation. In developing countries, where one faces strong corruption cartels, and weak rule of law with which to force compliance, independence is a major benefit.

Readers may be concerned that publication of these methods will provide potential fraudsters with the means to beat the monitors. They need not worry. Engineering a Benford-conforming dataset is a more challenging statistical exercise than is ensuring that digits are uniformly distributed. It would also require centralization across an organization, and matching of all supporting documentation, such as coordination of date-stamped receipts, cashbooks, vehicle logs, cancelled checks, and bank statements. Furthermore, each individual instructed to fabricate data would still face the same incentive to self-deal, which would undercut efforts to produce aggregate results consistent with Benford's Law. Such coordination would also expose leadership at high risk of detection.

Our methods are complementary to newly developed machine learning methods for fraud detection. Machine learning would not be appropriate on this data set, due to the small number of features (columns of data), the relatively small sample size, and the need for a training set of known outcomes, which these and similar datasets lack. However, digit analysis can be used to further machine learning techniques in other contexts. Fundamentally, machine learning relies on pattern detection. The more dimensions of analysis available, the more powerful machine learning becomes. Digit analysis is another dimension along which machine learning can be trained, and the patterns we have detailed in this study can be useful for even more sophisticated fraud-screening techniques.

By addressing external validity and expanding the capabilities of digit analysis, we hope to facilitate its broader use, especially where sample size is an issue, and data partitioning is desirable. The areas that might benefit, and where digit analysis has already been used, are in auditing, election fraud, scientific data fabrication, and the monitoring of enumerator integrity during survey research. The fact that digit analysis can be deployed without the cooperation of potential offenders is a significant advantage for many monitoring efforts. For example, our method could have been used in real time monitoring of this project to reduce potential fraud, or in the forensic audit of this project to identify and target the worst offending districts, three of which were missed in the World Bank audit sample. Such applications can potentially provide substantial savings. We also foresee use in a variety of new applications, for example, to check the authenticity of data supplied by governments in compliance with international economic, ecological and environmental agreements, or pollution and labor data supplied for treaty compliance. In the modern environment where big data proliferates, stronger tools to analyze these data for strategic and profitable manipulation are necessary.

References

Alvarez, R. Michael; Thad E. Hall and Susan D. Hyde eds. 2008. *Election Fraud: Detecting and Deterring Electoral Manipulation.* Brookings Institution Press.

Amiram, Dan; Zahn Bozanic and Ethan Rouen. 2015. "Financial Statement Errors: Evidence from the Distributional Properties of Financial Statement Numbers." *Review of Accounting Studies*, 20(4), 1540-93.

Avis, Eric; Claudio Ferraz and Frederico Finan. 2016. "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians." *National Bureau of Economic Research Working Paper Series*, No. 22443.

Beber, Bernd and Alexandra Scacco. 2012. "What the Numbers Say: A Digit-Based Test for Election Fraud." *Political Analysis*, 20(2), 211-34.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76(2), 169-217.

Boland, Philip and Kevin Hutchinson. 2000. "Student Selection of Random Digits." *The Statistician*, 49(4), 519-29.

Boyle, Jeff. 1994. "An Application of Fourier Series to the Most Significant Digit Problem." *The American Mathematical Monthly*, 101(9), 879-86.

Bredl, Sebastian; Peter Winker and Kerstin Kötschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." *Survey Methodology*, 38(1), 1-10.

Burgess, Robin; Matthew Hansen; Benjamin A Olken; Peter Potapov and Stefanie Sieber. 2012. "The Political Economy of Deforestation in the Tropics." *The*

Quarterly Journal of Economics, 127(4), 1707-54.

Chapanis, Alphonse. 1995. "Human Production of "Random" Numbers." *Perceptual and Motor Skills*, 81, 1347-63.

Cho, Wendy K Tam and Brian J Gaines. 2012. "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance." *The American Statistician*, 61(3), 218-23.

Debowski, Lukasz. 2003. "Benford's Law Number Generator," Polish Academy of Sciences, Institute of Computer Sciences,

Deckert, Joseph; Mikhail Myagkov and Peter Ordeshook. 2011. "Benford's Law and the Detection of Election Fraud." *Political Analysis*, 19, 245-68.

Diekmann, Andreas. 2007. "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data." *Journal of Applied Statistics*, 34(3), 321-29.

Duflo, Esther; Michael Greenstone; Rohini Pande and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *The Quarterly Journal of Economics*, 128(4), 1499-545.

Durtschi, Cindy; William Hillison and Carl Pacini. 2004. "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data." *Journal of Forensic Accounting*, V, 17-34.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *The Quarterly Journal of Economics*, 123(2), 703-45.

Fisman, David; Ray Fisman; Julia Galef; Rakesh Khurana and Yongxiang Wang. 2006. "Estimating the Value of Connections to Vice-President Cheney." *The B.E.*

Journal of Economic Analysis & Policy, 12(3).

Fisman, Raymond. 2001. "Estimating the Value of Political Connections." *The American Economic Review*, 91(4), 1095-102.

Hill, Theodore P. 1995. "A Statistical Derivation of the Significant-Digit Law." *Statistical Science*, 10(4), 354-63.

Integrity Vice Presidency of the World Bank and Internal Audit Department, Treasury, Government of Kenya. 2011. "Redacted Joint Review to Quantify Ineligible Expenditures for the Seven Districts and Headquarters of the Arid Lands Resource Management Program Phase 2 (Alrmp 2) for Fy07 & Fy08," Washington, DC:

Jacob, Brian A. and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics*, 118(3), 843-77.

Judge, George and Laura Schechter. 2009. "Detecting Problems in Survey Data Using Benford's Law." *Journal of Human Resources*, 44(1), 1-24.

Michalski, Tomasz and Gilles Stoltz. 2013. "Do Countries Falsify Economic Data Strategically? Some Evidence That They Might." *The Review of Economics and Statistics*, 95(2), 591-616.

Nagi, M.H; E.G. Stockwell and L.M. Snavley. 1973. "Digit Preference and Avoidance in the Age Statistics of Some Recent African Censuses: Some Patterns and Correlates." *International Statistical Review*, 41(2), 165-74.

Nigrini, M. 2012. *Benford's Law.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Nigrini, M and L Mittermaier. 1997. "The Use of Benford's Law as an Aid in Analytic Procedures." *Auditing: A Journal of Practice and Theory*, 16(2).

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115(2), 200-49.

Rath, Gustave J. 1966. "Randomization by Humans." *The American Journal of Psychology*, 79(1), 97-103.

Reinikka, Ritva and Jakob Svensson. 2006. "Using Micro-Surveys to Measure and Explain Corruption." *World Development*, 34(2), 359-70.

Republic of Kenya. 2006. " Arid Lands Resource Management Project (Phase Ii) Tana River

District Progress Report 2003-2006," 1-61.

Schräpler, Jörg-Peter. 2011. "Benford's Law as an Instrument for Fraud Detection in Surveys Using the Data of the Socio-Economic Panel (Soep)." *Journal of Economics and Statistics*, 231(5/6), 685-718.

Stefanovic, Michael. Former head of Investigations at INT (World Bank Integrity Vice Presidency). 2018. Interview with Jean Ensminger, Email.

UN Economic and Social Council Economic Comission for Africa. 1986. "Adjustment of Errors in the Reported Age-Sex Data from African Censuses," *Joint Conference of African Planners, Statisticians and Demographers.* Addis Ababa, Ethiopia:

Walter R. Mebane, Jr. 2011. "Comment on "Benford's Law and the Detection of Election Fraud"." *Political Analysis*, 19, 269-72.

World Bank. 2011. "Forensic Audit Report: Arid Lands Resource Management Project - Phase Ii," Redacted: World Bank,

_____. 2003. "Project Appraisal Document on a Proposed Credit in the Amount of Sdr 43.6 Million (Us$ 60m. Equivalent) to the Republic of Kenya for the Arid Lands Resource

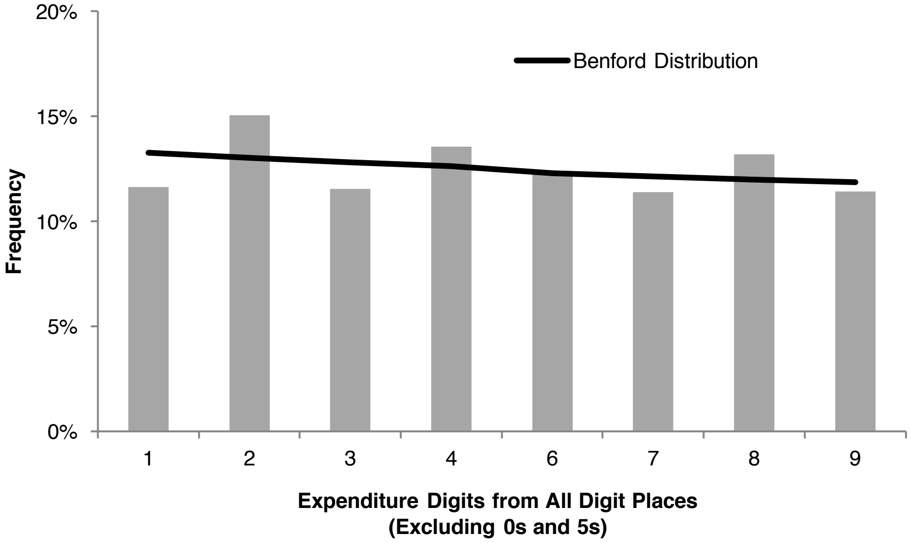
Management Project Phase Two," C. D. Eastern and Southern African Rural Development Operations, Africa Region, 31.

_____. 2007. "Project Appraisal Document on a Proposed Credit in the Amount of Sdr 57.8 Million (Us$ 86.0 Million Equivalent) to the Goverment of Kenya for a Western Kenya Community Driven Development and Flood Mitigation Project,"

World Bank Integrity Vice Presidency. 2018. "Redacted Investigation Reports," http://www.worldbank.org/en/about/unit/integrity-vice-presidency/redacted-investigation-reports. Accessed: 5/25/2018

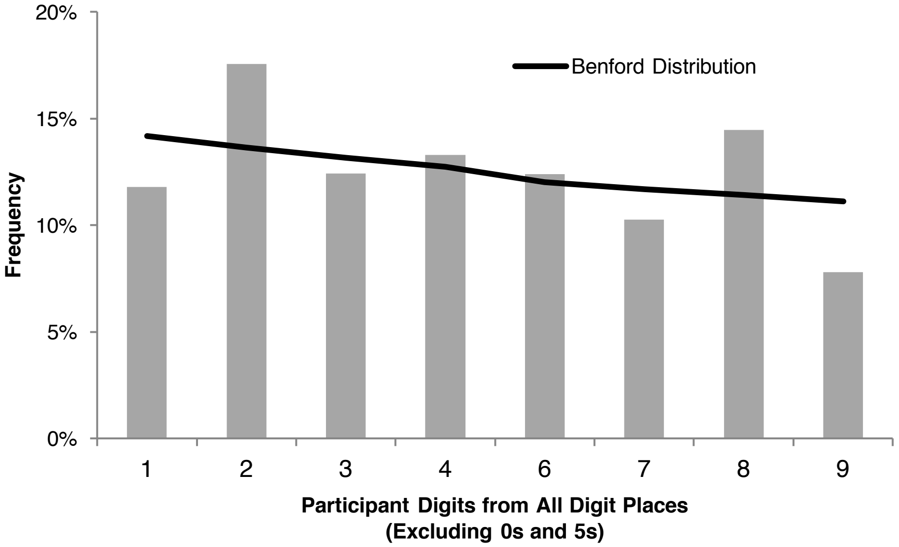
Figures and Tables

FIGURE 1: ALL DIGIT PLACES BEYOND THE FIRST AGAINST BENFORD'S LAW FOR EXPENDITURE DATA



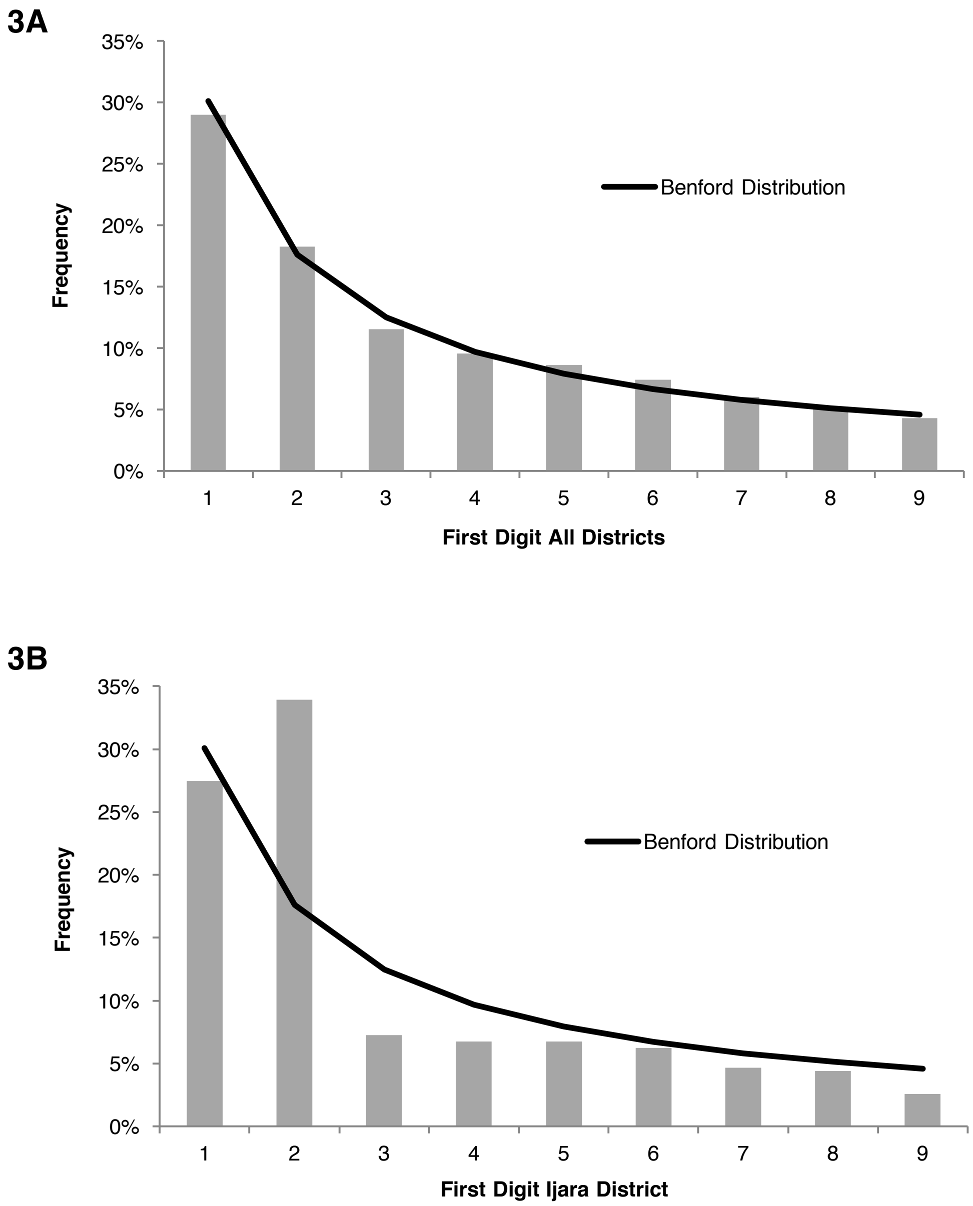
All districts combined ($p = 3.9 \times 10^{-15}$; $n = 9371$).

FIGURE 2: ALL DIGIT PLACES BEYOND THE FIRST AGAINST BENFORD'S LAW FOR PARTICIPANT DATA



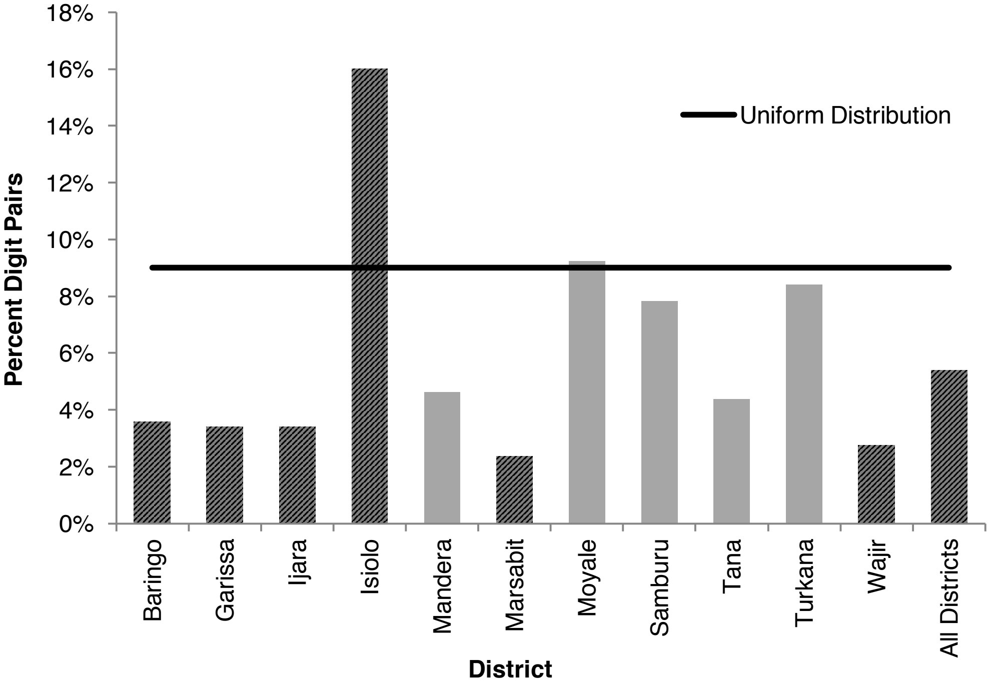
All districts combined ($p = 5.7 \times 10^{-51}$; $n = 7385$).

FIGURE 3AB: FIRST DIGIT EXPENDITURE DATA AGAINST BENFORD'S LAW



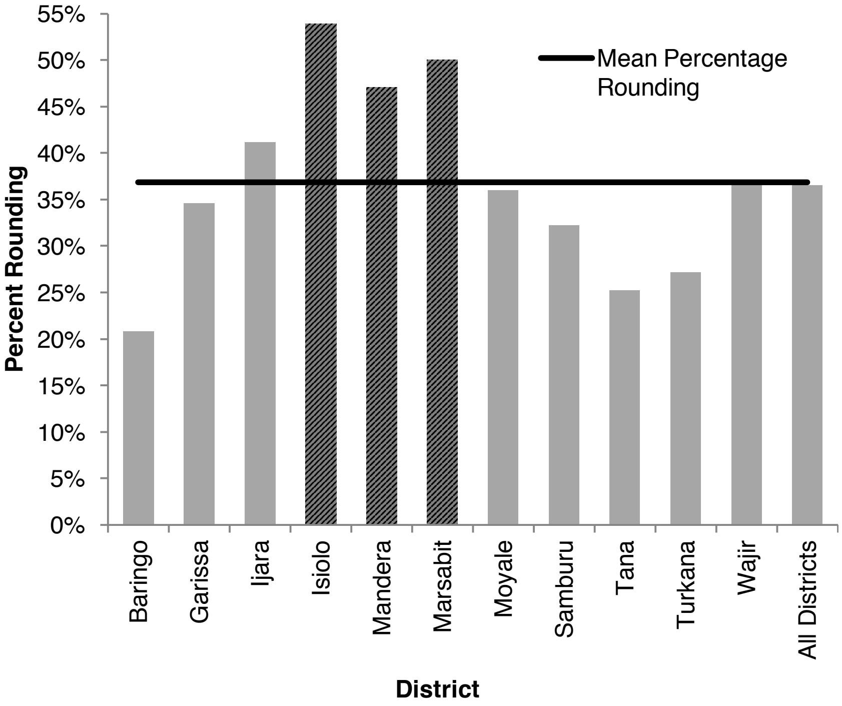
(A) All districts combined (*p* = 0.089; *n* = 4339). (B) Ijara District only (*p* = 2.3 × 10^{-13}; *n* = 386).

FIGURE 4: DIGITS PAIRS IN THE LAST TWO DIGITS FOR PARTICIPANT DATA BY DISTRICT



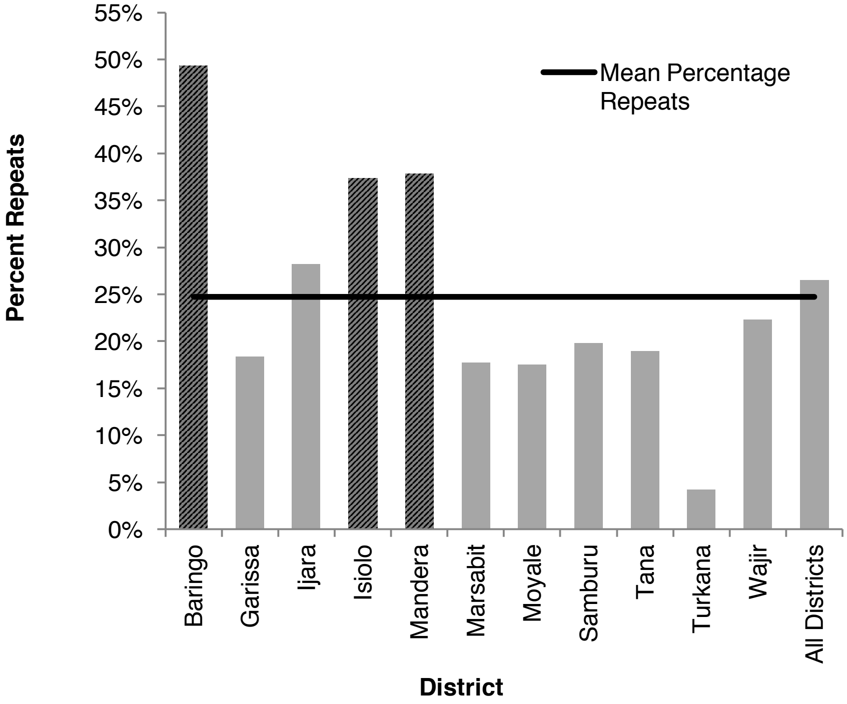
Crosshatched districts fail the binomial test at $p < 0.005$ by over or underutilizing digit pairs such as 11, 22, and 33 (Baringo p=8.2 × 10^{-4}, n = 251; Garissa p = 1.8 × 10^{-4}, n = 293; Ijara p = 0.0037, n = 176; Isiolo p = 0.0044, n = 125; Marsabit p = 0.0031, n = 126; Wajir p = 7.1 × 10^{-5}, n = 255; all districts p = 1.4 × 10^{-9}, n = 2059).

FIGURE 5: PERCENTAGE OF ROUNDED DIGITS IN EXPENDITURE DATA BY DISTRICT



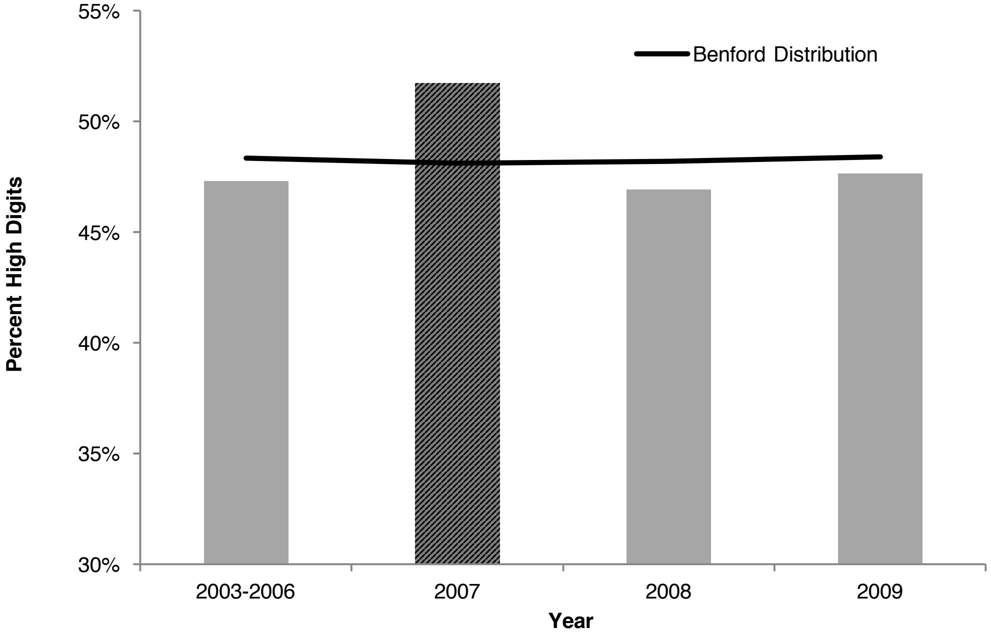
Percentage of digit places rounded in expenditure data by district. The three districts representing the top quartile are crosshatched.

FIGURE 6: PERCENTAGE OF REPEATED ENTRIES IN EXPENDITURE DATA BY DISTRICT



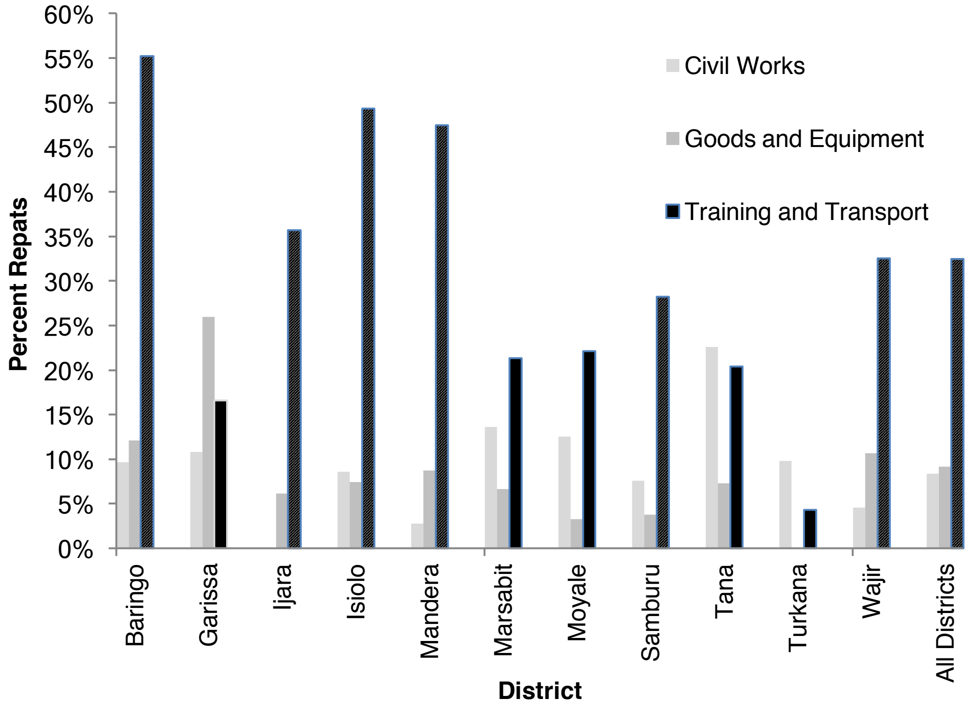
Percentage of exactly repeated expenditure entries by district for a given annual report. The three districts representing the top quartile are crosshatched.

FIGURE 7: ELECTION YEAR EFFECTS IN EXPENDITURE DATA



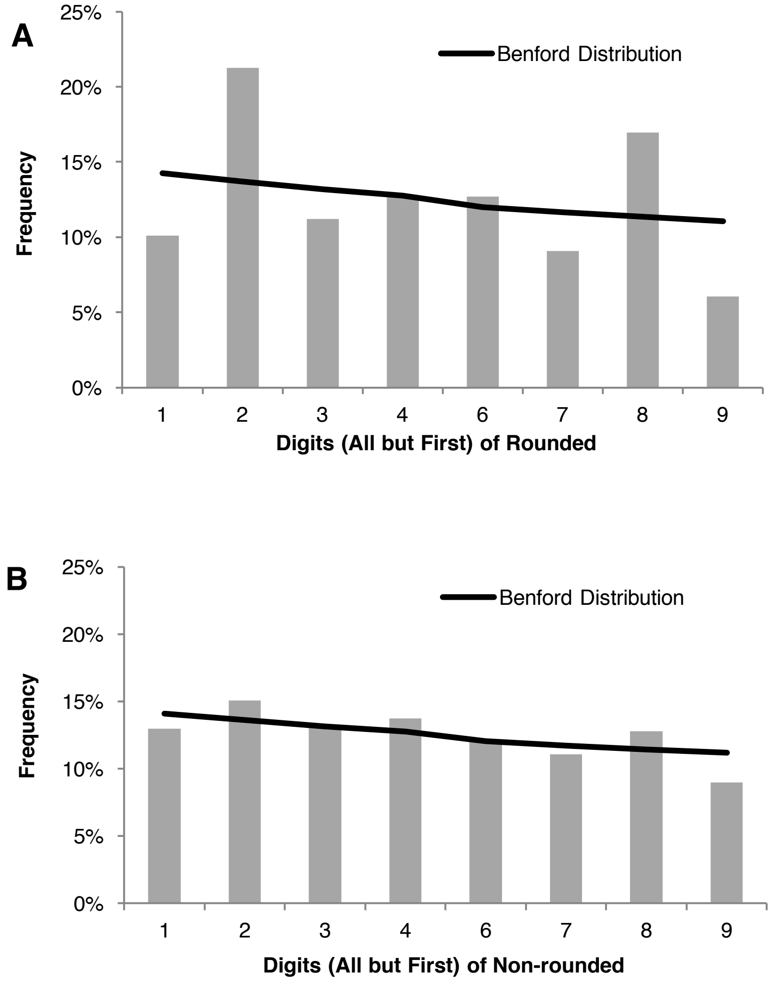
Percentage of high digits (6, 7, 8, 9 versus 1, 2, 3, 4) in all digit places but the first, for all districts, by year. 2007 was a Presidential election year. 2007 has a statistically significant presence of high digits in a ($p = 6.5 \times 10^{-6}$; $n = 1945$.)

FIGURE 8: SECTOR EFFECTS IN EXPENDITURES



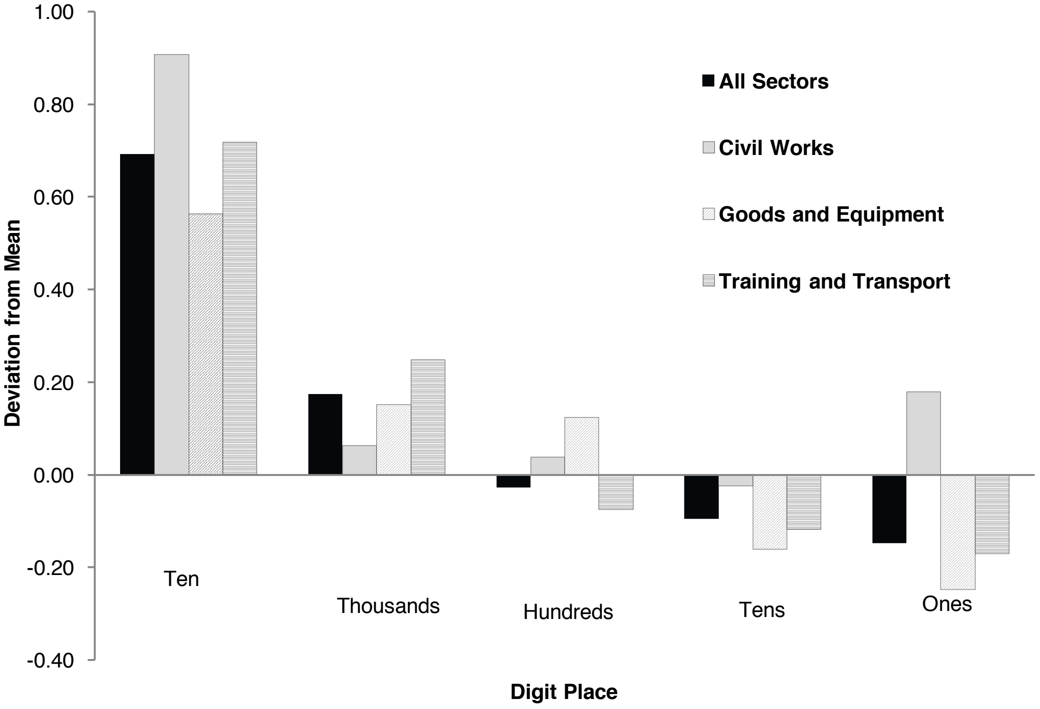
Percentage of line item expenditures repeated exactly, by district, year, and sector. Crosshatched districts report over three times as many repeated numbers in training and transport, versus other sectors, consistent with low-effort data manipulation.

FIGURE 9AB: UNPACKING ROUNDED AND UNROUNDED DIGITS IN PARTICIPANT DATA



(A) Digit breakout of all but the first digit (excluding 0s and 5s) when the total of male and female participants sums to a rounded number ($p = 2.6 \times 10^{-64}$; $n = 2975$). (B) Digit breakout of all but the first digit (excluding 0s and 5s) when the total of male and female participants sums to a non-rounded number ($p = 1.9 \times 10^{-6}$; $n = 4410$).

FIGURE 10: DEVIATION FROM BENFORD'S LAW MEAN IN EXPENDITURE DATA WITH MONTE CARLO SIMULATION



We compare the observed mean by digit place from the right to the Benford expected mean in each sector. Zero reflects conformance to the Benford expected mean. Positive values indicate the mean is higher than Benford's Law predicts. The observed pattern is consistent with a strategy of high digits in high digit value places and then underusing them in low digit value places to even out the digit distribution. We perform a Monte Carlo simulation of Benford-conforming datasets and compare our observed statistics to the simulated statistics to produce p-values. Compared to a sample of 100,000 simulations, using data from all sectors, we observe the following statistics: 10,000s place ($p = 1.0 \times 10^{-5}$), 1,000s ($p = 2.3 \times 10^{-4}$), 100s ($p = 0.33$), 10s ($p = 0.10$), 1s ($p = 0.061$).

TABLE 1: EXPECTED DIGIT FREQUENCIES UNDER BENFORD'S LAW

		Digit Place				
		1	2	3	4	5
	0	0.0000	0.1197	0.1018	0.1002	0.10002
	1	0.3010	0.1139	0.1014	0.1001	0.10001
	2	0.1761	0.1088	0.1010	0.1001	0.10001
	3	0.1249	0.1043	0.1006	0.1001	0.10001
Digit	4	0.0969	0.1003	0.1002	0.1000	0.10000
	5	0.0792	0.0967	0.0998	0.1000	0.10000
	6	0.0669	0.0934	0.0994	0.0999	0.09999
	7	0.0580	0.0904	0.0990	0.0999	0.09999
	8	0.0512	0.0876	0.0986	0.0999	0.09999
	9	0.0458	0.0850	0.0983	0.0998	0.09998

Source is Nigrini and Mittermaier (1997:54).

TABLE 2. SIGNIFICANCE OF DIGIT TESTS BY DISTRICT

Fig	Digit Test	Mandera	Isiolo	Baringo	Ijara	Wajir	Garissa	Samburu	Marsabit	Moyale	Turkana	Tana	All Districts
1	All Digit Places beyond the First: Expenditure	**3.6E-14** **846**	0.0082 437	**7.2E-17** **1352**	**2.6E-05** **769**	**1.9E-06** **1248**	**2.8E-08** **976**	0.020 848	**3.9E-04** **449**	**1.5E-14** **671**	0.40 907	**7.8E-04** **868**	**3.9E-15** **9371**
2	All Digit Places beyond the First: Participant	**9.0E-18** **886**	**6.1E-11** **478**	**2.1E-04** **674**	**5.9E-11** **765**	**6.5E-15** **731**	**6.2E-18** **858**	**2.3E-05** **639**	0.25 527	0.033 736	**0.0037** **591**	0.013 500	**5.7E-51** **7385**
3	First Digit: Expenditure Data	**1.4E-08** **489**	**5.5E-06** **308**	**1.4E-09** **488**	**2.3E-13** **386**	0.37 578	0.029 430	**5.7E-05** **359**	0.011 293	**1.9E-12** **319**	0.071 357	**0.0037** **332**	0.089 4339
4	Digit Pairs: Participant	0.0091 238	**0.0044** **125**	**8.2E-04** **251**	**0.0037** **176**	**7.1E-05** **255**	**1.8E-04** **293**	0.38 166	**0.0031** **126**	0.41 173	0.49 119	0.035 137	**1.4E-09** **2059**
5	Rounding Digits: Expenditure	**Top Quartile**	**Top Quartile**						**Top Quartile**				DNA
6	Repeating Numbers: Expenditure	**Top Quartile**	**Top Quartile**	**Top Quartile**									DNA
7	Year Effects (2007): Expenditure	0.18 117	0.22 98	0.0073 182	0.045 222	0.088 273	0.016 139	0.15 231	0.032 88	0.033 238	**0.0045** **192**	0.75 165	**6.5E-06** **1945**
8	Sector Effects: Expenditure	**> 3x**	**> 3x**	**> 3x**	**> 3x**	**>3x**		**> 3x**					DNA
9	Unpacking Rounded Numbers: Participant	**6.1E-21** **453**	**4.4E-13** **157**	0.0085 248	**1.2E-10** **298**	**7.6E-11** **433**	**5.9E-24** **459**	**3.9E-05** **179**	0.014 222	**0.0030** **179**	**3.1E-05** **205**	0.057 142	**2.6E-64** **2975**
10	Deviation from Mean in 10,000s Digit Place	**1.0E-05**	0.131	0.024	0.0054	**1.0E-05**	**0.0015**	**1.0E-05**	**1.0E-05**	**1.0E-05**	**1.0E-05**	**1.0E-05**	**1.0E-05**
	Number of Significant Tests p < 0 .005 (Out of 10)	8	7	6	6	6	5	5	4	4	4	3	6

We ran 10 digit tests on each of 11 districts. The tests were chosen to analyze different, non-overlapping aspects of the data. Given the large number of tests, a Bonferroni correction was used to establish 0.005 as the acceptable *p* – value for our tests. Failed tests at the 0.005 level are indicated in bold. For rounding and repeating (Figures 5 and 6), there is no theoretical means to establish the expected level and we work from the null hypothesis that there should be no significant difference between the districts. We flag the districts that are outliers in the upper quartile. Similarly, for the sector analysis (Figure 8), we flag the districts for which repeated numbers in sectors with higher risk of fraud (training and transport) are more than triple the level of other sectors (civil works and goods and equipment). We tabulate the number of significant tests for each district in the bottom row.

TABLE 3. DIGIT TESTS BY DISTRICT COMPARED TO WORLD BANK INT FORENSIC AUDIT RESULTS

	Digit Tests (Number Failed Out of 10)	INT Audit (Percent Suspected Fraudulent and Questionable Transactions)
Isiolo	7	74
Wajir	6	75
Samburu	5	68
Garissa	5	62
Tana	3	44
Mandera	8	Not Audited
Baringo	6	Not Audited
Ijara	6	Not Audited
Moyale	4	Not Audited
Marsabit	4	Not Audited
Turkana	4	Not Audited

Source for the INT forensic audit data is World Bank (2011).

Appendix

Appendix A: Last Digits

The literatures on both forensic auditing and election fraud emphasize analysis of the terminal digits, which should be uniformly distributed if they represent the fourth digit place or beyond (Beber and Scacco, 2012, Nigrini and Mittermaier, 1997). Results on the terminal digit are presented in Appendix Figure 1AB and show exceptional statistical significance for both expenditure and participant data. However, Benford's Law limits to the uniform distribution for digit places beyond the fourth, and therefore, tests of the final digit place are subsumed by the test of conformance to Benford's Law in our test of all digit places beyond the first.

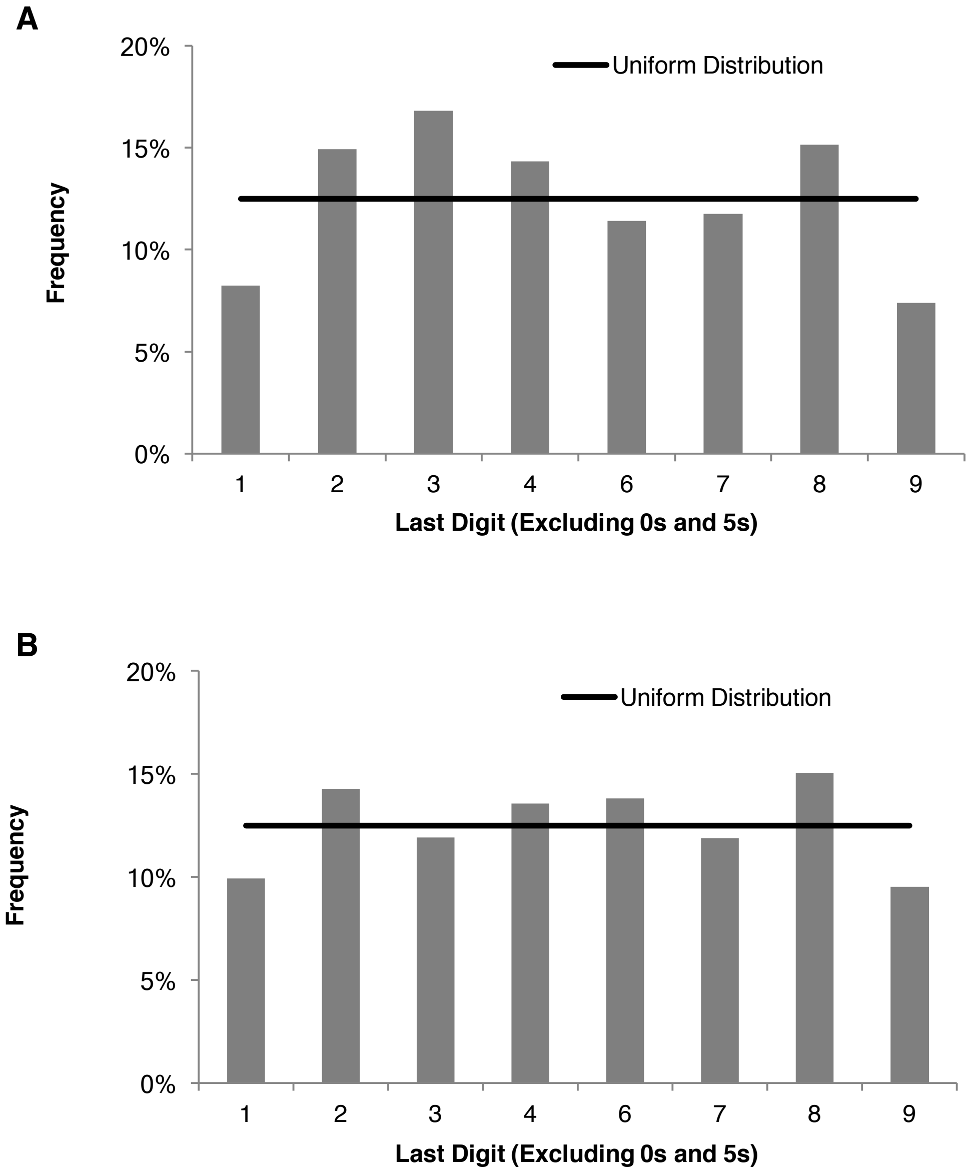
Appendix B: Difference of Means Statistics with Monte Carlo Simulations

We compute the mean by digit place in the last five digits from the right among 5, 6, and 7 digit numbers. We eliminate 0s and 5s from this computation and reweight Benford's Law as before. This gives us a mean for each of the 10,000s, 1,000s, 100s, 10s, and 1s digit places for those numbers that have all of these digit places. In each digit place from the right (1s, 10s, etc.), we compute the Benford expected mean as follows: for 5-digit numbers, the Benford mean in the 10,000s place is the mean of the 1st digit; for 6 digit numbers, the Benford mean in the 10,000s place is the mean of the 2nd digit; etc. For each number length and digit place from the right, we can compute an expected mean under Benford's Law.

We then combine our data from different string lengths, weighting the sample by how many numbers come from each length. This process gives us a mean of the digit place from the right, as well as an expected mean of the digit place from the right under Benford's Law. The difference in these values is the difference in means statistic. Positive values indicate a weighted mean that exceeds the weighted Benford's Law, indicating padding. Negative values indicate a weighted mean that is below the weighted Benford's Law, indicating overuse of low digits. In order to determine significance of each of our statistics, we perform a Monte Carlo simulation. We generate 100,000 datasets that are identical to the digit lengths observed in our dataset. Code for simulating Benford-distributed numbers was used with permission (Debowski, 2003). Code for matching Benford-conforming numbers with the lengths of our data was produced in Python. For each simulated dataset, we remove 0s and 5s and compute the means by digit place from the right as well as the Benford expected mean, identically to the above. For each of the 100,000 datasets, we produce a difference of means statistic. We then compare our observed difference of means statistic to these simulations. The p-values reported are the empirical cumulative distribution function (CDF) of our difference of means among the simulated statistics. That is, if our statistic exceeds 90% of the simulated values, its p-value is 0.10. Because there are 100,000 samples, there is a minimum p-value of 1 in 100,000.

Appendix Figures

APPENDIX FIGURE 1AB: LAST DIGIT EXPENDITURE AND PARTICIPANT DATA AGAINST THE UNIFORM DISTRIBUTION.



(A) Expenditure data ($p = 1.5 \times 10^{-9}$; $n = 851$). (B) Participant data ($p = 7.0 \times 10^{-26}$; $n = 5850$).