GEORGETOWN UNIVERSITY
McCourt School of Public Policy
MASSIVE DATA INSTITUTE

April 2024

The Challenges of Identifying Medical Debt in Court Records

James Carey, JD

Policy Fellow, Massive Data Institute

Margaret Haughney, MPH

Policy Analyst, Massive Data Institute

There is a medical debt crisis in America, but it is difficult to see due to a lack of data, especially around one of its greatest nexuses: our courts. This report examines the landscape of research on the impoverishing effects of medical debt and the hurdles to data-centered research in the field. Acknowledging that courts are the focus of a large amount of medical debt collection, it explores the challenges in using court data to investigate medical debt collection. One of the largest roadblocks is difficulty in identifying medical debt in court data, and this report evaluates good methods of doing so with low overhead. One potentially successful method was developed for the report and focuses on using expert consultation from a physician to identify plaintiffs.

The medical expertise of Dr. Fatu S. Conteh was invaluable in developing this method of identification, as were the data policy skills and work of Stephanie Straus and Jordan Rinaldi of the Massive Data Institute, as well as the guidance of Dr. Amy O'Hara of the Massive Data Institute and Professor Tanina Rostain of Georgetown University Law Center.

This research in this report was conducted under the Civil Justice Data Commons, a joint project between the Georgetown University Law Center and Georgetown University's Massive Data Institute that aims to create a secure, robust repository for civil legal data gathered from courts, legal services providers, and other civil law institutions.

This report was supported by a grant from the JPB Foundation.

Contents

Introduction

America has a crushing medical debt crisis. One in ten Americans owe significant medical debt, and one in a hundred owe medical debt over $10,000.[i] Medical debt contributes to racial inequity and can exacerbate poverty, and it can place even the most well-off in vulnerable positions due to its overwhelming nature.[ii] Medical debt can also weigh on Americans through constant bills and bad credit, requiring taking on another job or extra work hours and inducing general anxiety or worry. The impact of medical debt on community members has become a large enough problem that local governments around the U.S. have programs or proposals to purchase and forgive residents' medical debt.[iii] New York City has budgeted $18 million over three years to cancel $2 billion of medical debt, impacting up to 500,000 residents, through a partnership with the national nonprofit RIP Medical Debt.[iv]

A substantial amount of this medical debt is collected through lawsuits in state courts, but it is difficult to tell how substantial. The lack of state court data available to researchers makes it a hard question to answer. Even when state court case data are accessible, knowing which of those cases are medical debt collections is often difficult. The first step in researching how medical debt moves through courts is to be able to identify it.

This report examines the impoverishing effects of medical debt, providing an overview of research on the crisis, particularly work centered on data. It then details why court data is a key piece of this research, but also why it is a difficult resource to utilize. The report evaluates various methods of identifying medical debt in court data, which would unlock research. This evaluation includes a novel method that we developed which utilizes physician consultation in medical debt collections plaintiff identification, which performs well compared to other methods and may be promising for future research.

I. The Impoverishing Effects of Medical Debt

Medical debt is the most common type of debt tradeline found on credit reports, and the estimated amount of medical debt in collections ranges from $80 billion to $140 billion.[v] Recent research estimates that over a third of U.S. adults held a medical debt balance in 2021.[vi] [1]

Beyond the initial financial burden of a medical bill, patients may experience an array of negative consequences to their physical, emotional, and financial wellbeing. Patients with medical debt anecdotally report delaying medical care for other issues to avoid additional bills,[vii] and a 2023 Commonwealth Fund survey found that over a third of adults with medical or dental debt delayed or avoided seeking care or medication specifically because of their existing debt.[viii] In a 2016 Kaiser Family Foundation/New York Times survey, those with medical bill problems postponed health care at a rate almost three times greater than those with no medical bill problems.[ix] [2] Nearly 60% of adults in the Commonwealth Fund survey reported their health problem worsening after they delayed or skipped care due to cost.[x] Beyond medical care choices, people with medical debt sometimes have to make tough decisions between paying off their debt or paying for necessities; nearly 40% of adults with medical or dental debt had to decrease spending on their food, heat, or rent.[xi] Furthermore, those with medical debt have a higher risk of personal bankruptcy.[xii] It is not difficult to understand how one bill can result in bankruptcy when monthly payments put patients in the red.[xiii] Moving on from medical debt and bankruptcy can also be complicated by lower credit scores that result from those financial situations. Over a third of adults with medical or dental debt

[1] Although some studies report an overall decline over the past decade in debt metrics—such as mean debt in collections (Kluender, *end note v.*) or being in a family that has had problems paying medical bills (Robin A. Cohen & Emily P. Zammitti, Division of Health Interview Statistics, NCHS, *Problems Paying Medical Bills Among Persons Under Age 65: Early Release of Estimates from the National Health Interview Survey, 2011-June 2016*, NATIONAL CENTER FOR HEALTH STATISTICS (2026), available at http://www.cdc.gov/nchs/nhis/releases.htm)—the effects of medical debt are still felt strongly by patients struggling to navigate the medical and legal systems.

[2] Even in surveys of the general public, the percent of adults reporting that they have avoided or delayed medical care in the prior year due to costs ranges from 25% to 64% [Lunna Lopes, Alex Montero, Marley Presiado, & Liz Hamel, *Americans' Challenges with Health Care Costs*, KAISER FAMILY FOUNDATION (Mar. 1, 2024), https://www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/; Megan Brenan, *Record High in U.S. Put Off Medical Care Due to Cost in 2022*, GALLUP (Jan. 17, 2023), https://news.gallup.com/poll/468053/record-high-put-off-medical-care-due-cost-2022.aspx; Collins (*end note viii.*); Ann Waller Curtis, *New Care Payment Research Shows Americans Can't Afford Their Medical Bills*, BUSINESSWIRE (Feb. 14, 2018)].

said they received a lower credit rating, which can negatively impact getting a loan or a job. A single medical bill can create a cycle that hinders the ability to take care of oneself and one's family, stay healthy, maintain a stable income or pursue education or a higher-paying job, and ultimately be able to pay off debt.

Exacerbating the effects of medical debt are the disparities among lawsuits, judgments, and garnishments for those who are younger, Black, Latin/Hispanic, of low income, and located in the southern U.S.[xiv] These groups are more likely to have medical debt and to experience a greater frequency of lawsuits and wage garnishments. Waldman and Kiel reported medical debt plaintiffs sued defendants in majority-Black Census tracts over, on average, four times the rate they sued in majority-white tracts.[xv] In 2020, the Consumer Financial Protection Bureau found that "the mean medical debt in the lowest-income zip code decile in the U.S. was $677—more than five times higher than the mean medical debt in the highest-income zip code decile, which was $126." If a lawsuit results in wage garnishment, the extent to which wage garnishments can affect low-income earners is particularly extreme. There are accounts of paychecks, after garnishment and other deductions, totaling less than a dollar for over 60 hours of work.[xvi]

Some research suggests that being insured may not make a significant difference in medical debt. Studies following Medicaid expansion in 2014 showed reductions in medical debt in collections and new debt listed on credit reports while other types of debt remained similar between expansion and non-expansion states.[xvii] Himmelstein et al., however, found no difference in rates of bankruptcy due to medical debt between expansion and non-expansion states.[xviii] Their analysis could indicate that Medicaid expansion is insufficient to address how medical debt intersects with the insurance gap. This may be due to debt from out-of-network providers or services not fully covered under an insurance plan.[xix] High-deductible health plans (HDHPs) may also be a cause, as they can force patients into paying thousands of dollars in deductibles (as high as $7,900 in 2019) before their insurance starts covering costs.[xx] [3] More research is warranted to make strong claims about how Medicaid expansion and being insured affect medical debt.

[3] HDHPs been on the rise for several years (Cohen (*end note xx.*)). Even a hospital executive cited high deductibles as cause for reconsidering their hospital's payment policies (Jay Hancock & Elizabeth Lucas, *VCU Health Will Halt Patient Lawsuits, Boost Aid in Wake of KHN Investigation*,

II. Hurdles to Court Debt Collection Research

Researching medical debt in court cases is difficult due to a lack of available data and challenges in identifying medical debt cases from among the court data that are available. State and local court data are infamously unavailable for research.[xxi] Civil court data availability is more limited than criminal data and, within that narrowed subset, medical debt data are yet scarcer. Medical debt data face the limitations of the civil justice data they are a part of, as well as further complications. Civil justice data are lacking because there are important aspects of civil justice where records are not kept. Courts often do not record information such as demographics of parties or the origin of the debt. Conducting medical debt reform is also hampered by the lack of standardized and clear policies for health care institutions to offer financial assistance, communicate with patients who have medical bills, and collect debt. Different locales, sometimes even across hospitals in the same city, implement vastly different financial assistance, billing, and debt collection practices. Members of the American Hospital Association follow Internal Revenue Service guidelines, but the guidelines only require that hospitals have a financial assistance policy and that they make "reasonable efforts" to identify a patient's eligibility for assistance before they start debt collections.[xxii] Because federal medical debt protections are ineffective, states can implement protections to fill the gap.

Two institutions have recently released comprehensive resources that make it easier to review medical debt policies in the United States. The Commonwealth Fund report reviews the federal and state protection standards.[xxiii] They also created a supplemental interactive map that provides a state-by-state bulleted listing of what policies are in place for five categories: financial assistance standards, community benefit requirements, billing and collections, lawsuits, and reporting requirements.[xxiv] While The Commonwealth Fund laid out what currently exists, Innovation for Justice uses their Medical Debt Policy Scorecard to assess current policies against what they hope to achieve.[xxv] They assigned each state a ranking based on the extent

KFF HEALTH NEWS (Oct. 9, 2019), https://kffhealthnews.org/news/vcu-health-will-halt-patient-lawsuits-boost-aid-in-wake-of-khn-investigation/). In 2006 close to half of employer-sponsored health plans had a deductible, but in 2019 over 80% had a deductible (Kaiser Family Foundation, *2019 Employer Health Benefits Survey*, § 7 Employee Cost Sharing, KAISER FAMILY FOUNDATION (Sept. 25, 2019), available at https://www.kff.org/report-section/ehbs-2019-section-7-employee-cost-sharing/). The average deductible amount also increased in 2019, to over $1600, almost triple what it was in 2006. Unfortunately, HDHPs are also more common among lower-income workers because the plans are less costly for employers to provide.

to which a state meets a policy idea intended to advance a medical debt objective. For example, one objective is to "reduce the negative consequences for debtors after court", and implementing a policy to deny seizures of bank accounts would contribute towards achieving that objective. If a state does not have a policy addressing that, then they receive no points. Both of these scorecards provide a picture of each state's policies but are unable to show the effects of those policies on the ground.

The research which would show effects of policy changes seems to be dominated by investigative journalism and reports published by think-tanks or non-profits rather than peer-reviewed research studies. Additionally, not all of these reports and journalistic articles provide detailed methodology. From the reports that include their methods, it is clear that court case data lack the array of variables that would enable greater depth of research. Much of the peer-reviewed research using court data focuses mostly on statistics surrounding the number of lawsuits filed in a certain geography, the amount of debt under dispute, the amount requested for garnishment, and what variables may contribute to disparities. Going beyond the basic statistics would require courts to collect more information, change their methods of collection, or link their data to other administrative datasets.

For researchers, civil justice data are both inaccessible and lacking. The data are inaccessible because, where data exists, researchers cannot get at them. Courts keep records, but access to those records is often limited by the rules of the court (including charging large fees for access) or the media in which they are stored (such as paper records or antiquated electronic systems). Different data elements may have different accessibility in the same jurisdiction, e.g., some case information may be stored digitally while other information on the same case is stored in paper records. Courts also vary in what variables they collect in their documentation, often missing data that would provide greater understanding of how individuals experience medical debt in the court system. Cooper et al. were able to get "information on the plaintiffs and defendants, the filing date, the filing county, the court's decision, the judgment amount, and the court fees", but the court data lacked information on defendant race/ethnicity and whether an individual lawsuit was related to unpaid medical bills or another issue with the hospital.[xxvi] Ericson and Gross were able to collect the case filing date, hospital suit filing date, and the complaint amount from Maryland court case data.[xxvii] But because Maryland court data does not provide information on hospital collection rates, hospital contracts with attorneys, administrative costs, or revenue that hospitals received from each case, they had to estimate net revenue based on the complaint amounts. When that information is not recorded, it does not exist in the data. Minnesota case files contain the original amount in controversy, original creditor name, default judgment disposition, writ of execution, fines, and fees.[xxviii] However, the Minnesota State Bar Association Access to Justice Committee had to impute demographic information (e.g., race, ethnicity, and income) based on names and addresses because the civil courts do not collect demographic data. The type of debt is also not recorded in court documents; for the Committee to identify medical debt cases, they had to hand-review documents and classify the debt based on the original creditor. If researchers wish to explore garnishments or satisfaction of debts in Minnesota, they will find the court data insufficient. The garnishment process takes place outside of the court; rather the plaintiff's attorney has the duty of serving garnishment summons and does not have to file any record of it with the court. If a consumer decides to object to the garnishment, that paperwork also is sent to the plaintiff's team and not the court. When partial or full debt payments are made, reporting on satisfaction happens in less than half of debt lawsuits in Minnesota. This can extend negative consequences for consumers' credit because there is no filing in court records that documents their debt payment. This data gap in debt case outcomes also limits what research can be done and how much evidence is built to inform new court policies.

These two issues can also overlap. For example, key information such as the amount in controversy in a case may only be recorded in the plaintiff's pleading (the document filed to start the case) or in the judgment. In some jurisdictions, these documents may be accessed only by court employees and not outside researchers. When they are available to external parties, the data may be only available in pdfs or paper documents at the court and, thus, not machine-readable for data purposes.

Researchers for the Community Foundation of Greater Chattanooga were able to obtain case filing dates, case types, party names, defendant addresses, judgment amounts, and other details about service and garnishments from docket data via Tennessee Case Finder.[xxix] However, they had to hand-review court documents to get original claim amounts, requests for attorney's fees and post-judgment interest, and identities of original creditors in debt buyer cases. Another limitation in the Tennessee court files was the lack of details on whether the Civil Summons and other documentation were properly served to the correct person. Proper service is necessary for the defendant because the documentation notifies them of the lawsuit and the date, time, and location of the hearing. Improper service means that people may not even know there is a case against them until after garnishment has started or a lien has been placed on their property. Other states require officials to document the service method in greater detail, which can provide researchers with more insight into what is actually happening in the litigation process. Moreover, the documentation in Hamilton County sometimes provided very limited information on the actual lawsuit–only identifying the debt collector's name and the debt amount–which is unhelpful for the consumer to verify basic information on the case, such as the origin of the debt, whether the amount is correct, or if they actually owe anything at all.

States like North Carolina and Texas require debt collectors or buyers to provide proof of the validity of the debt, providing protection for consumers and an extra barrier to wasting time and resources of the court. In states like Minnesota that allow hip pocket filing, plaintiffs can serve consumers but do not need to provide proof that the debt is valid.[xxx] Only when a judgment occurs (either after the consumer responds or starting 21 days later with no response from the consumer) will the court issue a case number or require proof of valid debt. Moreover, proof of validity is only required from debt buyers in Minnesota, and those documents do not have to be provided to the consumer.

As medical debt cases are a subset of civil justice cases, the same barriers to access apply, and they are also inaccessible and lacking. However, even if those data-specific hurdles are overcome and access to civil justice cases is possible, researching medical debt cases is often extremely difficult because it is challenging to determine which of those civil court cases are medical debt cases.

Medical debt collection cases are rarely categorized as such in court data available to researchers. In our survey of court data made freely available for researchers by courts, only two states categorized medical debt cases at all: Arkansas and Connecticut. Arkansas uses a case type category tag for consumer debt that arises from medical expenses.[xxxi] Connecticut's categorization is more limited, as they label small claims debt collection (with separate categories for medical debt arising from a hospital vs. non-hospital source), but only small claims cases and not larger suits.[xxxii] Instead, they have a larger catch-all category for all collections. Although having some degree of categorization is valuable for these states, research into their medical debt cases is still hampered by the general difficulties of researching civil court data.

The other 48 states (and the District of Columbia) lack medical debt categorization of their researcher-accessible court case data. These states may have internal medical debt case categories, but these are not accessible for outside researchers. Even if there are states that may be open to providing those data to external researchers, it is difficult to target those states for research without extensive prior discussion with individual states.

Without official categorization, identification by researchers of which cases are medical debt collections is a vital step to understanding the extent of medical debt collection occurring in America's courts.

III. Methods for Identifying Medical Debt

Once access and authorization to research court data has been secured, a method of identifying which records in that court data are medical debt (med debt) collection is required. Focusing on efficient methods of string-based identification, we examined how different methods, focused on examining plaintiff names, performed when run on a verified dataset of med debt cases. Then, we ran those same methods on a larger dataset of debt cases, which included both med debt and other potential case types, and evaluated the methods on both number of records identified and accuracy. Several of the methods we looked at have been used previously in various settings to identify either med debt or debt cases as a whole. We also developed and examined a novel method, in conjunction with our physician collaborator, which focused on using subject matter medical knowledge to identify likely plaintiff names. The results of our case study indicate that this novel method was more successful than previously used methods at identifying med debt cases for the tested dataset (Connecticut court cases), and employing similar methods to this physician consultation method may prove fruitful for low-overhead court med debt identification in the future.

For many researchers, especially those with a background in policy instead of data, it is also important that this method have computation and labor efficiency. With that in mind, our research into identifying medical debt from court records focused on methods which do not require advanced computing power or specialized infrastructure aside from basic statistical programing tools. Researchers familiar with handling data in the Python or R programming languages should be able to employ these methods on a cloud computing platform such as the Civil Justice Data Commons or even a personal computer with appropriate batching for large datasets. The methods here are less complex and promise less revolutionary results than techniques such as Probabilistic Soft Logic or Topic Modeling, but they are more accessible for more research with sparser data, which we see as a valuable feature in med debt identification.[4]

[4] Variations and precursors to the methods used here, as well as more complex methods of identification, were highlighted at the Civil Justice Data Commons Clustering & Classifying Methods Convening in September 2022 [James Carey, *Convening of Civil Justice Experts on How to Cluster Data*, CJDC Blog (Nov. 11, 2022), https://www.law.georgetown.edu/tech-institute/initiatives/georgetown-justice-lab/civil-justice-data-commons/cjdc-blog/convening-of-civil-justice-experts-on-data-clustering/]. These included a Georgetown and University of Southern California collaboration on a Probabilistic Soft Logic method, which utilized the computing power of USC's Information Sciences Institute [Yixiang Yao, Weizhao Jin, Srivatsan Ravi, *Labeling without Seeing? Blind Annotation for Privacy-Preserving Entity Resolution*, arXiv:2308.03734 (Aug. 7, 2023), https://doi.org/10.48550/arXiv.2308.03734].

For each of these plaintiff name-based string identification methods, a matching algorithm was used to compare a set of strings to plaintiff names in the court case data.[5] The major difference between the different methods is how that set of strings was derived. Those methods are described below.

For this research, we used two different matching algorithms to attempt med debt identification. The first is a simple substring-based search.[6] In this algorithm, the method's set of strings is compared to all substrings in the potentially matching plaintiff name. If there was a match found, the case record is marked as identified. The second algorithm is a fuzzy matching algorithm based on Levenshtein distance, where a similar substring comparison occurs but the algorithm allows for the possibility of changes in the string, including substitution, insertion, or deletion of characters. This results in fuzzy matches, which do not have to match the searching string exactly and can allow for typos or other differences in how a string is recorded. In our case study, a similarity to the search string of 95% was required. This threshold is standard in fuzzy matching, though future research could experiment with different thresholds, or indeed fuzzy matching based on distances other than Levenshtein. These two algorithms are referred to as "simple" and "fuzzy" below.

The Methods

Physician Consultation Method

Our novel method of plaintiff name-based identification focuses on a set of strings derived through an iterative process with the help of a subject matter expert, physician Dr. Fatu S. Conteh. This process focused on court records from Oklahoma and built upon concepts used by Open Justice Oklahoma for their own med debt identification.[7] Dr. Conteh worked in conjunction with Massive Data Institute data science experts to iterate on R code which developed and tested this set of strings on the Civil Justice Data Commons' Oklahoma Datasets.[xxxiii] These iterations, which took place in conjunction with a series of

[5] In programming and data science, "strings" refers to strings of text characters, which could form parts of words, word themselves, or collections of words.

[6] A "substring" is a subset of the text characters in a string.

[7] Open Justice Oklahoma presented on these methods at the Civil Justice Data Commons Clustering & Classifying Methods Convening (*see footnote 4.*) and have made their R code for the process publicly accessible [medical-debt Repository, Open Justice Oklahoma Github, https://github.com/openjusticeok/medical-debt (last accessed Apr. 10, 2024)].

monthly consultations and joint programming work, were done on expanding sets of data and incorporated specific strings that, based on the Physician's expert knowledge, were likely to include med debt. The resulting set of strings included terms which are substrings of common organization names that med debt cases may be filed under but were honed to maximize matches while minimizing false identification. Examination of the identified cases in each iteration also resulted in the determination of which cases were being falsely identified. The set of strings was pruned to minimize these cases, and an 'excluded' set of strings was also created. The matching algorithm takes this excluded set of strings and excludes any resulting matches from being marked as med debt. The sets of strings created by this process were then exported to be used with matching with other datasets, as we did in this case study.

Machine Learning Generated Terms Method

Researchers at the University of Guelph have worked in the Civil Justice Data Commons to conduct similar plaintiff name-based medical debt identification based not on the input of a subject matter expert but instead machine learning.[xxxiv] [8] Their work involved running machine learning Natural Language Processing algorithms on known med debt cases to identify common substrings in them, which were then used as signifiers for med debt in other data. Like our physician consultation method, this machine learning ("ML") method was initially developed using data from the Civil Justice Data Commons' repository of Oklahoma court data. Like our physician consultation method, an 'excluded' set of strings was also developed, to help prevent false matches. The sets of strings that were identified by machine learning were then exported to be used with string matching algorithms. Although the sets of strings were generated by machine learning, there is no active machine learning iteration involved in their use. This lowers the technical and computational overhead of their use, but it also means this method is less adaptable than active Natural Language Processing run directly on a target court dataset.

Manual Review of Top Cases Method

One especially low-overhead, though potentially labor-intensive, method of identifying medical debt plaintiffs is through manual review. January Advisors, data science consultants who frequently work on court data, used this method in their work on consumer debt in Michigan.[xxxv] They identified which plaintiffs

[8] Eric Sartor and Julia Hohenadel from the University of Guelph worked on this method, which was also presented on at the Civil Justice Data Commons Clustering & Classifying Methods Convening (*see footnote 4.*).

were most likely to be the source of consumer debt by conducting a manual review of the top 100 filers of civil cases, flagging those that appeared to be collecting on consumer debt.

For our plaintiff name-based identification of med debt, we used a similar process as this method. We reviewed the top 100 and top 1000 filers of our civil cases for the target jurisdiction (Connecticut civil cases for our case study). This manual review identified the names of the plaintiffs who were filing the greatest number of cases and, based on manual review by our researchers, high volumes of cases were filing medical debt collections among the frequent-filer plaintiffs. Those manually identified plaintiff names were then used as a set of strings for matching. As might be expected, these strings were less pruned and honed than those found during other methods, as they were directly taken from the plaintiff names in the case records. However, they do have the advantage of capturing plaintiffs with names that do not appear to be medical organizations based on their face, but local knowledge gained from examining cases indicate that they are in fact a medical organization filing med debt claims. An example of this would be a plaintiff called Local Hospital Network which filed cases under the name "LHN." The manual review for this method took several hours of focused legal research, which would be the same for every new jurisdiction it is performed on, but had lower overhead than the months of physician-based consultation in our novel method. For this method, there were no excluded sets of strings.

CMS Facility Names Method

Finally, we used records of medical facility names provided to the public by the Centers for Medicare & Medicaid Services as a set of strings, narrowed to fit our target jurisdiction.[xxxvi] This method uses CMS's register of hospital facilities similar to the way the manual review method uses top plaintiff names. The downside of this method is that the names that facilities register with Medicare may not match with the name they use when filing a med debt claim. However, it has the benefit of being an authoritative list that includes many facilities which may not be in the top 100 or top 1000 filers of civil cases by volume in a jurisdiction. Future research could explore the CMS Medicare register as a resource, going beyond hospital facilities and also including individual providers, though research would be required to ensure overmatching is pruned. For this method, there were no excluded sets of strings.

Case Study[9]

For this report, our case study was Connecticut court data in the Civil Justice Data Commons. As mentioned above, Connecticut is one of the few states that has any researcher-accessible data, which includes a case type for medical debt. This made it an ideal dataset to test different identification methods, as they could be compared against this verified subset. However, Connecticut only has a med debt case type for small claims, i.e., cases under an amount in demand of $5,000.[xxxvii] This means there is also a wider range of possible cases that are med debt collection but categorized as other debt case types because their amount in demands exceeds the threshold for small claims. We used the methods detailed above to attempt to identify these additional cases, once they had been tested against the verified medical debt.

For each method, we ran the set of strings with both the simple algorithm and fuzzy algorithm. For those methods which included a set of excluded strings, we ran fuzzy matching both with and without these set of strings. Some of the excluded strings were derived with the intent of preventing matches of strings that are close to the included strings but do not represent medical debt. When these are used with the fuzzy matching, in some cases they may be too close to the included set of strings and, thus, result in missing accurate identifications.

Figure 1 shows the number of case records identified using each method/algorithm combination when ran only on the verified medical debt cases, based on case type. The novel physician consultation method, when combined with fuzzy matching, identified the greatest number of cases at 80%. This, alongside the method used with the other algorithms, shows a marked increase in cases identified over the other methods. The Machine Learning-based method identified around a third of all cases when used with fuzzy matching, and other methods did less well.

[9] The code used in this case study is available at on the Civil Justice Data Commons GitHub at https://github.com/Civil-Justice-Data-Commons/Case_Studies_in_Medical_Debt_Identification/.

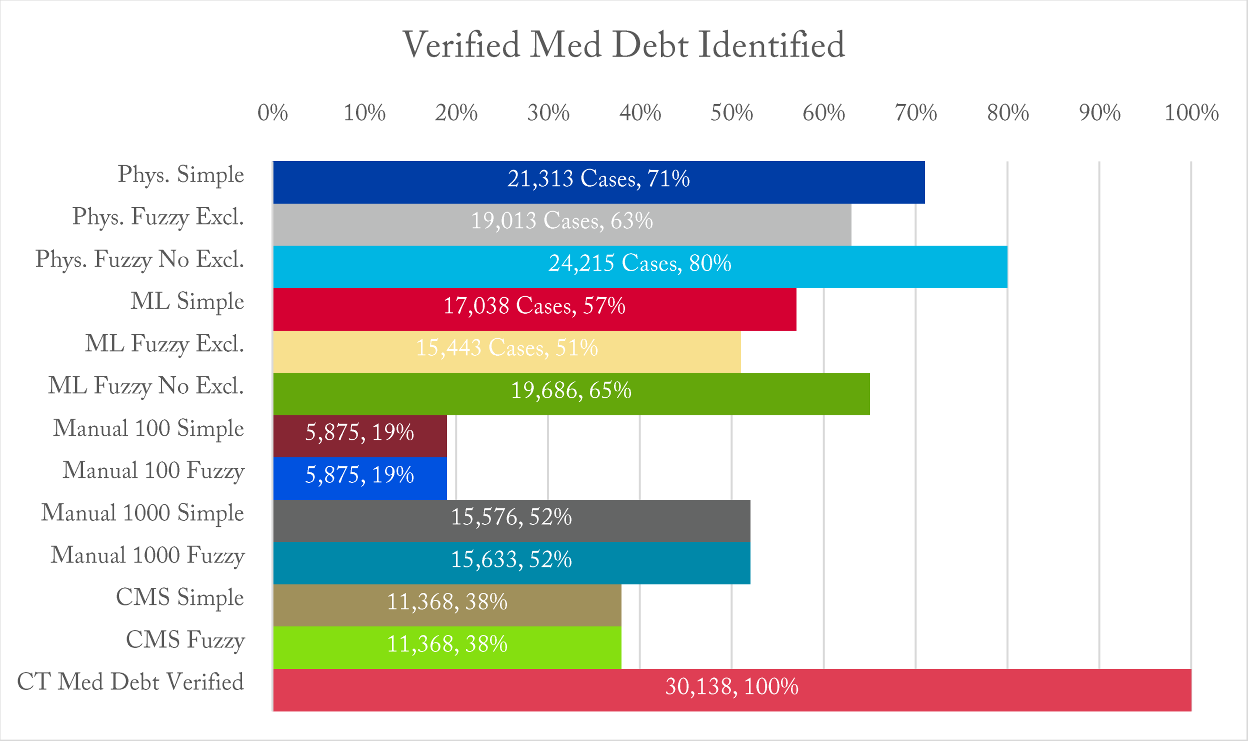

Figure 1: Verified Med Debt Identified

Table 1, included at the end of this report, included details on all overlaps between identified records. From this table, it can be seen that there is a significant overlap between all other methods and the cases captured by the physician consultation method. This indicates that the method is capturing most of the cases the other methods are, in addition to cases that they are not. However, there is an approximate 15% subset of cases which the machine learning methods identify but the physician consultation method does not.

Figure 2 shows the number of verified cases each method identified, separated by year. The Connecticut dataset is not of constant size year-over-year and is less complete for some time periods, so this figure should not be taken to indicate that there are no med debt cases in the early 2000s. What it does show, however, is that general trends in case numbers are consistently reflected across the methods, even if the magnitude of cases is not.

In the Connecticut case data, small claims cases include the amount demanded (which is always less than $5,000 for small claims). Using this data element, we can examine the average dollar value of the cases each method identified. Figure 3 shows this examination. Based on this data, it appears that most methods slightly overestimate the average dollar amount. This indicates that plaintiffs with names that are more recognizable as medical organizations may be filing high dollar value claims. Once again, the physician consultation method combined with fuzzy matching gets closest to the actual average from verified cases. However, all methods have their average clustered close to $1,500, their Q1 close to $1,000, and their Q3 close to $2,500. For a rough approximation of the average value of claims, even the least accurate methods may suffice. This contrasts with estimating the overall number of med debt cases, where methods such as the Manual Top 100 fail to capture close to the current number of cases.

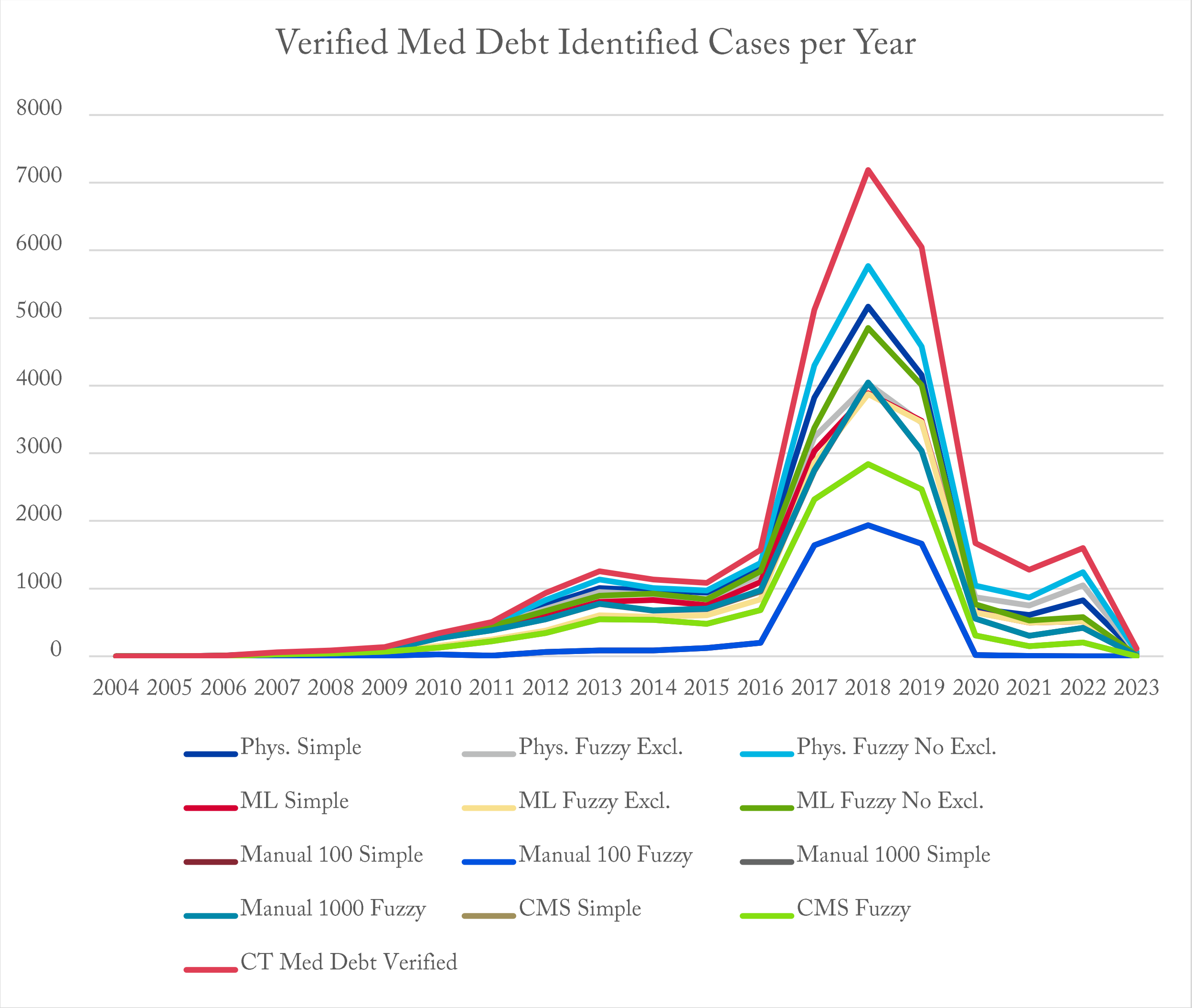

Figure 2: Verified Med Debt Identified Cases per Year

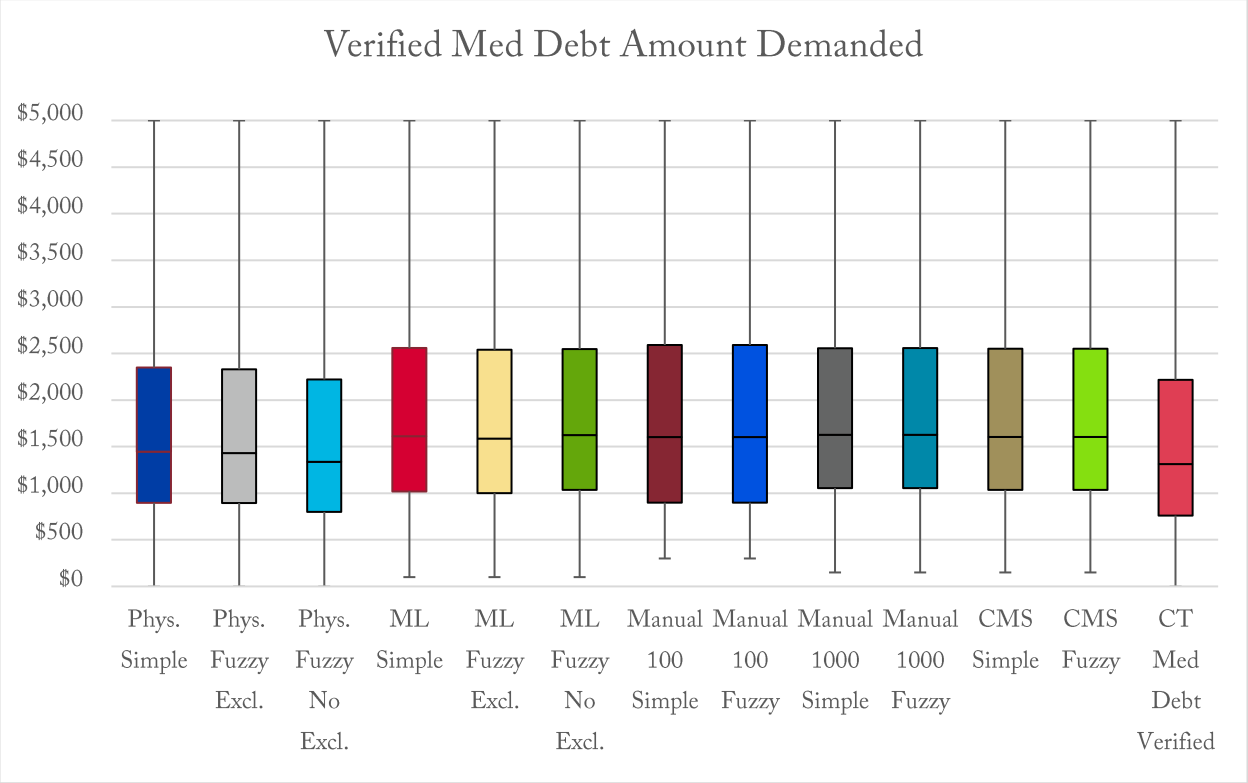

Figure 3: Verified Med Debt Amount Demanded

In addition to running the methods on the subset of cases which were verified med debt based on case type, we also ran them on a larger set of civil cases that could possibly be medical debt. The first narrowing step was to limit cases to those coded with a case type which could possibly contain medical debt collection, including the broader Collections case type and cases which had been transferred from Small Claims to the broader civil court.

Figure 4 shows the total number of cases identified from this dataset that includes all the possible locations of med debt of any dollar amount. The general ratios of cases identified between the different methods are fairly consistent with those when the methods were run against only verified med debt records. Likewise, Figure 5 indicates that the year-over-year matches closely to that of the verified med debt cases. More years are included as the Civil Justice Data Commons has case records for years further in the past for non-small-claims cases.

Table 2, at the end of this report, likewise shows that the overlap percentages between identified records are similar. Once again, the physician consultation method combined with fuzzy matching seems to produce the greatest number of identified records.

However, this may be due to incorrectly identified records. To account for this, we conducted a manual review of cases records identified to estimate the accuracy of each method. Figure 6 shows the results of this manual review. For each method and algorithm combination, we performed 10 samples of 100 records each and conducted manual review on the identified cases to verify if they were med debt or not. Overall, every method is largely accurate. Although the physician consultation method with fuzzy matching is less accurate than some of the other methods, it still averages over 97% accuracy in its case identification. Surprisingly, this review also indicates that the use of excluded sets of strings may not result in a meaningful accuracy increase.

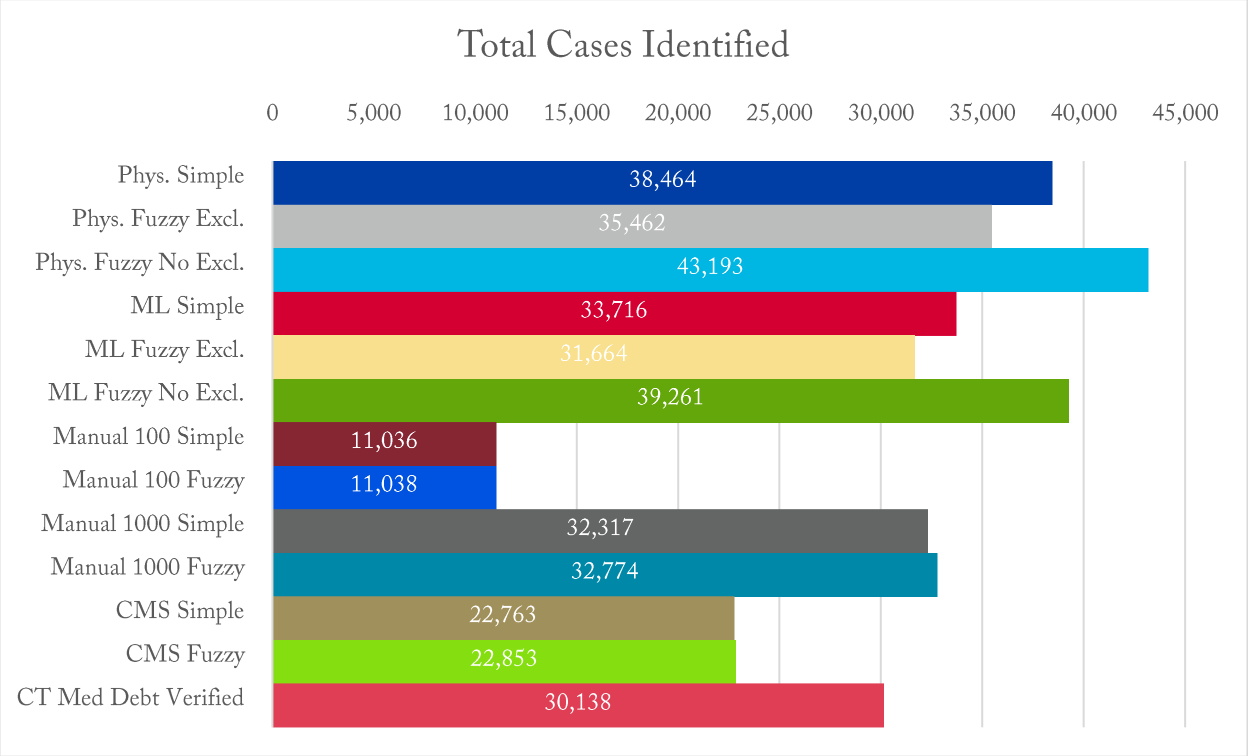

Figure 4: Total Cases Identified

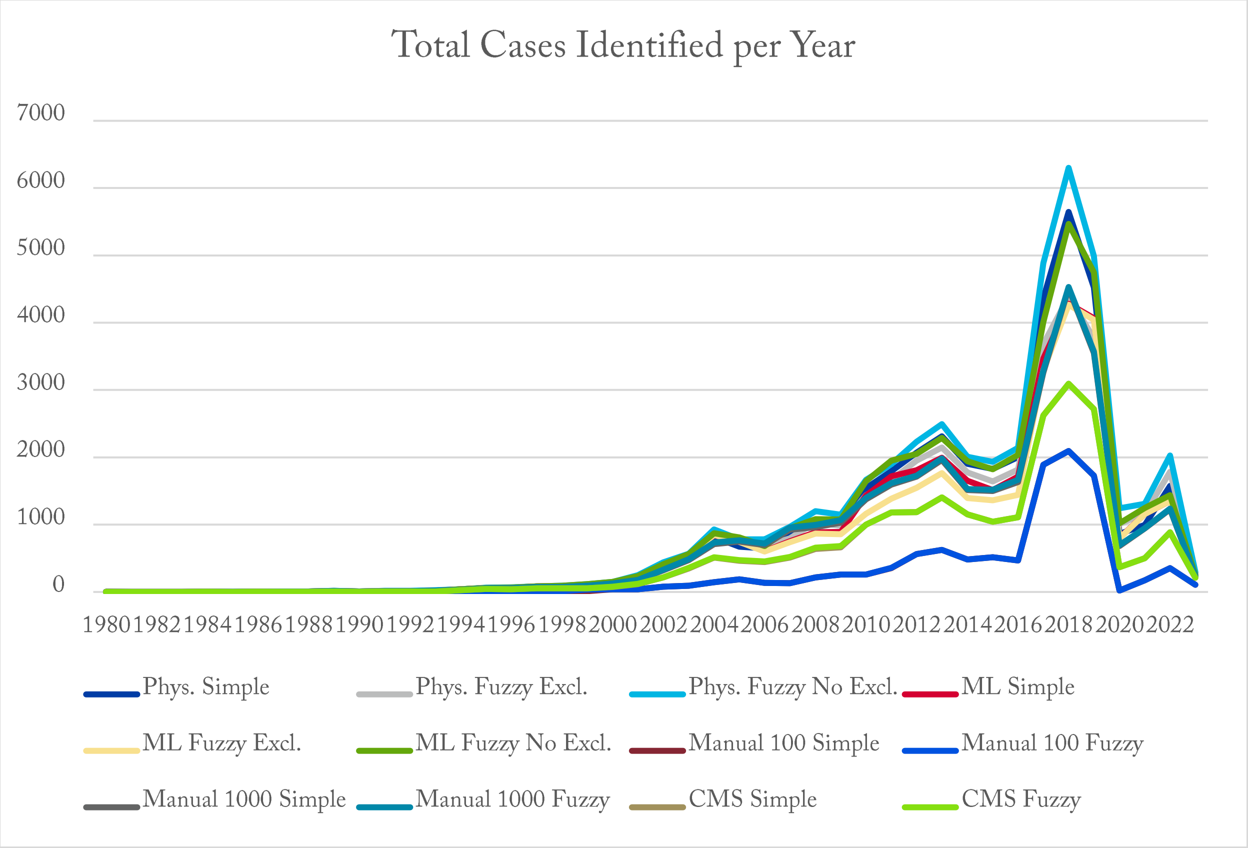

Figure 5: Total Cases Identified per Year

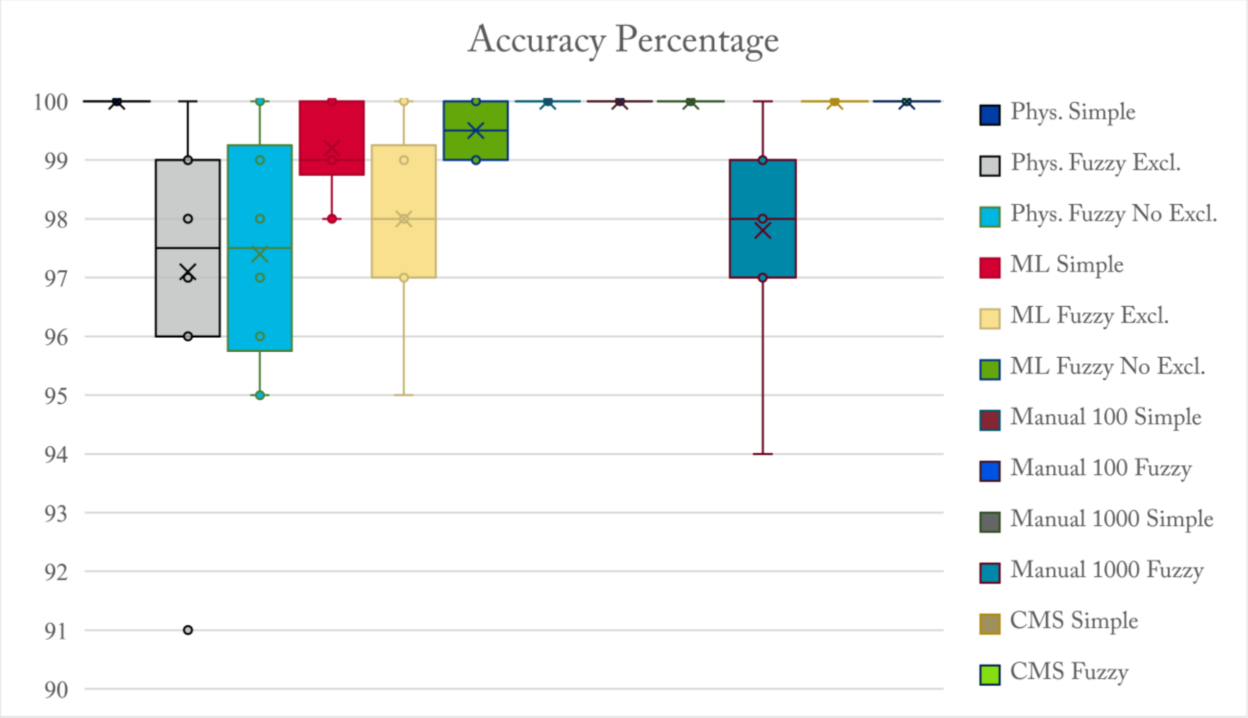

Figure 6: Accuracy Percentage

Takeaways and Further Research

Our case study indicates that our physician consultation method shows promise for identifying medical debt collection cases with a low overhead for new jurisdictions, as it functioned well on Connecticut court data after being developed using Oklahoma data. The method of employing machine learning-derived sets of strings also shows potential, though in this case it underperformed the physician consultation method. Additionally, the matching algorithm used does make a non-trivial difference, with fuzzy matching consistently performing better than simple string matching.

However, none of the methods were perfect, and even in the best case only found 80% of verified med debt cases. This shows both how messy and difficult to parse court data can be, as well as why researcher med debt identification is currently not a substitute for improvements in court data, particularly more complete case categorization around medical debt. Without greater advances, we can only see a fraction of the picture of med debt in America's courts, and study of the medical debt collections crisis cannot be complete. Novel methods of case identification by researchers, though, may help fill in the blanks, so long as those researchers have access to court data through programs such as the Civil Justice Data Commons and through courts opening their data to outside research.

Table 1: Verified Med Debt Cases Identified Overlaps

	Overlap with:	Phys. Simple		Phys. Fuzzy Excl.		Phys. Fuzzy No Excl.		ML Simple		ML Fuzzy Excl.		ML Fuzzy No Excl.		Manual 100 Simple		Manual 100 Fuzzy		Manual 1000 Simple		Manual 1000 Fuzzy		CMS Simple		CMS Fuzzy		CT Med Debt Verified	
Method	Total Cases	21,313		19,013		24,215		17,038		15,443		19,686		5,875		5,875		15,576		15,633		11,368		11,368		30,138	
Phys. Simple	21,313	21,313	100%	17,164	90%	21,313	88%	14,308	84%	12,761	83%	16,713	85%	5,875	100%	5,875	100%	14,594	94%	14,644	94%	11,368	100%	11,368	100%	21,313	71%
Phys. Fuzzy Excl.	19,013	17,164	81%	19,013	100%	19,013	79%	13,908	82%	12,373	80%	14,972	76%	5,316	90%	5,316	90%	12,862	83%	12,900	83%	11,091	98%	11,091	98%	19,013	63%
Phys. Fuzzy No Excl.	24,215	21,313	100%	19,013	100%	24,215	100%	14,327	84%	12,774	83%	16,760	85%	5,875	100%	5,875	100%	14,783	95%	14,833	95%	11,368	100%	11,368	100%	24,215	80%
ML Simple	17,038	14,308	67%	13,908	73%	14,327	59%	17,038	100%	15,443	100%	17,038	87%	5,316	90%	5,316	90%	12,430	80%	12,480	80%	10,254	90%	10,254	90%	17,038	57%
ML Fuzzy Excl.	15,443	12,761	60%	12,373	65%	12,774	53%	15,443	91%	15,443	100%	15,443	78%	5,316	90%	5,316	90%	12,035	77%	12,085	77%	10,214	90%	10,214	90%	15,443	51%
ML Fuzzy No Excl.	19,686	16,713	78%	14,972	79%	16,760	69%	17,038	100%	15,443	100%	19,686	100%	5,875	100%	5,875	100%	14,534	93%	14,585	93%	11,368	100%	11,368	100%	19,686	65%
Manual 100 Simple	5,875	5,875	28%	5,316	28%	5,875	24%	5,316	31%	5,316	34%	5,875	30%	5,875	100%	5,875	100%	5,875	38%	5,875	38%	4,455	39%	4,455	39%	5,875	19%
Manual 100 Fuzzy	5,875	5,875	28%	5,316	28%	5,875	24%	5,316	31%	5,316	34%	5,875	30%	5,875	100%	5,875	100%	5,875	38%	5,875	38%	4,455	39%	4,455	39%	5,875	19%
Manual 1000 Simple	15,576	14,594	68%	12,862	68%	14,783	61%	12,430	73%	12,035	78%	14,534	74%	5,875	100%	5,875	100%	15,576	100%	15,576	100%	11,276	99%	11,276	99%	15,576	52%
Manual 1000 Fuzzy	15,633	14,644	69%	12,900	68%	14,833	61%	12,480	73%	12,085	78%	14,585	74%	5,875	100%	5,875	100%	15,576	100%	15,633	100%	11,281	99%	11,281	99%	15,633	52%
CMS Simple	11,368	11,368	53%	11,091	58%	11,368	47%	10,254	60%	10,214	66%	11,368	58%	4,455	76%	4,455	76%	11,276	72%	11,281	72%	11,368	100%	11,368	100%	11,368	38%
CMS Fuzzy	11,368	11,368	53%	11,091	58%	11,368	47%	10,254	60%	10,214	66%	11,368	58%	4,455	76%	4,455	76%	11,276	72%	11,281	72%	11,368	100%	11,368	100%	11,368	38%
CT Med Debt Verified	30,138	21,313	100%	19,013	100%	24,215	100%	17,038	100%	15,443	100%	19,686	100%	5,875	100%	5,875	100%	15,576	100%	15,633	100%	11,368	100%	11,368	100%	30,138	100%

Table 2: All Identified Cases Overlaps

	Overlap with:	Phys. Simple		Phys. Fuzzy Excl.		Phys. Fuzzy No Excl.		ML Simple		ML Fuzzy Excl.		ML Fuzzy No Excl.		Manual 100 Simple		Manual 100 Fuzzy		Manual 1000 Simple		Manual 1000 Fuzzy		CMS Simple		CMS Fuzzy		CT Med Debt Verified	
Method	Total Cases	38,464		35,462		43,193		33,716		31,664		39,261		11,036		11,038		32,317		32,774		22,763		22,853		30,138	
Phys. Simple	38,464	38,464	100%	32,123	91%	38,464	89%	29,098	86%	27,018	85%	33,134	84%	11,036	100%	11,036	100%	29,100	90%	29,153	89%	22,755	100%	22,755	100%	21,313	71%
Phys. Fuzzy Excl.	35,462	32,123	84%	35,462	100%	35,462	82%	28,328	84%	26,349	83%	29,671	76%	9,779	89%	9,781	89%	25,985	80%	26,046	79%	22,113	97%	22,203	97%	19,013	63%
Phys. Fuzzy No Excl.	43,193	38,464	100%	35,462	100%	43,193	100%	29,147	86%	27,136	86%	33,364	85%	11,036	100%	11,038	100%	29,347	91%	29,422	90%	22,763	100%	22,853	100%	24,215	80%
ML Simple	33,716	29,098	76%	28,328	80%	29,147	67%	33,716	100%	31,525	100%	33,716	86%	9,779	89%	9,779	89%	26,439	82%	26,492	81%	21,453	94%	21,454	94%	17,038	57%
ML Fuzzy Excl.	31,664	27,018	70%	26,349	74%	27,136	63%	31,525	94%	31,664	100%	31,664	81%	9,779	89%	9,781	89%	25,653	79%	25,725	78%	21,135	93%	21,199	93%	15,443	51%
ML Fuzzy No Excl.	39,261	33,134	86%	29,671	84%	33,364	77%	33,716	100%	31,664	100%	39,261	100%	11,036	100%	11,038	100%	29,759	92%	29,832	91%	22,763	100%	22,853	100%	19,686	65%
Manual 100 Simple	11,036	11,036	29%	9,779	28%	11,036	26%	9,779	29%	9,779	31%	11,036	28%	11,036	100%	11,036	100%	11,036	34%	11,036	34%	7,909	35%	7,909	35%	5,875	19%
Manual 100 Fuzzy	11,038	11,036	29%	9,781	28%	11,038	26%	9,779	29%	9,781	31%	11,038	28%	11,036	100%	11,038	100%	11,036	34%	11,038	34%	7,909	35%	7,911	35%	5,875	19%
Manual 1000 Simple	32,317	29,100	76%	25,985	73%	29,347	68%	26,439	78%	25,653	81%	29,759	76%	11,036	100%	11,036	100%	32,317	100%	32,317	99%	22,262	98%	22,306	98%	15,576	52%
Manual 1000 Fuzzy	32,774	29,153	76%	26,046	73%	29,422	68%	26,492	79%	25,725	81%	29,832	76%	11,036	100%	11,038	100%	32,317	100%	32,774	100%	22,267	98%	22,331	98%	15,633	52%
CMS Simple	22,763	22,755	59%	22,113	62%	22,763	53%	21,453	64%	21,135	67%	22,763	58%	7,909	72%	7,909	72%	22,262	69%	22,267	68%	22,763	100%	22,763	100%	11,368	38%
CMS Fuzzy	22,853	22,755	59%	22,203	63%	22,853	53%	21,454	64%	21,199	67%	22,853	58%	7,909	72%	7,911	72%	22,306	69%	22,331	68%	22,763	100%	22,853	100%	11,368	38%
CT Med Debt Verified	30,138	21,313	55%	19,013	54%	24,215	56%	17,038	51%	15,443	49%	19,686	50%	5,875	53%	5,875	53%	15,576	48%	15,633	48%	11,368	50%	11,368	50%	30,138	100%

End Notes

[i] Shameek Rakshit, Matthew Rae, Gary Claxton, Krutika Amin, and Cynthia Cox, *The Burden of Medical Debt in the United States*, Peterson-KFF Health System Tracker (Feb. 12, 2024), https://www.healthsystemtracker.org/brief/the-burden-of-medical-debt-in-the-united-states/.

[ii] Quentin Fottrell, *The 'Tragedy' of American Healthcare: Olympic Gymnast Mary Lou Retton's Family is Crowdsourcing for Her Hospital Bills. She's Not Alone.*, MarketWatch (Oct. 11, 2023), https://www.marketwatch.com/story/the-tragedy-of-american-healthcare-olympic-gymnast-mary-lou-rettons-family-is-relying-on-gofundme-for-her-hospital-bills-shes-not-alone-9ee22cb2.

[iii] Anna Claire Vollers, *Governments Can Erase Your Medical Debt for Pennies on The Dollar — And Some Are*, Stateline (Feb. 13, 2024), https://stateline.org/2024/02/13/governments-can-erase-your-medical-debt-for-pennies-on-the-dollar-and-some-are/.

[iv] Associated Press, *New York City Plans to Wipe Out $2 billion in Medical Debt for 500,000 Residents*, Associated Press (Jan. 22, 2024), https://apnews.com/article/rip-medical-debt-new-york-city-adams-1f39530cd79937ced52f47ab4749fb58.

[v] Consumer Financial Protection Bureau, *Medical Debt Burden in the United States*, Consumer Financial Protection Bureau (2022), (available at https://files.consumerfinance.gov/f/documents/cfpb_medical-debt-burden-in-the-united-states_report_2022-03.pdf); Raymond Kluender, Neale Mahoney, Francis Wong, et al, *Medical Debt in the US, 2009-2020*, 326(3) JAMA 250-256 (2021).

[vi] Dan Grunebaum, *Americans Knowingly Going into Medical Debt: Survey*, Healthcare.com (Mar. 15, 2022), https://www.healthcare.com/americans-knowingly-going-into-medical-debt-survey-481593; Sara R. Collins, Gabriella N. Aboulafia, Munira Z. Gunja, *As the Pandemic Eases, What is the State of Health Care Coverage and Affordability in the U.S.?*, Commonwealth Fund (Jul. 16, 2021), https://www.commonwealthfund.org/publications/issue-briefs/2021/jul/as-pandemic-eases-what-is-state-coverage-affordability-survey.

[vii] Farah Hashim et al, *Eroding the Public Trust: A Report of Texas Hospitals Suing Patients*, Arnold Ventures (2020), available at https://a2e0dcdc-3168-4345-9e39-788b0a5bb779.filesusr.com/ugd/29ca8c_095296028da54e778dbfb34987c3cc9c.pdf; S. Kliff, *With Medical Bills Skyrocketing, More Hospitals are Suing for Payment*, New York Times (Nov. 8, 2019), https://www.nytimes.com/2019/11/08/us/hospitals-lawsuits-medical-debt.html.

[viii] Sara R. Collins, Shereya Roy, Relebohile Masitha, *Paying for It: How Health Care Costs and Medical Debt are Making Americans Sicker and Poorer*, Commonwealth Fund (Oct. 26, 2023), https://www.commonwealthfund.org/publications/surveys/2023/oct/paying-for-it-costs-debt-americans-sicker-poorer-2023-affordability-survey.

[ix] Liz Hamel, Mira Norton, Karen Pollitz, Larry Levitt, Gary Claxton, & Mollyann Brodie, *The Burden of Medical Debt: Results from the Kaiser Family Foundation/New York Times Medical Bills Survey*, The Henry J. Kaiser Family Foundation (Jan. 2016), https://www.kff.org/wp-content/uploads/2016/01/8806-the-burden-of-medical-debt-results-from-the-kaiser-family-foundation-new-york-times-medical-bills-survey.pdf.

[x] Collins (*end note viii.*).

[xi] Bram Sable-Smith, *'You've Been Served': Wisconsin Hospitals Sue Patients Over Debt—Even During Pandemic*, Wisconsin Watch (Apr. 1, 2020), https://wisconsinwatch.org/2020/04/hospitals-sue-patients-during-pandemic/; Collins (*end note viii.*); Kliff (*end note vii.*).

[xii] Carlos Dobkin, Amy Finkelstein, Raymond Kluender, & Mathew J. Notowidigdo, *Myth and Measurement: The Case of Medical Bankruptcies*, 378 N. Eng. J. Med. 1076-1078 (Mar. 21, 2018), available at https://www.nejm.org/doi/10.1056/NEJMp1716604.

[xiii] Kliff (*end note vii.*); Sable-Smith (*end note xi.*).

[xiv] Annie Waldman & Paul Kiel, *Racial Disparity in Debt Collection Lawsuits: A Study of Three Metro Areas*, ProPublica (Oct. 8, 2015), available at https://static.propublica.org/projects/race-and-debt/assets/pdf/ProPublica-garnishments-whitepaper.pdf; Michael Batty, Christa Gibbs, & Benedic Ippolito, *Unlike Medical Spending, Medical Bills in Collections Decrease with Patient's Age*, 37 Health Affairs (Jul. 25, 2018), available at https://www.healthaffairs.org/doi/10.1377/hlthaff.2018.0349; National Nurses United, AFL-CIO, & Coalition for Humane Hopkins, *Taking Neighbors to Court: John Hopkings Hospital Medical Debt Lawsuits*¸ National Nurses United (May 2019), available at https://www.nationalnursesunited.org/sites/default/files/nnu/documents/Johns-Hopkins-Medical-Debt-report.pdf; Neil Bennett, Jonathan

Eggleston, Laryssa Mykyta, & Briana Sullivan, *Who Had Medical Debt in the United States?: 19% of U.S. Households Could Not Afford to Pay for Medical Care Right Away*, U.S. Census Bureau (Apr. 7, 2021), https://www.census.gov/library/stories/2021/04/who-had-medical-debt-in-united-states.html; Zach Cooper, James Han, & Neale Mahoney, *Hospital Lawsuits Over Unpaid Bills Increased by 37 Percent in Wisconsin from 2001 to 2018*, 40 Health Affairs (Dec. 2021), available at https://www.healthaffairs.org/doi/10.1377/hlthaff.2021.01130; CFPB (*end note v.*); Jay Hancock & Elizabeth Lucas, *'UVA Has Ruined Us': Health System Sues Thousands of Patients, Seizing Paychecks and Claiming Homes*, KFF Health News (Sept. 10, 2019), https://kffhealthnews.org/news/uva-health-system-sues-patients-virginia-courts-garnishment-liens-bankruptcy/; Alec MacGillis, *One Thing the Pandemic Hasn't Stopped: Aggressive Medical-Debt Collection*, ProPublica (Apr. 28, 2020), https://www.propublica.org/article/one-thing-the-pandemic-hasnt-stopped-aggressive-medical-debt-collection; Abell Foundation, *Limits on Medical Debt Lawsuits*, Abell Foundation (Apr. 27, 2021), https://abell.org/publication/limits-on-medical-debt-lawsuits/.

[xv] Waldman (*end note xiv.*).

[xvi] Kliff (*end note vii.*).

[xvii] Kluender (*end note v.*).

[xviii] David U. Himmelstein, Robert M. Lawless, Deborah Thorne, Pamela Foohey, & Steffie Woolhandler, *Medical Bankruptcy: Still Common Despite Affordable Care Act*, 109(3) Am. J. Public Health 431-433 (Mar. 2019), available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6366487/.

[xix] Kate Dore, *More than 1 in 3 U.S. Adults Carry Medical Debt, Survey Finds*, CNBC (Nov. 24, 2021), https://www.cnbc.com/2021/11/24/more-than-1-in-3-us-adults-carry-medical-debt-survey-finds.html.

[xx] Elizabeth Cohen & John Bonifield, *When Some Patients Don't Pay, This Hospital Sues*, CNN (Sept. 10, 2019), https://www.cnn.com/2019/09/10/health/carlsbad-new-mexico-hospital-eprise/index.html.

[xxi] Tanina Rostain & Amy O'Hara, *The Civil Justice Data Gap*, in Legal Tech and the Future of Civil Justice (ed. David Freeman Engstrom, Feb. 2, 2023), available at https://www.cambridge.org/core/books/legal-tech-and-the-future-of-civil-justice/civil-justice-data-gap/D9C1D9EAD0E2E8ACF1DF7FDEB2BB0875.

[xxii] Hancock (*end note xiv.*).

[xxiii] Maanasa Kona & Vrudhi Raimugia, *State Protections Against Medical Debt: A Look at Policies Across the U.S.*, Commonwealth Fund (Sept. 7, 2023), https://www.commonwealthfund.org/publications/fund-reports/2023/sep/state-protections-medical-debt-policies-across-us.

[xxiv] Kona (*end note xxi.*).

[xxv] Innovation for Justice, *Medical Debt Policy Scorecard*, U. of AZ and U. of UT (2022), https://medicaldebtpolicyscorecard.org/.

[xxvi] Cooper (*end note xiv.*).

[xxvii] Ericson (*end note xiv.*).

[xxviii] Minnesota State Bar Association Access to Justice Committee, *Minnesota Consumer Debt Litigation: A Statewide Access to Justice Report*, Minnesota State Bar Association (2023), https://www.mnbar.org/about-msba/what-we-stand-for/access-to-justice/debt-litigation-report.

[xxix] Community Foundation of Greater Chattanooga, *The Impact of Debt Collection Lawsuits in Hamilton County, TN*, Community Foundation of Greater Chattanooga (Mar. 12, 2024), https://www.cfgc.org/index-entry/the-impact-of-debt-collection-lawsuits-in-hamilton-county-tn.

[xxx] Minnesota State Bar Access to Justice Committee (*end note xxviii.*).

[xxxi] *Arkansas Judiciary Case Search*, Arkansas Courts, https://caseinfo.arcourts.gov/opad (last accessed Apr. 10, 2024) (Case type "CB" is "Debt Medical Expenses").

[xxxii] *Party Name Search*, State of Connecticut Judicial Branch, https://civilinquiry.jud.ct.gov/PartySearch.aspx (last accessed Apr. 10, 2024); State of Connecticut Judicial Branch, *Understanding the Display of Case Information*, State of Connecticut Judicial Branch (Aug. 14, 2017), available at https://civilinquiry.jud.ct.gov/Understanding%20Display%20of%20Case%20Information.pdf; State of Connecticut Judicial Branch, *E-File Small Claims Matters*, State of Connecticut Judicial Branch (Jun. 1, 2018).

[xxxiii] Civil Justice Data Commons, *Georgetown Civil Justice Data Commons*, Redivis, https://redivis.com/cjdc (last accessed Apr. 10, 2024).

[xxxiv] Julia Hohenadel, *Furthering the Analysis of Medical Debt in Civil Court Data with Natural Language Processing*, U. of Guelph (2022).

[xxxv] Michigan Justice for All Commission, *Advancing Justice for All in Debt Collection Lawsuits*, Michigan Courts (Nov. 2022), available at https://www.courts.michigan.gov/4ac33d/siteassets/reports/special-initiatives/justice-for-all/jfa_advancing_justice_for_all_in_debt_collection_lawsuits.pdf.

[xxxvi] Centers for Medicare & Medicaid Services, *Hospital General Information*, DATA.CMS.GOV, https://data.cms.gov/provider-data/dataset/xubh-q36u (last accessed Apr. 10, 2024).

[xxxvii] State of Connecticut Judicial Branch, *E-File Small Claims Matters* (*end note xxxii.*).