
Open Hospital Prices: Reducing Healthcare Costs via Price Transparency

1 Project Description

This Open Hospital Prices initiative aims to reduce hospital bills in the United States through price transparency. While hospitals are mandated to release their prices, this data comes in a variety of machine-readable formats that are not accessible to an average person. This Magic Grant will enable anyone in the U.S. to assess how price varies by hospital, region, and service, and to create accessible interfaces that empower people to make better care decisions.

Our project has three parts:

- Build an aggregate data archive of hospital prices and the tools for automatically adjudicating differences in terminology
- Write news articles and create resources for data journalists to connect the data with anecdotal evidence
- Make an actionable toolkit for consumers looking to reduce their health care bills: a search engine on top of the data archive, and calling scripts populated by an automatic negotiation platform

Making a difference in a problem of this scale requires a mix of technical and human approaches: our work begins with data curation and machine learning, and ends with practical tools in the hands of well-informed consumers.

1.1 Background

Healthcare Spending Americans spend \$10,966 per person on healthcare annually, or twice the \$5,697 of comparable OECD countries, a gap that has increased since 1980. Despite this spending difference, Americans have not seen improved life expectancy over their peers [4]. More than 50% of Americans carry medical debt, with total outstanding debt estimated at \$200BN-\$1TN [3]. Americans avoid medical help when they can't afford it, skipping checkups and choosing to live with symptoms for fear of surprise charges. Those who are forced to seek care can end up with crippling debt. Worse, research is limited in its ability to assess the effects of pricing structures on medical professionals, structures that may lead to over-prescription, upcoding, over-diagnosis, or unnecessary visits. All of this is related to the fact that hospital administrators and insurers understand the recommended charges of the billing desk much better than consumers do.

Congressional Bill and Lack of Transparency In 2020, the U.S. Congress passed S.1524, the Health Care PRICE Transparency Act, which requires that all hospitals post their prices online. Per the sponsor: "these solutions will bring transparency to the healthcare system and in turn create more competition ... knowing the cost of health care will empower Americans to shop for the services best suited to their needs" [1]. However, this bill has had limited success. Independent studies note compliance rates as low as 5.6% [8]. Furthermore, in preliminary analyses for this Magic Grant, we have identified that the following difficulties remain unresolved by the bill:

- *Billing Items are Indecipherable to Existing Tools.* Under the sub-field of artificial intelligence known as natural language processing (NLP), we recognize that matching dense bill descriptions like "HC NMR LIPO PRFL" to standardized entries to be an ontology mapping problem, an extension of named entity recognition. The best tools for doing this for medical terms are CTakes and MetaMap, but we find neither is able to match more than 5-10% of rows between two hospital price lists.
- *Disorganized or Incomplete Data.* The bill contains no stipulations for the 13,944 hospitals [7] to post data in a common format. In many cases, we find hospitals intentionally mask or hide their data, with malformed JSON, extensive trainings required to access the data, or differences in insurance billing notations.

Strong Price Variations are not Publicly Known

Our initial research finds many differences in prices by providers, with some as high as 2-4x the cost of neighboring hospitals as shown in Table 1. Such alarming differences exist because consumers have no visibility into what they'll be charged; if prices were more obvious, consumers would be empowered to shop around.

1.2 Deliverables

Outcome: A reduction in the median hospital bill price paid by consumers through the following resources:

Tools and Datasets for Hospital Price Analysis

- **Tools to Compare Raw Pricing Data** Tools which can parse dense service descriptions like for clinically recognized terms and compare between hospitals.
- **A Database of Hospital Prices** A database with the maximum number of hospitals, such that any two hospitals are comparable by the service.

Reducing the Cost of Health Care Through Increased Price Transparency

Hospital A	Hospital B	Medication Price (A/B)	Procedure Price (A/B)	Total Visits/Year
NYU Langone	NewYork-Presbyterian	1.6	1.1	8.0M
NYU Langone	Lenox Hill	1.8	1.5	8.0M
NYU Langone	Staten Island University Hospital	1.9	1.8	6.3M
NYU Langone	Jamaica Hospital MC	2.1	3.7	6.4M
NewYork-Presbyterian	Lenox Hill	2.3	1.8	2.3M
NewYork-Presbyterian	Staten Island University Hospital	3.0	2.1	2.4M
NewYork-Presbyterian	Jamaica Hospital MC	4.5	2.3	2.35M
Lenox Hill	Staten Island University Hospital	1.6	1.4	0.7M
Lenox Hill	Jamaica Hospital MC	2.0	1.2	0.65M

Table 1: **Pairwise price comparisons of major area hospitals reveal significant discoveries in the Open Hospital Prices proof of concept for the Greater New York City Area.** A/B = 2 implies hospital A costs twice as much as hospital B. Medications and Procedures are baskets of terms automatically identified by the software pipeline built for this Magic Grant proof of concept.

Empowering Data Journalists with Interactive Price Archives

- **A Data Journalism Resource** Starter code, data schema definitions, and data exploration tools will help local news organizations write articles on health-care prices in their community.
- **Consumer Interest Articles** The proof of concept work led us to heuristics like medication costing 2x at one hospital versus another, or the advice to not to stay overnight in a hospital where you can't afford a neighboring hotel. Guidance like this will influence consumer behavior to avoid price gouging.

An example article might be one where OBGYNs anonymously comment on the pressure to add procedures during labor, which links to our tools showing the prices of those procedures at local hospitals.

- **National Policy Articles** Trends in aggregate pricing data surfaced by the interactive data archives can guide healthcare regulators to improve price transparency and reduce costs.

Consumer-Oriented Website to Achieve Better Price Outcomes

- **A Bill Negotiation Platform** A website which provides data-driven insights to help consumers negotiate their bills, including calling scripts.
- **A Price Shopping Platform** A website which allows consumers to shop for local hospital services and make an informed decision on which hospital to visit for a given procedure.

1.3 Proof of Concept and Prior NLP Work

Table 1 shows selected results from our prior [study](#) of price differences in NYC hospitals. NYU Langone and NewYork-Presbyterian hospitals both cost 2-3x their Queens and Staten Island counterparts when measured by standard medications or procedures, with the Lenox Hill hospital falling in the middle.

In writing this Magic Grant, we were concerned to learn how the existing clinical text parsing tools do not use modern NLP techniques like BERT [2]. Such models dramatically outperform prior techniques on natural language understanding tasks like extracting data or answering questions [10], indicating that a much higher degree of accuracy is possible than has been attempted in this domain.

Our team has the unique experience to accomplish an NLP task of this scale. We have overseen the data labeling and model training for thousands of documents and built software applications that help lawyers negotiate. We've authored highly cited papers in Clinical NLP [5] and written code that analyzed over 100M patient records through network studies at Columbia.

2 Timeline

Q4, 2022 *Data:* scale web scraping pipeline to all 10k+ hospitals in the U.S.

Annotation: hire experts (physicians, medical trainees, billing staff) using the pipelines we've built for data annotation to rectify disparate information across institutions.

Journalism: conduct stakeholder interviews with hospital employees and members of the public for UX research and longer features (e.g. in the areas of maternal health and medical debt). Screen newsrooms like ProPublica for partnership in storytelling and data visualization.

Q1, 2023 *Machine learning:* train NLP models on annotated datasets and assess the accuracy of the named entity recognition system using the measure in Sec 4.

Web development: UI/UX research for initial release of Open Hospital Prices archive.

Journalism: finalize partnerships with newsrooms for longer feature stories.

Q2, 2023 *Partnerships:* outreach to hospitals in the dataset to correct any errors that could not be fixed with human annotators.

Web development: deploy optical character recognition tool for analyzing user-uploaded medical bills; UI/UX testing for ensuring users can readily explore the database in a manner tailored to their medical bills.

Entrepreneurship: initial sales calls to sell pricing data. This represents a unique business opportunity for the hybrid nonprofit and for-profit structure described in Sec 5.

Q3, 2023 *Journalism:* Publish initial reporting, data visualization platform, interactive data archive, and open source tooling.

Entrepreneurship: onboard initial industry clients onto interactive data archive. Prototype free-to-use medical bill dispute system similar to [WinIt](#) or [GoodRx](#) to help consumers navigate prices.

3 Innovation and Originality

In preparation for this grant, we made a sales call to UnitedHealthCare insurance to learn that [FAIR Health](#) is the vendor major insurers use for price data. However, in a subsequent introductory sales call with FAIR on March 30, 2022, we learned that their geo-level data only includes the first 3 digits of ZIP codes and they are thus unable to release facility-level info like our data archives enable.

Smaller-scale projects of this nature have been successful. [HospitalPrices.wtf](#) uses a 2017 dataset that underwent a similar curation localized to California. [GoodRx](#) is a tool which compares pharmacy prices. [Never Pay the First Bill](#) [6] outlines ways to use billing codes and insurance data to negotiate charges.

Yet the full data targeted in this project has only been released in the last 1-2 years, and there are not yet the models and tools which can effectively combine it. No matter who builds a website for this dataset, the accuracy and usefulness of their insights will be limited by the fact that comparing pricing terms between hospital systems is a largely unsolved NLP problem to date. We've verified this through 50+ informational interviews with major and minor payors, providers, and purveyors of interoperability software.

4 Impact

We will evaluate the success and impact of our project by first measuring what percentage of all hospitals in the United States we have collected price transparency data from, and then by assessing the accuracy of the machine learning system required to harmonize data across institutions. We will assess the impact of our work by the number of people that have uploaded hospital bills into our tool that enables easy price comparison.

Measure: Reduction in Median Hospital Bill The overall success of this project will be tracked by if hospital bill prices reduce relative to inflation.

Measure: Accuracy and Quality of Data Analysis Tools For our database, we will evaluate using completeness. For the ontology mapping tools, we will compute standard measures of accuracy like inter-annotator agreement in labeling the dataset, and calculate the F1 score for correctly identifying terms.

Measure: Usage of Consumer-Oriented Tools and Articles For our internet-facing tools, we will measure effectiveness in terms of page visits, social media shares, and the readership of our articles and articles written using data we produced.

5 Entrepreneurship

We are working with Carly Leinheiser, J.D., whose [firm](#) specializes in the social impact sector, and [Aaron Williamson](#), J.D., who has worked with the Software Freedom Law Center. Leinheiser and Williamson helped several open source software projects become sustainable

through structures that enable both consulting revenue and grants revenue. The legal structure that our counsel recommended for this work was a 501(c)3 tax-exempt organization with a Limited Liability Corporation as a wholly-owned subsidiary. This is known as a contract hybrid and the firm created [best practices](#) for this structure that enable social impact organizations to also generate business revenue through the LLC.

One such entity could be [One Fact, Inc.](#) (as one fact, such as the unequal price of medications between hospitals, is all it can take to create policy change). Before incorporating, we worked with filing FOIA requests to build and publish the first [national database](#) of all assisted living facilities in the country [9], and is now working with policymakers to expand access to housing for vulnerable populations. This Magic Grant would seed the first significant project for this entity and enable downstream impact on data journalism and policy in health care through the sustainable building of open source machine learning and journalism tools as digital public goods.

References

- [1] *Braun reintroduces healthcare transparency bills | Senator Mike Braun.*
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *arXiv:1810.04805* (2019).
- [3] D. Gordon. *50% Of Americans Now Carry Medical Debt, A New Chronic Condition For Millions.*
- [4] *How does health spending in the U.S. compare to other countries?*
- [5] K. Huang, J. Altosaar, and R. Ranganath. "Clinical-BERT: Modeling Clinical Notes and Predicting Hospital Readmission". *arXiv:1904.05342 [cs]* (2020).
- [6] Marshall. *Never Pay the First Bill: And Other Ways to Fight the Health Care System and Win.*
- [7] *Number of hospitals and hospital employment in each state in 2019 : The Economics Daily: U.S. Bureau of Labor Statistics.*
- [8] *Press Release Semi Annual Compliance Report 2022.*
- [9] A. Stengel, J. Altosaar, R. Dittrich, and N. Elhadad. "Assisted Living in the United States: an Open Dataset". *NeurIPS Machine Learning for Public Health*. 2021.
- [10] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". *arXiv:1804.07461* (2019).