The impact of payer status on hospital admissions: Evidence from an academic medical center

Abstract

Background: Inequity exists in accessing to care for patients with different payer statuses. However, there are few studies on the difference of hospital admissions. This study aims to examine how the payer status affects patients hospitalization from the perspective of a safety-net hospital.

Methods: We extracted all patients with visiting record in this medical center between 5/1/2009-4/30/2014, and then linked the outpatient and inpatient records three year before target admission time to patients. We conduct a retrospective observational study using a conditional logistic regression methodology. To control the illness of patients with different diseases in training the model, we construct a three-dimension variable with data stratification technology. The model is validated on a dataset distinct from the one used for training.

Results: Payer status is strongly associated with a patient’s admission. Patients covered by private insurance or uninsured are less likely to be admitted than those totally or partially insured by government. For uninsured patients, inequity in access to hospitalization is observed. Among all non-clinical influential factors considered in our study, payer status is a significant important factor.

Conclusion: Attention is needed on improving the access to care for vulnerable (low-income) patients, for example, by actively advertising free care programs, reaching out to community organizations with better access to these individuals, or offering assurances that access to care is not linked to immigration procedures. Also, in order to reduce preventable admissions, basic preventive care services should be enhanced.

Keywords: payer status; admissions; low-income; uninsured
Background

Based on their primary insurance payers, patients can be classified into five groups: Medicare, Medicaid, Commercial, Uninsured, and Other, among which Medicare and Medicaid are government administered programs whereas commercial insurance is provided by private insurers. Medicare is available for patients 65 or older, younger people with disabilities, and dialysis patients; Medicaid is available for low-income individuals or pays for costs as a supplement to Medicare. Patients who are insured by Medicaid are usually under-served comparing to patients with other payer status. Populations insured by government or uninsured are a significant focus of recent healthcare reform in U.S.[1]

In Massachusetts, health insurance coverage and gaps in access to care among different insurance programs are key points in the process of health reform since the passage of Massachusetts’ 2006 health care reform law[2]. According to the MA Health Reform survey, the coverage since 2008 has already reached 95% in MA, which is about 7% higher than the average coverage in U.S. However, according to this survey, there were 46.9% of insured adults in MA who have met difficulties in access to care in 2015. In California, there is an increasing trend of physician offices and hospitals refusing care for Medicaid patients, driven by the apparent low or delayed reimbursement for their services[3]. In this context, it becomes evident that strategies are needed to improve patients’ access to care. This work focuses on revealing the gaps in access to hospitalization from the perspective of payer status.

Recent work on this problem can be classified into two streams: (1) studying the inequities in access to care and treatment; and (2) demonstrating how policy changes influence access to care. Generally, equity in healthcare encompasses timely access, equivalence of care, and absence of avoidable or remediable differences among groups of people, pertinent to distinct social, economic, demographical or geographical criteria[4]. Newacheck et al.[5] investigated the effect of insurance on children’s
access to primary care. Recent efforts about primary payers were focused on the patients’ treatment, resource utilization and outcomes within a specific population[1]. Allen et al.[6] investigated the impact of payers on a patient’s in-hospital experience. They found that Medicaid is often associated with longer hospitalizations and high in-hospital mortality than other insurance types. Roetzheim et al.[7] and Adepoju et al.[8] investigated the effect of payer status on treatment and outcomes for patients with cancer. They found that the uninsured and Medicaid patients are less likely to receive breast-conserving treatment (BCT), a preferred method of surgical treatment in the early stage of breast cancer, than patients with private insurance. Giacovelli et al.[9] investigated the effect of insurance status on access to care and the outcomes of vascular disease. They found that Medicaid and uninsured patients present more severe disease and experience worse prognosis. Khouja et al. [10] investigated the effect of policy changes in Medicaid on low-income children’s access to preventable dental care.

In the second stream of research, Dzordzormenyoh [11] investigated the impact of Medicaid expansion to patients’ access to care from three aspects: access to a physician, access to basic healthcare, and access to specialized care. They concluded that the Medicaid expansion has improved the access for low-income patients. Mazurenko et al.[12] conducted a systematic review and concluded that the Medicaid expansion is associated with improvement in access to care and outcomes. Weissman et al.[13] found that the hospitalization rate for Medicaid and uninsured patients is higher than privately insured patients. Ferro et al.[14] found the Hospital Readmissions Reduction Program, which reduces payment to hospitals with excess readmissions, plays a role in avoiding readmissions.

Similar to the purpose of various programs expanding insurance coverage, safety-net providers are institutions aiming at providing care to all patients regardless of their ability to pay. Safety-net hospitals are usually associated with increased access
and service, lower rates of preventable hospitalizations, and better health outcomes for low-income patients. Sutton et al. [15] found that over 40% of inpatients in safety-net hospitals were either covered by Medicaid (34.7%) or uninsured (6.7%), compared to just 20.7% of inpatients in non-safety-net hospitals (16.8% Medicaid and 3.9% uninsured). Some researchers investigated driving factors of admission to a safety-net provider for uninsured patients; for example, Hadley and Cunningham [16] found that closer proximity to a safety-net provider increases access to care for uninsured patients, but the improvements in access to care are relatively small compared to the impact of similar factors on access to care for insured patients. Nevertheless, few studies have considered how payer status affects admissions to hospitals.

Our research focuses on investigating the impact of payer status on hospitalizations in an academic safety-net medical center and comparing the relative importance on influencing a patient’s hospitalization. It is helpful to find the inherent difference in hospital admission via studying from the level of a facility [17]. By analyzing the potential mechanism, we can obtain different policy implication for patients with different payer status. Unlike most studies focusing on a specific group of patients or a specific disease, we conduct this study from the perspective of a facility. We stratified patients into five groups according to their insurance sponsor: Government sponsored, Private-sponsored, Multiple-sponsored, Uninsured, and Other.

**Methods**

**Data**

The data we used come from a large academic medical center in the U.S. Only patients who had visit records during the period of 5/1/2009-4/30/2014 were included in our study, resulting in a dataset containing information for 490,761 patients. For each patient in this set, we extracted their demographic characteristics and medical
history (outpatient/emergency room visit records, diagnoses, medications, hospitalizations) during the period of 5/1/2006-4/30/2014. Table 1 describes the extracted features corresponding to each patient.

Dependent Variable

A binary variable Admission is defined indicating whether a patient has been hospitalized during the entire time-period we considered.

We find there exists significant imbalance between the sample size of the admitted and non-admitted patients. In order to reduce the imbalance, we introduce a “floating” time node – ”target time”, which is similar to the ”target year” used in Brisimi et al.[18] Initially, we set all patients’ target time to the upper limit of the period we consider. For patients who have ever been admitted during this period, we revise the target time as the time (year) of their last hospitalization. The model we will derive seeks to predict admission at the target time. With this method, the number of patients labeled as admitted accounts for 16.53% of the total number of patients.

Independent Variable

There are 8 payers in our sample: Medicare, Medicaid, Commonwealth care (offered by the state), Health safety net, Accident, Commercial, Self-pay and Other. According to the nature of payers, we classified them into 5 groups: government-private-sponsored, government-sponsored, private-sponsored, uninsured, and other (see Appendix A in Additional file 1 for detailed information on group composition). We use a categorical variable Payer to denote a patient’s payer status.

We note that some patients have multiple payers during the specific time-period. Since the insurance status of a patient may change over time, we consider the latest status as their payer status.
Covariates

In order to organize all the available information in a uniform way for all patients, some preprocessing of the data is needed. Pre-processing steps include imputing missing values and summarizing historical factors.

Sample imputation

Missing data imputation seeks to replace missing data with substituted values. We note that some records lack the information on the primary diagnosis. We use a matching algorithm to impute the missing values based on the following rules: (1) for patients who have prescriptions, we select the most common disease treated with their prescriptions as the substituted value; (2) for patients who have no prescriptions, we select the most common disease according to the history of their medical visits; and (3) for patients who have neither prescriptions nor prior medical visits, we label their diagnosis as "Unknown."

We note that a patient may seek care for different causes that correspond to different primary diagnoses in their visiting records. In order to keep each patient’s primary diagnosis unique in our sample, we choose the most common visit (or hospitalization) cause as the primary diagnosis for a patient who has never (ever) been hospitalized. We define Diagnosis as an unordered categorical variable denoting a patient’s primary diagnosis (23 categories; cf. Table 2).

Summarization of historical factors

We summarize records three years before a patient’s target time. We assigned a patient’s most common primary diagnosis in the variable we call Diagnosis. In order to include other diagnoses, we introduce a variable called Combinations, indicating the total number of other diseases for each patient. PrscNum denotes the total number of prescriptions in the record for this patient that are attributed to the primary diagnosis, which takes values in the interval 0 to 30, with 30 including cases of 30 or more prescriptions. Prior Admissions and ER Visits are both ordered
categorical variables coded 0 to 2, with 0 representing no earlier visit/admission records, 1 representing 1 earlier visit/admission records, and 2 representing more than 1 earlier visit/admission records. *Chronic* is a binary indicator variable indicating whether a patient’s disease is chronic.

We also include other demographic factors: *Age, Race, Marital Status*. *Age* is an ordinal category variable, coded 1 to 8, with 1 indicating ages no more than 10 years, 2 indicating ages from 10 to 20 years, etc., up to 7 corresponding to ages from 60 to 70 years, and 8 indicating ages more than 70 years. We convert categorical features—including *Race, Marital Status, Diagnosis* and *Payer*—into a set of binary variables with one-hot encoding. For all binary variables, zero indicates lack of this feature. In total, there are about 50 features for each patient. We present a summary description of the variables in Appendix D in Additional file 1.

**Patient and feature selection**

We remove features which are only available for a small number of patients. Patients who are under 3 years old are also removed as they have not enough historical records to indicate their physical condition until the selected target time. There are 462,809 patients retained after the steps outlined above.

**Statistical analysis**

In the following analysis, we randomly select 70% of the patients for training the logistic regression model and retain 30% for testing the performance of the trained model. The model predicts admission at the target year for each patient.

First, we present the descriptive characteristics of key factors. A Chi-square test is implemented to compare the frequency of various categories for categorical variables in the admitted and non-admitted cohorts. For continuous variables, a t-test is used to accept or refute the null hypothesis that the corresponding means of the variables are equal in the admitted and non-admitted cohorts.
Second, we utilize a statistical hypothesis test comparing the sample difference of proportions between the insured and uninsured patients.[18] With this method, we split our data set into two groups of size $N_1$ and $N_2$, respectively, where the first group includes insured patients and the second group the uninsured. Correspondingly, the admission rates are denoted by $p_1$ and $p_2$. Suppose that whether or not a patient has been insured does not influence their hospitalization. Then, $p_1$ should be statistically similar to $p_2$. Under the null hypothesis that $p_1 = p_2$, the difference between $p_1$ and $p_2$ approximately complies with a normal distribution, whose mean $\mu$ and deviation $\sigma$ equal to 0 and $PQ(1/N_1 + 1/N_2)$, respectively, where $P = (p_1N_1 + p_2N_2)/(N_1 + N_2)$ and $Q = 1 - p$. We can then use the estimator $z = (p_1 - p_2)/\sigma$ to assess whether the null hypothesis holds or not.

Then, we conduct a conditional logistic regression to further elaborate how a patient’s hospitalization was affected. Conditional logistic regression is a specialized type of logistic regression employed when the number of parameters is large relative to the number of samples[19]. It allows us to adjust for confounding at the design stage by stratifying samples according to one or more confounding factors. The distributions of cases and controls in different strata are usually substantially different. We stratify our sample based on the severity of illness. Lab tests are direct quantitative indicators reflecting the severity and complexity a patient’s health condition. However, we exclude these factors due to the serious deficiencies of lab test values and the difference in examination items, and use a three-dimensional tuple consisting of Age, ER Visits, and Prior Admissions as an indicator of severity. In our dataset, there are 72 different age-severity strata. The smallest and largest sample size in each stratum is 23 and 64,324, respectively.
The regression model is

\[
\log\left(\frac{\text{Prob(Admission)}}{1 - \text{Prob(Admission)}}\right) = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Race} + \beta_3 \text{Marital} \\
+ \beta_4 \text{PrscNum} + \beta_5 \text{Chronic} + \beta_6 \text{Diagnosis} \\
+ \beta_7 \text{Combinations} + \beta_8 \text{Payer} + \beta_9 \text{Severity}
\]

where Prob(Admission) denotes the probability of hospitalization in the target time, \( \beta = (\beta_0, \beta_1, ..., \beta_9) \) are unstandardized coefficients.

Finally, a Receiver Operating Characteristic (ROC) curve, which plots the sensitivity (or detection rate, or recall) as a function of the false positive rate (equal to one minus the specificity) is presented to demonstrate the performance of the model.

**Results**

After the data pre-processing, our study population consisted of 462,809 patients, 67,332 of whom were admitted during the target year. Table 2 presents summary statistics of our sample. Patients who have been admitted are more likely to be male, white, and be divorced or separated. The number of patients insured by a government program is much higher than those with other payer status in our sample, which is consistent with a safety-net health care provider. This is mainly the result of the mission of safety-net hospitals to provide healthcare to all populations regardless of their payer status or ability to pay. The severity of illness, as defined earlier, is significantly different between admitted and non-admitted. Appendix C in Additional file 1 illustrates how the three factors we used to define severity affect admissions.

As outlined in Methods, we first use a hypothesis test to compare admission rates between the insured patients and the uninsured. In the hypothesis test, we obtain \( z \)
= 49.3, which means that the probability that $p_1 = p_2$ is much smaller than 0.0001. Therefore, we can reject the null hypothesis and assert that insurance status affects hospitalizations.

Then, we examine how payer status affects admissions with a conditional logistic regression model. Table 3 presents our results. We use Private as the reference payer status. The significantly positive unstandardized coefficients of Multiple Payers and Government indicate that the admission probability for patients who are totally or partially insured by government, controlling for other variables in the model, is higher than that for patients who are insured by private insurers. Similarly, the significantly negative unstandardized coefficient of Uninsured indicates that the uninsured patients are less likely to be admitted than patients insured by private insurers.

The specific influence of payer status on admission can be further elaborated by the corresponding odds ratios. (1) The odds ratio of Multiple Payers is 1.956, greater than 1, indicating that the odds of being admitted increase by 95.6% when the variable Multiple Payers increases. Putting it another way, the odds of being admitted for patients with multiple payers is 95.6% higher than the odds for patients with a single payer or no payer. (2) The odds ratio of Government (odds ratio = 1.214) indicates that the odds of being admitted for patients who are insured by the government is 21.4% higher than the rest of the population. (3) The odds ratio of Uninsured (odds ratio = 0.815) is less than 1, indicating that uninsured patients are less likely to be admitted than insured patients when controlling for other variables.

The standardized coefficients in the 5th column in Table 3 reveal the relative importance of different variables on admissions. (For a graphical version, see Appendix B in Additional file 1). They indicate how many standard deviations of change in logit(Admissions) are associated with one standard deviation increase in the independent variables. For payer status, Multiple Payers has the greatest impact on
admissions, followed by the status of totally insured by Government. Uninsured status has the smallest impact on admissions.

Validity check

We validate the estimated model via its prediction accuracy on a dataset distinct from the one used for training. Figure 1 plots the ROC curve for a random split of the data into a training and test set. The model has an Area Under the ROC Curve (AUC) of 92%, indicating excellent predictive power.

Discussion

In an analysis of the relationship between admissions and payer status in a safety-net hospital, we found that uninsured patients are still less likely to be admitted than insured patients after controlling for demographics and prior medical conditions. These results are consistent with a streamline of work demonstrating inequity in access to health care[13, 11]. Generally, most uninsured people are in low-income families and may not be receiving public financial assistance for various reasons (e.g., age, income cutoff of financial assistance, undocumented immigrants, mental illness). As a consequence, it is also possible that they may not seek timely health care or receive treatment due to their economic, legal, or mental illness condition.

Further, we also found that patients who are partially or totally insured by government are more likely to be admitted than those who are insured by private insurers. Though we have controlled for prior medical conditions and age, still the association is significant. According to Exhibit A1 in the Additional file 1, among patients insured by government, 77.18% are totally or partially covered by Medicaid, therefore, we can infer that the high odds ratio of admission for patients insured by government are mainly driven by patients with Medicaid. Since our findings are not causal, this relationship only demonstrates an association between payer status and admissions. There may be several plausible explanations for this association: (1) Low-income patients tend to experience worse health quality and delayed diagnosis
and treatment, leading to worse health conditions when they are forced to seek health care. This is consistent with Adepoju et al.[8] and Giacovelli et.[9]. From this perspective, there may be more opportunities to design preventive care programs for patients that are insured by the government, compared to those insured by private insurers. (2) Patients with enough ability to pay may elect to seek care at a non-safety-net hospital setting, and these admissions are not present in our dataset. As suggested by Sutton et al.[15], inpatients in a safety-net hospital are usually associated with pregnancy and injuries, whereas inpatients in non-safety-net hospitals are more likely to be related to surgery. (3) Patients insured by Medicare and Medicaid include patients with disabilities who may require more frequent hospitalizations. However, this effect is mitigated by the fact that our model controls for the complexity of a patient’s health condition.

Our results also highlight additional information on factors that can influence a patient’s hospitalization. Demographics play a role in explaining the relationship between admissions and payer status. Among all factors included in our analysis, the number of other diseases for a patient (combinations) contributes the most to a hospitalization since this variable has the largest standardized coefficient. In addition, the standardized coefficients suggest that demographics are not a major factor.

There exist significant differences among disease types in admissions. Specifically, the top 3 disease types contributing to admission are Pregnancy, Injury & Poison and Endocrine (see Table 3 or Appendix B in Additional file 1). Unlike other similar studies focusing on a specific disease, our analysis includes all disease types, which can avoid potential selection biases.

Our results can lead to two policy recommendations. (1) Attention is needed on improving the access to care for vulnerable (low-income) patients, for example, by actively advertising free care programs, reaching out to community organizations
with better access to these individuals, or offering assurances that access to care is not linked to immigration procedures. (2) In order to reduce preventable admissions, basic preventive care services should be enhanced.

**Limitations**

First, the differences among payer groups may be biased by not controlling for lab tests and other unaccounted factors in our analysis, including disability status for those insured by government programs. Though we use surrogates for illness severity, they may not fully account for the true health status of a patient.

Second, the identification of admissions is imprecise. Admitted patients in our sample are limited to those who were actually admitted, omitting those who were given recommendations to be admitted but did not follow through, or elected to be admitted at a different hospital, or even moved or died. This type of label bias makes our results underestimate the probability of admission.

Third, the lack of patient source. Due to the lack of information on whether a patient has primary care though the hospital we considered, we cannot separate patients who get their primary care and those who come just for a hospitalization. This may lead to deficiency in historical information for some admitted patients.

Fourth, some estimates of independent variable coefficients may not be accurate, in part due to the dependence between payer status and disease types. Nevertheless, according to a rough rule in estimating the severity of collinearity resulting from dependence between variables,[20] the collinearity can be tolerated if the standardized coefficient is smaller than 1 or the unstandardized coefficient is smaller than 2. In our analysis, the standardized coefficients are all smaller than 1 except for Combinations, which suggests that the results are plausible.

**Conclusions**

In this paper, we focused on exploring the impact of payer status to a patient’s hospitalization. As far as we have known about this topic, this paper is one of the
few researches conducted from the level of a facility. Based on the insurance status, we stratified patients into five groups: government-sponsored, private-sponsored, multiple-sponsored, uninsured and other. We then used a conditional logistic regression model that is able to control for the influence of a patient’s illness severity to investigate the influence of other potential social factors. We found that coverage by Medicaid or Medicare plays a significant role in improving access to care (e.g., hospitalization) for low-income patients, but there might exist preventable admissions for this group of patients. For uninsured patients, inequity in terms of hospitalization still exists. Therefore, strategies to prevent hospitalizations for low-income insured patients and providing help for uninsured patients may be advisable.

Acknowledgements

The authors would like to thank Bill Adams for useful discussions.

Abbreviations

ICD9: International classification of diseases 9th version; HIV: human immunodeficiency virus; ER: emergency room; ROC curve: Receiver Operating Characteristic curve.

Funding

IP was supported by the National Science Foundation under grants IIS-1914792, DMS-1664644, and CNS-1645681, by the Office of Naval Research under grant N00014-19-1-2571, by the National Institutes of Health under grant R01-GM135930. JH was supported by the National Natural Science Foundation of China under grants 71720107003 and 72033003.

Authors’ contributions

YZ analyzed the data and drafted the manuscript. IP contributed to data acquisition, study design, and performed the critical revision of the manuscript. JH coordinated the study and revised the draft. All authors read and approved the final manuscript.

Declarations

Ethics approval and consent to participate

The data in this study is anonymous, therefore there are no informed consent of the person included in this study. The data protection office of this medical center approved the use of this study.

Consent for publication

Not applicable.
Availability of data and materials

The data that support the findings of this study are available from a Medical Center in Massachusetts but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Medical Center.

Competing interests

The authors declare that they have no competing interests.

Author details

References

1. LaPar DJ, Bhamidipati CM, Mery CM, Stukenborg GJ, Jones DR, Schirmer BD, et al. Primary payer status affects mortality for major surgical operations. Annals of surgery. 2010;252(3):544.

2. Long SK, Dimmock TH. Summary of Health Insurance Coverage and Health Care Access and Affordability in Massachusetts: 2015 Update. Urban Institute: Massachusetts; 2015. https://bluecrossmafoundation.org/sites/default/files/download/publication/MHRS_2015_Summary_FINAL_v02.pdf. Accessed November 2, 2020.

3. Hinman A, Bozic KJ. Impact of payer type on resource utilization, outcomes and access to care in total hip arthroplasty. The Journal of arthroplasty. 2008;23(6):9–14.

4. Petrou P, Samoutis G, Lionis C. Single-payer or a multipayer health system: a systematic literature review. Public health. 2018;163:141–152.

5. Newacheck PW, Stoddard JJ, Hughes DC, Pearl M. Health insurance and access to primary care for children. New England Journal of Medicine. 1998;338(8):513–519.

6. Allen LA, Smoyer Tomic KE, Wilson KL, Smith DM, Agodoa I. The inpatient experience and predictors of length of stay for patients hospitalized with systolic heart failure: comparison by commercial, Medicaid, and Medicare payer type. Journal of medical economics. 2013;16(1):43–54.

7. Roetzheim RG, Gonzalez EC, Ferrante JM, Pal N, Van Durme DJ, Krischer JP. Effects of health insurance and race on breast carcinoma treatments and outcomes. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2000;89(11):2202–2213.

8. Adepoju L, Wanjiku S, Brown M, Qu W, Williams M, Redfern RE, et al. Effect of insurance payer status on the surgical treatment of early stage breast cancer: data analysis from a single health system. JAMA surgery. 2013;148(6):570–572.

9. Giacovelli JK, Egorova N, Nowygrod R, Gelijns A, Kent KC, Morrissey NJ. Insurance status predicts access to care and outcomes of vascular disease. Journal of vascular surgery. 2008;48(4):905–911.

10. Khouja T, Burgette JM, Donohue JM, Roberts ET. Association between Medicaid expansion, dental coverage policies for adults, and children’s receipt of preventive dental services. Health services research. 2020;55(5):642–650.

11. Dzordzormenyoh MK. The Impact of Medicaid Expansion on Access to Healthcare in America: An Andersenian Perspective. University of South Dakota; 2019.

12. Mazurenko O, Balio CP, Agarwal R, Carroll AE, Menachemi N. The effects of Medicaid expansion under the ACA: a systematic review. Health Affairs. 2018;37(6):944–950.

13. Weissman JS, Gatsonis C, Epstein AM. Rates of avoidable hospitalization by insurance status in Massachusetts and Maryland. Jama. 1992;268(17):2388–2394.
14. Ferro EG, Secemsky EA, Wadhera RK, Choi E, Strom JB, Wasfy JH, et al. Patient Readmission Rates For All Insurance Types After Implementation Of The Hospital Readmissions Reduction Program. Health Affairs. 2019;38(4):585–593.

15. Sutton J, Washington R, Fingar K, Elixhauser A. Characteristics of Safety-Net Hospitals, 2014. HCUP Statistical Brief. 2016;(213).

16. Hadley J, Cunningham P. Availability of safety net providers and access to care of uninsured persons. Health services research. 2004;39(5):1527–1546.

17. Cai S, Miller SC, Nelson DL, Mukamel DB. The impact of Medicaid payer status on hospitalizations in nursing homes. Medical care. 2015;53(7):574.

18. Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. Proceedings of the IEEE. 2018;106(4):690–707.

19. Wooff D. Logistic regression: a self-learning text. JSTOR; 2004.

20. Menard S. Applied logistic regression analysis. vol. 106. Sage; 2002.

21. Organization WH. International Classification of Disease, 9th Revision(ICD9). Word Health Organization; 1999. https://simba.isr.umich.edu/restricted/docs/Mortality/icd_09_codes.pdf. Accessed November 2, 2020.

22. Agresti A. An introduction to categorical data analysis. John Wiley & Sons; 1996.

**Figures**

**Figure 1** ROC curve to reveal the model’s prediction performance

![ROC Curve](image-url)
Table 1  Features extracted for each patient.

| Ontology             | Number of Factors | Examples                                                                 |
|----------------------|-------------------|--------------------------------------------------------------------------|
| Demographics         | 6                 | Gender; Age; Race; Marital Status; Language; Zip-Code                    |
| Payer                | 5                 | Government-Sponsored (Medicare, Medicaid, Health safety net, commonwealth care); Private-sponsored (Commercial, Accident); Uninsured(Self-pay, Unknown); Other; Multiple-sponsor(Insured by both government and private sponsors) |
| Primary Diagnosis    | 22                | e.g., Infections (ICD9: 001-139); Neoplasms (ICD9: 140-239); Endocrine (ICD9: 240-259); Nutrition (ICD9: 260-269); Blood (ICD9: 280-289); Circulatory (ICD9: 390-459); Injury&Position (ICD9: 800-999); Illdefined (ICD9: 780-799); Supplementary 1 (ICD9: V01-V91) |
| Chronic Disease      | 1                 | e.g., Hypertensive disease (ICD9: 401.0-401.99); Diabetes mellitus (ICD9: 250.00-250.13, 250.22, 250.40-250.93); HIV (ICD9: 042.0-043.9, 079.53, v65.44); etc. |
| Prescription         | 1                 | Whether there are prescriptions or not                                  |
| Service by depart-   | 3                 | Inpatient, Outpatient, Emergency Room                                   |
| Admissions           | 22                | The cause of hospitalization.                                           |

1 ICD9 is a commonly used medical coding system for disease. See website [21] for full list of diseases.
2 Supplementary 1 is a disease type in ICD9, defined as Supplementary classification of factors influencing health status and contact with health services.
Table 2: Descriptive Characteristics

| Variables          | Admissions<sup>1</sup> | P-value<sup>2</sup> |
|--------------------|------------------------|---------------------|
|                    | =1                     | =0                  |
| Gender             |                        | 0.000               |
| Female             | 48.1                   | 51.9                |
| Race               |                        | 0.000               |
| Asian              | 2.77                   | 5.25                |
| Hispanic           | 19.1                   | 19.1                |
| Other Race         | 5.08                   | 7.69                |
| Unknown Race       | 0.7                    | 7.11                |
| White              | 37.4                   | 33.1                |
| Marital            |                        | 0.000               |
| Divorced           | 5.75                   | 2.72                |
| Other Marital      | 1.92                   | 5.05                |
| Separated          | 2.53                   | 1.36                |
| Single             | 56.1                   | 65.9                |
| Unknown Marital    | 0.02                   | 0.13                |
| Widow              | 5.47                   | 1.58                |
| PrscNum            | 0.52                   | 0.39                | 0.000               |
| Chronic            | 19.7                   | 80.3                | 0.000               |
| Combinations       | 6.65                   | 2.18                | 0.000               |
| Diagnosis          |                        | 0.000               |
| Blood              | 0.96                   | 0.36                |
| Circulatory        | 13.3                   | 4.12                |
| Congential Anomalies | 0.21               | 0.32                |
| Digestive          | 9.7                    | 6.97                |
| Endocrine          | 2.33                   | 3.04                |
| Genitourinary      | 4.72                   | 5.45                |
| Ildefined          | 4.38                   | 12.2                |
| Infections         | 4.02                   | 2.94                |
| Injury & Poison    | 13.1                   | 8.92                |
| Mental             | 1.64                   | 6.56                |
| Metabolic & Immune | 1.35                   | 1.38                |
| Musculoskeletal    | 5.81                   | 10.77               |
| Neoplasms          | 6.37                   | 2.94                |
| Nutrition          | 2.68                   | 7.73                |
| Pregnancy          | 16.0                   | 1.44                |
| Respiratory        | 6.18                   | 3.42                |
| Secondary          | 1.33                   | 0.50                |
| Skin               | 2.75                   | 4.64                |
| Supplementary 1    | 1.9                    | 13.0                |

*Continues*
Table 2 Descriptives Characteristics (Continued)

| Variables          | Admissions\(^1\) | P-value\(^2\) |
|--------------------|------------------|--------------|
|                    | =1               | =0           |
| Unknown Diagnosis  | 0.08             | 3.13         | -            |
| Payer              |                  | 0.000        |
| Multiple payers    | 8.24             | 5.08         | -            |
| Government         | 65.7             | 50.8         | -            |
| Private            | 23.1             | 35.0         | -            |
| Other Payers       | 1.51             | 3.70         | -            |
| Uninsured          | 1.44             | 5.44         | -            |
| ER Visits          |                  | 0.000        |
| =0 ER visit record | 86.2             | 73.9         | -            |
| =1 ER visit record | 8.22             | 19.6         | -            |
| >1 ER visit records| 5.61             | 6.48         | -            |
| Prior Admissions   |                  | 0.000        |
| =0 prior admissions| 66.6             | 99.3         | -            |
| =1 prior admissions| 16.6             | 0.53         | -            |
| >1 prior admissions| 16.7             | 0.15         | -            |

\(^1\) The 3rd and 4th column report the percentage of admitted and non-admitted patients, respectively, with the given feature for categorical variables or the mean over admitted patients for continuous features.

\(^2\) A Chi-square test is used to check whether the difference of proportions (percentages) among different categories of a categorical variable is significant. A t-test is used to check the null hypothesis that means of continuous variables among different categories are equal. The p-values are presented in the 5th column.
| Variables        | $\beta^1$ | $\beta^2$ | odds ratio$^3$ | P-value |
|------------------|-----------|-----------|----------------|---------|
| **Payer**        |           |           |                |         |
| Multiple payers  | 0.671     | 0.153     | (1.846, 2.072) | 0.000   |
| Government       | 0.194     | 0.097     | (1.176, 1.253) | 0.000   |
| Other Payers     | 0.503     | 0.091     | (1.496, 1.829) | 0.000   |
| Uninsured        | -0.204    | -0.044    | (0.744, 0.893) | 0.000   |
| **Gender**       |           |           |                |         |
| Female           | -0.286    | -0.143    | (0.731, 0.773) | 0.000   |
| **Race**         |           |           |                |         |
| Asian            | -0.207    | -0.045    | (0.754, 0.876) | 0.000   |
| Hispanic         | 0.021     | 0.008     | (0.983, 1.061) | 0.290   |
| Other Race       | -0.312    | -0.081    | (0.690, 0.756) | 0.000   |
| Unknown Race     | -1.182    | -0.286    | (0.273, 0.345) | 0.000   |
| White            | 0.315     | 0.149     | (1.324, 1.418) | 0.000   |
| **Marital**      |           |           |                |         |
| Divorced         | 0.162     | 0.028     | (1.099, 1.257) | 0.000   |
| Other Marital    | -0.281    | -0.059    | (0.694, 0.822) | 0.000   |
| Separated        | 0.044     | 0.005     | (0.948, 1.151) | 0.000   |
| Single           | -0.077    | -0.037    | (0.895, 0.957) | 0.000   |
| Unknown Marital  | -0.932    | -0.032    | (0.151, 1.030) | 0.000   |
| Widow            | 0.338     | 0.049     | (1.296, 1.516) | 0.000   |
| **PrescNum**     | -0.199    | -0.309    | (0.813, 0.827) | 0.000   |
| **Chronic**      | -0.589    | -0.194    | (0.527, 0.584) | 0.000   |
| **Combinations** | 0.337     | 1.113     | (1.395, 1.408) | 0.000   |
| **Diagnosis**    |           |           |                |         |
| Circulatory      | 0.543     | 0.124     | (1.457, 2.036) | 0.000   |
| Congential Anomalies | -0.654   | -0.036    | (0.392, 0.689) | 0.000   |
| Digestive        | 0.198     | 0.052     | (1.036, 1.436) | 0.018   |
| Endocrine        | -1.080    | 0.182     | (0.284, 0.406) | 0.000   |
| Genitourinary    | -0.586    | -0.132    | (0.471, 0.659) | 0.000   |
| Ildefined        | -1.356    | -0.426    | (0.218, 0.305) | 0.000   |
| Infections       | -0.455    | -0.079    | (0.533, 0.755) | 0.000   |
| Injury & Poison  | 0.835     | 0.245     | (1.958, 2.713) | 0.000   |
| Mental           | -1.835    | -0.431    | (0.133, 0.191) | 0.000   |
| Metabolic & Immune | -0.294   | -0.036    | (0.624, 0.890) | 0.001   |
| Musculoskeletal  | -1.081    | -0.324    | (0.288, 0.400) | 0.000   |
| Neoplasms        | 0.244     | 0.044     | (1.079, 1.511) | 0.005   |
| Nutrition        | -1.767    | -0.059    | (0.097, 0.300) | 0.000   |
| Pregnancy        | 2.581     | 0.477     | (11.19, 15.58) | 0.000   |
| Respiratory      | 0.236     | 0.045     | (1.069, 1.498) | 0.006   |

*Continues*
### Table 3 Descriptives Characteristics (Continued)

| Variables             | $\beta^1$ | $\beta^2$ | odds ratio$^3$ | P-value |
|-----------------------|-----------|-----------|----------------|---------|
| Secondary             | 0.034     | 0.003     | (0.842, 1.270) | 0.747   |
| Skin                  | -0.561    | -0.115    | (0.480, 0.678) | 0.000   |
| Supplementary1        | -2.346    | -0.745    | (0.080, 0.114) | 0.000   |
| Unknown Diagnosis     | -2.928    | -0.473    | (0.037, 0.076) | 0.000   |

$^1$ $\beta$ corresponds to the unstandardized coefficients from the regression model.

$^2$ $\beta^*$ corresponds to the standardized coefficients from the regression model. $\beta^*$ is calculated with the method suggested by Agresti[22] and Menard[20]: $\beta^* = \beta S_x$, where $S_x$ is the standard deviation of predictor $x$.

$^3$ The 5th column presents the 95% confidence interval of odds ratio.

### Additional Files

Additional file 1 — Appendix

This file contains the following content: (1) Summary characteristics of Payers; (2) The order of importance on influencing hospital admission; (3) Relationship between admissions and severity representatives; (4) The descriptive characters of variables.