EasyChair Preprint

№ 4945

Personalized Predictions for Unplanned Urinary Tract Infection Hospitalizations with Hierarchical Clustering

Lingchao Mao, Kimia Vahdat, Sara Shashaani and Julie Swann

January 30, 2021

Personalized Predictions for Unplanned Urinary Tract Infection Hospitalizations with Hierarchical Clustering

Lingchao Mao[*1], Kimia Vahdat[1], Sara Shashaani[1], Julie L. Swann[1]

[1]Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC, 27695, US

ABSTRACT

Urinary Tract Infection (UTI) is the one of the most frequent and preventable healthcare-associated infections in the US and an important cause of morbidity and excess healthcare costs. This study aims to predict the 30-day risk of a beneficiary for ***unplanned hospitalization for UTI****. Using 2008-12 Medicare fee-for-service claims and several public sources, we extracted 784 features, including patient demographics, clinical conditions, healthcare utilization, provider quality metrics, and community safety indicators. To address the challenge of high heterogeneity and imbalance in data, we propose a* ***hierarchical clustering*** *approach that leverages existing knowledge and data-driven algorithms to partition the population into groups of similar risk, followed by building a LASSO-Logistic Regression (LLR) model for each group. Our prediction models are trained on 237,675 2011 Medicare beneficiaries and tested on 230,042 2012 Medicare beneficiaries. We compare the clustering-based approach to a baseline LLR model using five performance metrics, including the area under the curve (AUC), the True Positive Rate (TPR), and the False Positive Rate (FPR). Results show that the hierarchical clustering approach achieves more accurate and precise predictions (AUC 0.72) than the benchmark model and offers more granular feature importance insights for each patient group.*

Keywords: Statistics; Cluster analysis; Data Analysis; Health care; Hospitals

1. INTRODUCTION

Clinical predictive analysis is of increasing interest to policy-makers, healthcare providers, and researchers with the potential to reduce healthcare costs and improve care quality [1–3]. Even a small reduction in potentially avoidable hospitalizations (PAH) would result in substantial savings in economic and human costs [4]. This study aims to build personalized prediction models for unplanned hospital admissions for Urinary Tract Infection (UTI). UTI is the most frequent and preventable healthcare-associated infection (HAI) in the US, one of the five most common ambulatory care-sensitive conditions (ACSC), and an important cause of morbidity and excess healthcare costs [4–8]. UTI not only results in patient discomfort but also increases the risk of PAH and discharge delays [5, 8].

One of the main challenges of predictive modeling using healthcare data is the large heterogeneity of patient profiles coupled with low disease occurrence rates. This heterogeneity and sparsity pose difficulties for conventional classification algorithms to achieve good classification performance without becoming overly complex. Bertsimas et al. partitioned the study population into five cost buckets to alleviate the heterogeneity of cost patterns; and used classification trees and clustering methods to divide data into more uniform groups, which improved predictions for healthcare costs [9]. Elbattah et al. used unsupervised learning to find coherent clusters of patients and showed that the clustering-aided models achieved higher accuracy in predicting the length of stay of hip fractures [10]. Beyan et al. proposed a hierarchical decomposition method that partitions data into smaller subsets, each with a different feature space, and showed improvement in classification performance with over twenty imbalanced data sets [11]. Therefore, we claim that by dividing the population into similar risk groups based on patient demographics, medical history, care quality, and environmental factors we can build more effective prediction models for each group.

Although a number of clustering algorithms have been applied to healthcare data, such as partition clustering, agglomerative clustering, and density clustering, they are purely data-driven and highly dependent on the major patterns of the data [1–3, 11–13]. In addition to traditional data-driven algorithms, we aim to leverage existing knowledge from literature and domain to define representative patient groups. Our proposed framework adopts a hierarchical structure because of its advantage to model dependence relationships between levels of the hierarchy.

The contributions of this paper are two-fold:

[*] Corresponding author: E-mail: lmao3@ncsu.edu

- a hierarchical clustering framework that leverages both existing knowledge and data-driven patterns to group patients with similar risk levels with respect to unplanned UTI admissions. This approach can also be applied to non-healthcare problems where data is highly heterogeneous and imbalanced, and domain knowledge can be used to guide focused modeling; and
- monthly probabilistic predictions for Medicare beneficiaries' risk for unplanned UTI admissions and interpretable insights about the most relevant variables, which may facilitate the design of interventions.

To our best knowledge, our study is the first to predict UTI hospitalization as small as monthly intervals. The closest in the literature is a study by Carter, which focuses on the nursing home population and provides quarterly predictions (pseudo R squared 0.0931) [14]; and another by Saver et al., which predicts several acute and chronic hospitalizations for a year-long interval (AUC 0.87) [15]. While these studies have longer prediction intervals and a smaller study population, they used logistic regression and a similar set of variables.

The remaining of this paper is organized into four sections (Figure 1). In Section 2 we discuss the data used in this study. In Section 3 we provide details about our hierarchical clustering method. Section 4 presents a summary of results and comparison with a baseline model. Lastly, we conclude the paper with a summary of findings.

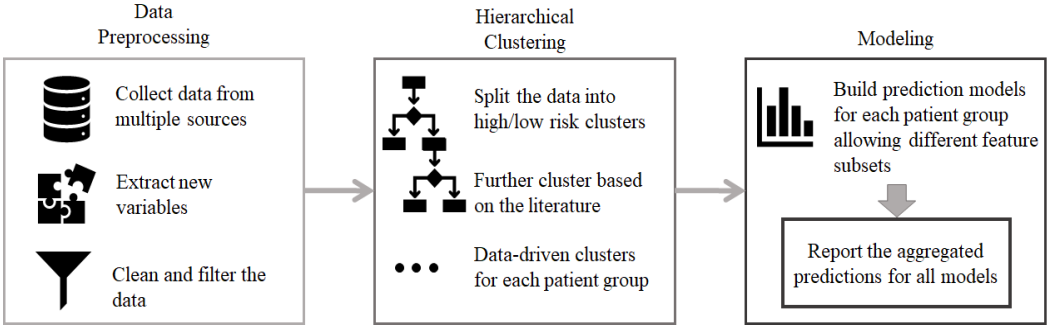

Figure 1: Road map of the modeling procedure

2. DATA COLLECTION AND VARIABLE QUANTIFICATION

This study uses 2008-2012 Medicare Limited Data Sets (LDS) which contain a 5% random sample of seven types of fee-for-service claims annually. Data from 2008-2011 are used to capture patient clinical history; data from April to November of 2011 are used for model training; and from 2012 are used for model testing and performance evaluation. In addition to medical claims, we collect multiple public data [16–25] to create relevant predictors. From the prepossessing, we obtain a patient-month data set (where rows correspond to patient's data for each month of the year) with 784 features including patient demographics, clinical history, healthcare utilization and spending, provider quality metrics, and community safety metrics (Table 1) for beneficiaries who had at least one inpatient, outpatient, or Skilled Nursing Facility (SNF) claim during the year. These features were identified from many studies, more detail is provided below.

Table 1: Summary of the predictors considered in the model

Demographics	Age, gender, race, low income, managed care, supplemental insurance, socioeconomic status, smoking, obesity
Clinical History	Acute and chronic CCS conditions and their aggregates (neuro, heart, diabetes, cognitive, alcohol substance abuse, cancer), ESRD, immunocompromised, post transplant, number of CCS conditions
Healthcare Utilization	Number of inpatient, outpatient, SNF, carrier, Durable Medical Equip-ment (DME), homehealth, hospice claims in the last one, three, six, and twelve months [26]; past and current nursing homestay; elixhauser comorbidity index [27]; number of specialty visits in the last month (allergy, neurology, endocrinology, car-diology); number of emergency room, physician, hospitalization, ICU, CCU, and Oncology stays in the last one and three months; length of stay in hospital and SNF in the last one and three months [28, 29]
Healthcare Spending	Medicare and non-medicare paid costs of inpatient, outpatient, SNF, carrier, DME, homehealth, hospice claims in the last one, three, six, and twelve months [28, 29]
Most Recent Provider's Quality Metrics	Hospital overall rating, number of beds, count of outpatient procedures, emergency room volume [14]; several disease-specific death rates, complications rates, postoperative complication rates, infection rates, and readmission rates [21, 22]

Community Quality Metrics	Rural indicator, household income [15]; state-level flu activity, vaccine effectiveness, air quality [17, 25]; region safety scores; population statistics about race, education, income, access to care and food, etc. [18, 19, 23]

To define our target event, unplanned hospitalization for UTI, we analyze whether a patient's inpatient claim satisfies the Prevention Quality Indicators (PQI) criteria put forth by Agency for Healthcare Research and Quality [30]. We use the Clinical Classification System (CCS) developed by AHRQ to compute 285 CCS variables based on ICD-9 diagnosis codes. The data for acute CCS conditions are transformed into two binary indicators of length since the first diagnosis (less than six months ago or not). We aggregate specific CCS variables to alleviate data sparsity while making the presentation clinically meaningful (Table 1).

3. METHOD

In this section, we discuss the hierarchical clustering modeling approach and the evaluation criteria employed.

3.1. HIERARCHICAL CLUSTERING APPROACH

As discussed in Section 1, the main challenge of developing predictive models using healthcare data is the heterogeneity of patterns coupled with scarcity of events. We propose a novel approach to address this challenge, referenced as hierarchical clustering. This approach partitions patient-month data points into more uniform groups, then builds targeted prediction models with coefficients and feature sets unique to each group. The key advantage of this approach is the ability to use known relationships identified from literature and domain knowledge to categorize archetypical patient groups meaningful for providers and based on their resemblance in risk of UTI hospitalization.

The model building process involves four steps:

Step 1. Overall partitioning into high versus low percentage of event occurrence. Since data is highly imbalanced, the desired effect is to first separate the population into two groups such that the majority of event occurrences are concentrated in a small group. To identify the best partition rule, we use the R implementation of Classification and Regression Trees (CART) in the *caret* library with adjusted cost functions to emphasize more on correctly identifying events than non-events [31]. We include high level variables that indicate healthcare utilization and UTI history so that the results are applicable to a larger population. The variables we include in CART are the number of inpatient, outpatient, carrier, SNF, hospice, homehealth, and DME claims in the previous three, six, and twelve months; the Medicare and non-Medicare paid costs of these seven types of claims in the previous three, six, and twelve months; and previous UTI.

Step 2. Categorizing archetypical patient groups meaningful for providers. For the subset with the highest prevalence of events from Step 1, we categorize archetypical patient groups intended to be meaningful for providers. The goal is to define types of patient populations that have fundamentally different conditions that may drive differences in regression models. To promote understandability, the choice and order of these branches are based on domain knowledge and results from the literature.

The first differentiating group we define is those who are on Medicare because they have End-Stage Renal Disease (ESRD). These individuals may be younger than 65, and they have been identified to have a significant disease that may relate to UTI risk [4, 15, 32]. In the next level of the hierarchy, we consider people who have been in a nursing home [33, 34] identified through the algorithm developed by Koroukian et al. [35]. Nursing home residents are likely those who need help with activities of daily living (ADLs) and/or have difficulties with walking, hearing or seeing [36]. These conditions cause them to be at greater risk for admissions or adverse events [8]. This setting is different from Skilled Nursing Facilities (SNF) because the residents may be self-financing, and nursing homes tend to be for longer occupation. Literature has also shown that UTIs can be associated with urinary-related cancer [4, 15, 28] and with mental conditions such as dementia, delirium, and Parkinson's disease [15, 34, 37–39]; so these patient groups are identified for lower levels of the hierarchy.

Step 3. Data-driven clustering to improve the predictive power of the models. We apply CART on each patient group identified in Step 2 to obtain candidate data-driven clusters. The variables we include in CART consists of the set of features discussed in Step 1, which capture general healthcare utilization. To decide whether to employ a cluster and when to stop branching, we build brute-force regression models for the two children nodes and the parent node before branching and choose the option that achieves higher AUC on the hold-out data set.

Step 4. Regularized regression model to provide monthly risk predictions. We build prediction models for each resulting cluster. We choose the Lasso-Logistic Regression model (LLR) because it has been shown effective in largely imbalanced data [40]. The penalty parameter is tuned separately for each cluster using 3-fold cross-validation such that the chosen parameter minimizes the deviance of the predicted values from the logistic regression model [41].

The 784 features from Table 1 are provided to all levels of the hierarchy, and L1 norm regularization selects the most relevant features in each cluster. Youden's index is used to select the probability threshold for each cluster [44].

3.2. BASELINE APPROACH

As a baseline model, we run LLR using the same settings described in Step 4 on the training data before clustering. In other words, the baseline approach builds one prediction model for the entire population. Parameter tuning and threshold selection are also performed only once. This is the most relevant benchmark, considering modeling procedure and the data, that we found in the literature as described in Section 1.

3.3. EVALUATION METRICS

We use a combination of discrimination and calibration measures to assess model performance. The Area Under the Curve (AUC) evaluates the likelihood that the predicted probability of an event instance is higher than that of a non-event instance. However, AUC fails to measure the goodness of fit when data is imbalanced [42]. Therefore, we include the TPR and FPR, which are especially useful for imbalanced class problems [43]. The former indicates the percent of events *correctly* predicted by the model out of the total event instances, and the latter measures the percent of non-events that the model *incorrectly* predicts as positive. We use TPR and FPR to understand the unplanned admissions that are captured by the model (TPR), as well as the potential cost associated if interventions are used unnecessarily (FPR). In addition, we report the Sensitivity at Low Alert Rates (SLA) at 1%, which measures the TPR for instances that are given the highest risks. Lastly, we include accuracy, which is another common metric to measure the percentage of correctly predicted event and non-event over all data points.

4. RESULTS

Our predictions focus on patients who had at least one inpatient or SNF claim during the year. To ensure complete health profiles, we exclude beneficiaries who are Medicare part A and B enrollees for only part of the year, enrolled in managed care, with supplemental insurance, or with disability [44]. Table 2 shows summary statistics of the preprocessed data.

Table 2: Descriptive statistics of the 2011 and 2012 study population.

	2011		**2012**	
	Count	**Percentage**	**Count**	**Percentage**
Total beneficiaries	1,257,485	100%	1,274,142	100%
Age above 65	1,133,412	90%	1,149,054	90%
Disability	229,377	18%	245,031	19%
Male	518,698	41%	528,557	41%
Previous unplanned admission due to UTI	10,203	1%	10,099	1%
Had at least an inpatient claim	348,866	28%	327,198	26%
Had at least an SNF claim	96,163	8%	90,572	7%
After exclusions	237,675		230,042	

We obtain a hierarchical structure with 12 knowledge-based and data-driven clusters from training data, as visualized in Figure 2. The most important partition based on event prevalence identified by CART (Section 3.1 step 1) is whether the patient had a historical diagnosis of UTI or any medical claims in the past year. For the lower event occurrence group, data-driven clustering suggests assessing whether the beneficiary had been admitted to SNF in the last month or to ICU in the last three months. For people who had no urinary-related cancer but had cognitive conditions, this step suggests grouping patients based on inpatient costs in the last 12 months. Similarly, for the beneficiaries who did not have an inpatient or SNF visit in the last 12 months, this step suggests using more than 10 carrier claims (which include physician visits) in the last year as the best split criteria.

For each cluster, the model intercept proxies the base risk level, and the set of selected features shows which factors are most likely to be associated with UTI hospitalization for that particular patient group. The features' estimated coefficients indicate their quantified impact towards the base risk of their cluster. A positive coefficient indicates an increase in risk, and a negative coefficient is associated with a decrease in risk. Note that the base risk should be interpreted in combination with the variability of the coefficients; a cluster with few features, that have a small effect on the predicted probability, may have a lower risk than a cluster with highly variable coefficients. Therefore, we calculate the summation of the intercept and coefficient variance as the underlying risk of each cluster, which is shown with color-coding in Figure 2; the darker the cluster, the higher the risk.

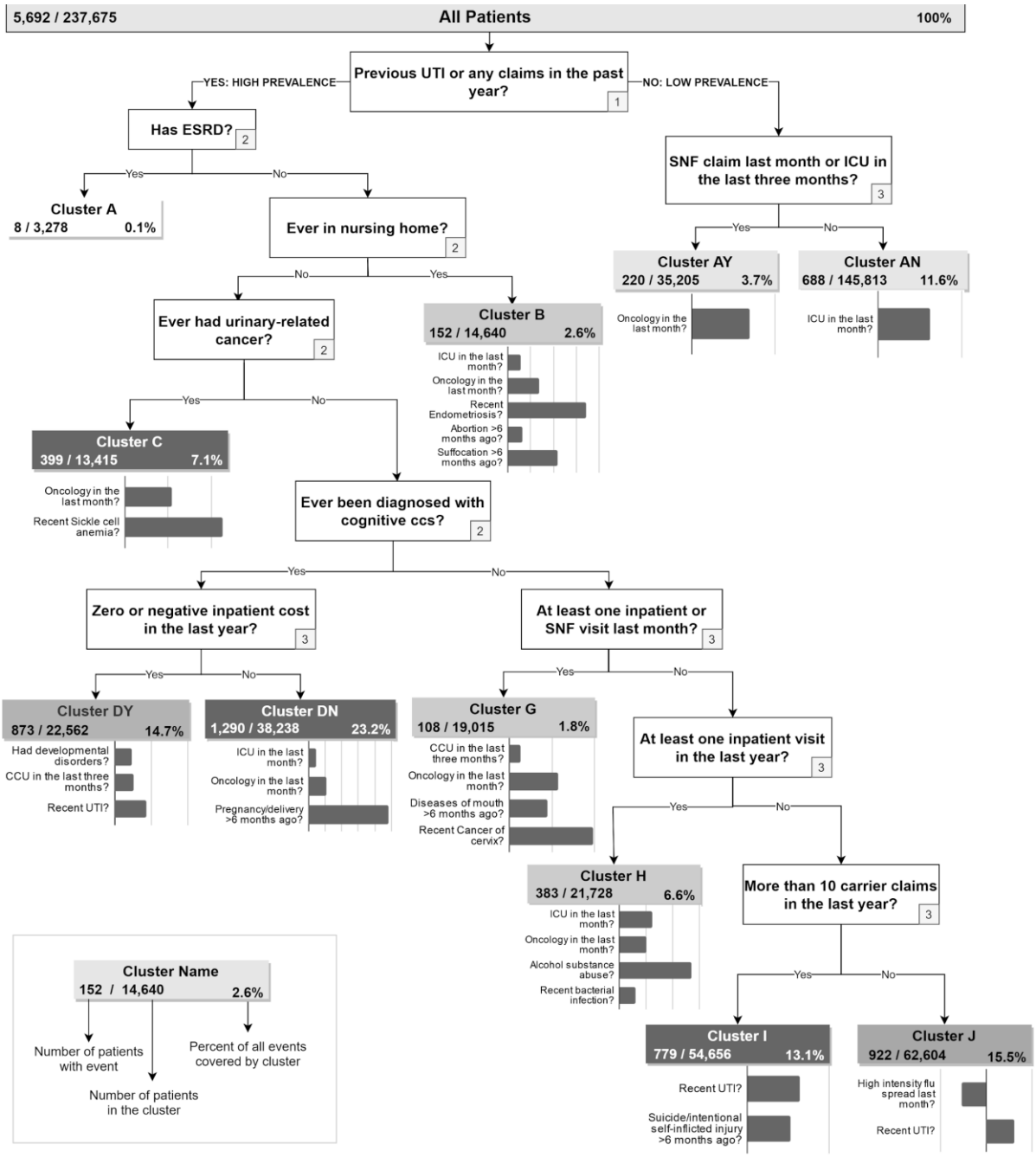

Figure 2: Hierarchical structure is visualized in the form of a tree. For each cluster, the ratio of the number of patients with the event to the number of patients in the cluster is shown on the lower-left corner of the box; the percentage of all event rows contained in the cluster is shown on the lower right corner. The small square in the lower right corner of each decision node indicates which step (defined in Section 3.1) it belongs to. The plots display the relative coefficient values for the selected features (note that only features with a coefficient greater than 0.5 are displayed for visualization). Clusters are color-coded based on the average risk obtained from the model intercept and coefficient variance; darker colors correspond to higher risk. Cluster A received zero predictions due to its low event occurrence rate and low event coverage percentage.

The most selected features across clusters include a previous month visit to oncology, ICU, or CCU, and whether the patient had a recent UTI. Other variables specific to each cluster include endometriosis, developmental disorders, pregnancy or delivery, abortion, cancer of the cervix, bacterial infection, alcohol substance abuse, sickle cell anemia, self-inflicted injury, suffocation, and flu intensity last month. Additional considerations should be taken into account while interpreting the selected features. Some of these variables may not be directly related to UTI but are proxies for their underlying health and/or environmental conditions that are associated with a risk to UTI. For example, the flu

intensity variable can be interpreted as a proxy for weather seasonality, which is shown to be correlated with UTI hospitalizations in previous studies [45, 46]. Another example is suffocation, which may be an indicator of specific patient characteristics or behaviors such as intentional injuries [47]. The recency in the wording of these variables relies on health history recorded by Medicare claims. If a patient had their diagnosis more than six months ago but is added to the system not long ago, our data will still indicate that their first diagnosis was recent. Therefore, conditions that have very low probabilities of development after the age of 65, such as endometriosis or pregnancy, may be aliases for a recent addition of the patient's data or diagnosis code to the system. These variables can still be proxies for recent visits with providers who recorded previous health history.

In the baseline model, the variables with the highest coefficients selected are previous month visits to oncology and UTI history, which coincide with the top features selected from the hierarchical clustering approach. One advantage of the latter is providing more personalized feature importance summaries for each group of patients. The prediction performance of these two modeling approaches is summarized in Table 3. The clustering-based approach achieves a higher AUC (0.72) than the baseline (0.63), which means the model is more likely to predict a higher risk for instances that UTI admission actually occurred than a non-event instance. The higher accuracy score also indicates that the clustering-based approach predicts both events and non-events more accurately than the baseline approach. Although the baseline model achieves a slightly higher TPR (0.84), the model overpredicts many patients resulting in a high percentage of false positives (0.67). The clustering-based approach significantly reduces false positives (FPR 0.43) while maintaining a reasonable TPR (0.77). Therefore, we conclude that the hierarchical clustering approach achieves more accurate and precise predictions than the approach without clustering.

Table 3: Comparing results between baseline and hierarchical clustering-based LLR. The latter achieves higher AUC, SLA, accuracy, and a lower FPR than the former. Slightly higher TPR in baseline LLR is due to classifying a lot of data points as positive, indicated by high FPR.

	AUC	**SLA 1%**	**TPR**	**FPR**	**Accuracy**
Baseline LLR	0.63	0.03	0.84	0.67	0.33
Hierarchical clustering-based LLR	**0.72**	**0.04**	0.77	**0.43**	**0.57**

5. CONCLUSION

One of the main challenges of predictive healthcare analytics is the large heterogeneity of patient patterns coupled with high data imbalance. The hierarchical clustering approach proposed in this paper tackles this challenge by leveraging existing knowledge about UTI as well as data-driven algorithms to identify representative patient groups, then building personalized prediction models for each group. This approach starts by separating patients into two major clusters differentiated by high and low event prevalence. Then knowledge from literature and domain are used to define archetypical patient groups intended to be meaningful to providers. These rules include whether a patient has ESRD, nursing home residence, urinary-related cancer, and cognitive diseases. These are either disease-based characteristics that are often positively correlated with risk to UTI hospitalizations or frailty indicators that suggest the patient's vulnerability. The lower levels of the hierarchy are data-driven clusters that are associated with general healthcare utilization, such as inpatient, SNF, and carrier visits.

The prediction performance shows that the hierarchical clustering-based models achieve more accurate and precise predictions than the approach without clustering. Another advantage of this approach is to provide more personalized insight on which factors are most relevant to each patient group, instead of a single feature importance summary for the entire population. The variables most associated with UTI hospitalizations amongst all patient groups are whether the patient had a recent UTI diagnosis, as identified by previous studies [4]; or at least one oncology, ICU, or CCU visit in the previous month. This result agrees with studies that showed that about 15% of the patients admitted to acute hospitals receive a urinary catheter during their stay, after which infection frequently occurs, as ICU and CCU visits proxy the use of catheters [5]. Additional feature insights for each of the twelve patient groups are discussed in Section 4.

The structure we have chosen for the tree is subject to our literature review and domain knowledge. For instance, we locate the nursing home variable at a higher level than urinary-related cancer because we believe that the frailty condition associated with nursing home residency dominates the specific health characteristics of urinary cancer. Other researchers could make different choices based on the knowledge they gather. In future studies, we suggest this framework to be used with more rigorous causal tools. The key contribution of the hierarchical clustering approach is providing a framework that can leverage existing knowledge to identify target groups meaningful to practitioners and that can be integrated with data-driven algorithms to build personalized prediction models for each representative group.

Although the LLR model was used to compare the performance of the hierarchical clustering approach with the non-clustering approach, other machine learning models may be used with the hierarchical clustering framework by modifying Step 4 in Section 3. The hierarchical clustering approach can also be applied to non-healthcare problems where data is highly heterogeneous and imbalanced, and domain knowledge is available to guide focused modeling.

6. Limitations and Future Research

In this study, data were limited to Medicare fee-for-service beneficiaries so health insights may not apply to all populations. We also rely on the accuracy and completeness of diagnosis from the claims to compute our predictors. In addition, studies have shown that the usage of urinary catheter is closely associated with UTI [5, 33]. Usage is not indicated in all claims like inpatient, however, adding an indicator for catheter would further improve the model predictions. Future studies may take these into account to build more accurate models for UTI.

Acknowledgements

The authors gratefully acknowledge the Centers for Medicare and Medicaid Services (CMS) for providing medical claims data used in this study. Partial support was provided by Dr. Joseph Agor from Oregon State University, and graduate students James McKenna from Oregon State University, and Mina Mohammadi, Akash Pateria, and Prasanth Yadla from North Carolina State University for data preprocessing.

References

[1] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," *Advances in Neural Information Processing Systems, Nips*, pp. 3512–3520, 2016.

[2] O. Hasan *et al.*, "Hospital readmission in general medicine patients: a prediction model," *J. Gen. Intern. Med.*, vol. 25, no. 3, pp. 211–219, 2010.

[3] J. Donzé, D. Aujesky, D. Williams, and J. L. Schnipper, "Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model," *JAMA Intern. Med.*, vol. 173, no. 8, pp. 632–638, 2013.

[4] E. G. Walsh, J. M. Wiener, S. Haber, A. Bragg, M. Freiman, and J. G. Ouslander, "Potentially avoidable hospitalizations of dually eligible Medicare and Medicaid beneficiaries from nursing facility and home-and community-based services waiver programs," *J. Am. Geriatr. Soc.*, vol. 60, no. 5, pp. 821–829, 2012.

[5] S. Saint, J. A. Meddings, D. Calfee, C. P. Kowalski, and S. L. Krein, "Catheter-associated urinary tract infection and the Medicare rule changes." American College of Physicians, 2009.

[6] S. Saint *et al.*, "Translating health care--associated urinary tract infection prevention research into practice via the bladder bundle," *Jt. Comm. J. Qual. Patient Saf.*, vol. 35, no. 9, pp. 449–455, 2009.

[7] J. Billings, L. Zeitel, J. Lukomnik, T. S. Carey, A. E. Blank, and L. Newman, "Impact of socioeconomic status on hospital use in New York City," *Health Aff.*, vol. 12, no. 1, pp. 162–173, 1993.

[8] K. T. Unroe, J. L. Carnahan, S. E. Hickman, G. A. Sachs, Z. Hass, and G. Arling, "The complexity of determining whether a nursing home transfer is avoidable at time of transfer," *J. Am. Geriatr. Soc.*, vol. 66, no. 5, pp. 895–901, 2018.

[9] D. Bertsimas *et al.*, "Algorithmic Prediction of Health-Care Costs," *Oper. Res.*, vol. 56, no. 6, pp. 1382–1392, Dec. 2008.

[10] M. Elbattah and O. Molloy, "Clustering-Aided approach for predicting patient outcomes with application to Elderly Healthcare in Ireland," 2017.

[11] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, pp. 1653–1672, 2015.

[12] R. Nithya, P. Manikandan, and D. Ramyachitra, "Analysis of clustering technique for the diabetes dataset using the training set parameter," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 9, pp. 166–169, 2015.

[13] G. Ogbuabor and F. N. Ugwoke, "Clustering algorithm for a healthcare dataset using silhouette score value," *International Journal of Computer Science & Information Technology*, vol. 10, no. 2, pp. 27–37, 2018.

[14] M. W. Carter, "Factors Associated with Ambulatory Care—Sensitive Hospitalizations among Nursing Home Residents," *J. Aging Health*, vol. 15, no. 2, pp. 295–331, May 2003.

[15] B. G. Saver, C.-Y. Wang, S. A. Dobie, P. K. Green, and L.-M. Baldwin, "The central role of comorbidity in predicting ambulatory care sensitive hospitalizations," *Eur. J. Public Health*, vol. 24, no. 1, pp. 66–72, 2014.

[16] "CMS' SSA to FIPS State and County Crosswalk," *The National Bureau of Economic Research*. 2011, [Online]. Available: https://data.nber.org/data/ssa-fips-state-county-crosswalk.html.

[17] "Immunization," *Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System*. 2011, [Online]. Available: https://www.cdc.gov/brfss/index.html.

[18] "County Health Rankings," *County Health Rankings and Roadmaps*. 2011, [Online]. Available: https://www.countyhealthrankings.org/.

[19] “Population Census Elderly Living alone,” *United States Census Bureau*. 2011, [Online]. Available: https://data.census.gov/cedsci/.
[20] “Population Census,” *United States Census Bureau*. 2011, [Online]. Available: https://data.census.gov/cedsci/.
[21] “Hospital Compare Dataset,” *Centers for Medicare & Medicaid Services*. 2011, [Online]. Available: https://data.medicare.gov/data/hospital-compare.
[22] “Nursing Home Compare Dataset,” *Centers for Medicare & Medicaid Services*. 2011, [Online]. Available: https://data.medicare.gov/data/nursing-home-compare.
[23] “Public Use Files HRR Table for Beneficiaries 65 and older,” *Centers for Medicare & Medicaid Services*. 2011, [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV_PUF.
[24] “Diabetes Atlas,” *United States Diabetes Surveillance System (USDSS)*. 2011, [Online]. Available: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html.
[25] “Weekly U.S. Influenza Surveillance Report,” *Centers for Disease Control and Prevention*. 2011, [Online]. Available: https://www.cdc.gov/flu/weekly/index.htm.
[26] C. Ma, M. D. McHugh, and L. H. Aiken, “Organization of hospital nursing and 30-day readmissions in Medicare patients undergoing surgery,” *Med. Care*, vol. 53, no. 1, p. 65, 2015.
[27] J. C. Will, I. A. Nwaise, L. Schieb, and Y. Zhong, “Geographic and racial patterns of preventable hospitalizations for hypertension: Medicare beneficiaries, 2004--2009,” *Public Health Rep.*, vol. 129, no. 1, pp. 8–18, 2014.
[28] Z. Moghadamyeghaneh, M. J. Stamos, and L. Stewart, “Patient Co-morbidity and functional status influence the occurrence of hospital acquired conditions more strongly than hospital factors,” *J. Gastrointest. Surg.*, vol. 23, no. 1, pp. 163–172, 2019.
[29] J. K. Chan, A. B. Gardner, A. K. Mann, and D. S. Kapp, “Hospital-acquired conditions after surgery for gynecologic cancer—An analysis of 82,304 patients,” *Gynecol. Oncol.*, vol. 150, no. 3, pp. 515–520, 2018.
[30] A. Q. Indicators, “Prevention Quality Indicators Technical Specifications,” *Department of Health and Human Services. Agency for Healthcare Research and Quality*, 2001.
[31] M. Kuhn, “Building Predictive Models in R Using the caret Package,” *Journal of Statistical Software, Articles*, vol. 28, no. 5, pp. 1–26, 2008.
[32] S. B. Naqvi and A. J. Collins, “Infectious complications in chronic kidney disease,” *Adv. Chronic Kidney Dis.*, vol. 13, no. 3, pp. 199–204, 2006.
[33] H. L. Wald, A. Ma, D. W. Bratzler, and A. M. Kramer, “Indwelling urinary catheter use in the postoperative period: analysis of the national surgical infection prevention project data,” *Arch. Surg.*, vol. 143, no. 6, pp. 551–557, 2008.
[34] D. C. Grabowski, A. J. O’Malley, and N. R. Barhydt, “The costs and potential savings associated with nursing home hospitalizations,” *Health Aff.*, vol. 26, no. 6, pp. 1753–1761, 2007.
[35] S. M. Koroukian, F. Xu, and P. Murray, “Ability of Medicare claims data to identify nursing home patients: a validation study,” *Med. Care*, vol. 46, no. 11, p. 1184, 2008.
[36] “Nursing Homes.” https://www.healthinaging.org/age-friendly-healthcare-you/care-settings/nursing-homes (accessed Oct. 15, 2020).
[37] A. W. Willis *et al.*, “Neurologist-associated reduction in PD-related hospitalizations and health care expenditures,” *Neurology*, vol. 79, no. 17, pp. 1774–1780, 2012.
[38] K. Dharmarajan *et al.*, “Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia,” *JAMA*, vol. 309, no. 4, pp. 355–363, 2013.
[39] E. L. Sampson, M. R. Blanchard, L. Jones, A. Tookman, and M. King, “Dementia in the acute hospital: prospective cohort study of prevalence and mortality,” *Br. J. Psychiatry*, vol. 195, no. 1, pp. 61–66, 2009.
[40] H. Wang, Q. Xu, and L. Zhou, “Large unbalanced credit scoring using lasso-logistic regression ensemble,” *PLoS One*, vol. 10, no. 2, p. e0117844, 2015.
[41] P. D. Allison and Others, “Measures of fit for logistic regression,” in *Proceedings of the SAS global forum 2014 conference*, 2014, pp. 1–13.
[42] D. W. Homer, S. Lemeshow, and R. X. Sturdivant, “Applied logistic regression, 3rd edn. Hoboken.” NJ: Wiley, 2013.
[43] S. Raschka, “An Overview of General Performance Metrics of Binary Classifier Systems,” *arXiv [cs.LG]*, Oct. 17, 2014.
[44] E. Mokyr Horner and M. R. Cullen, “Linking individual medicare health claims data with work-life claims and other administrative data,” *BMC Public Health*, vol. 15, p. 995, Sep. 2015.
[45] J. E. Anderson, “Seasonality of symptomatic bacterial urinary infections in women,” *J. Epidemiol. Community Health*, vol. 37, no. 4, pp. 286–290, Dec. 1983.
[46] P.-C. Hsu, Y.-C. Lo, P.-Y. Wu, J.-W. Chiu, and M.-J. Jeng, “The relationship of seasonality and the increase in urinary tract infections among hospitalized patients with spinal cord injury,” *J. Chin. Med. Assoc.*, vol. 82, no. 5, pp. 401–406, May 2019.
[47] R. Sasso, R. Bachir, and M. El Sayed, “Suffocation Injuries in the United States: Patient Characteristics and Factors Associated with Mortality,” *West. J. Emerg. Med.*, vol. 19, no. 4, pp. 707–714, Jul. 2018.