Published in final edited form as:
Intell Based Med. 2023 ; 7: . doi:10.1016/j.ibmed.2023.100093.

Novel integration of governmental data sources using machine learning to identify super-utilization among U.S. counties

Iben M. Ricket[a,*], **Michael E. Matheny**[b,c,d,e], **Todd A. MacKenzie**[f], **Jennifer A. Emond**[f,g], **Kusum L. Ailawadi**[h], **Jeremiah R. Brown**[a,f]

[a]Department of Epidemiology, Dartmouth Geisel School of Medicine at Dartmouth, Hanover, NH, USA

[b]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[c]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[d]Division of General Internal Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[e]Geriatric Research Education and Clinical Care Center, Tennessee Valley Healthcare System VA, Nashville, TN, USA

[f]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

[g]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

[h]Tuck School of Business at Dartmouth, Hanover, NH, USA

Abstract

Background: Super-utilizers consume the greatest share of resource intensive healthcare (RIHC) and reducing their utilization remains a crucial challenge to healthcare systems in the United States (U.S.). The objective of this study was to predict RIHC among U.S. counties, using routinely collected data from the U.S. government, including information on consumer spending,

[*]Corresponding author. Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hinman Box 7920, Hanover, NH, 03755, USA. Iben.ricket.gr@dartmouth.edu (I.M. Ricket).

CRediT author statement
Iben M. Ricket: conceptualization, methodology, software, formal analysis, investigation, resources, data curation, writing-original draft, writing-reviewing and editing.
Michael Matheny: conceptualization, methodology, formal analysis, writing-reviewing, and editing.
Todd A. MacKenzie: conceptualization, methodology, software, formal analysis, writing-reviewing and editing.
Jennifer A. Emond: conceptualization, methodology, validation, supervision, formal analysis, writing-original draft, writing-reviewing and editing.
Kusum L. Ailawadi: conceptualization, methodology, supervision, formal analysis, writing-original draft, writing-reviewing and editing.
Jeremiah R. Brown, conceptualization, methodology, resources, supervision, writing-reviewing and editing.

Declaration of competing interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2023.100093.

offering an alternative method for identifying super-utilization among population units rather than individuals.

Methods: Cross-sectional data from 5 governmental sources in 2017 were used in a machine learning pipeline, where target-prediction features were selected and used in 4 distinct algorithms. Outcome metrics of RIHC utilization came from the American Hospital Association and included yearly: (1) emergency rooms visit, (2) inpatient days, and (3) hospital expenditures. Target-prediction features included: 149 demographic characteristics from the U.S. Census Bureau, 151 adult and child health characteristics from the Centers for Disease Control and Prevention, 151 community characteristics from the American Community Survey, and 571 consumer expenditures from the Bureau of Labor Statistics. SHAP analysis identified important target-prediction features for 3 RIHC outcome metrics.

Results: 2475 counties with emergency rooms and 2491 counties with hospitals were included. The median yearly emergency room visits per capita was 0.450 [IQR:0.318, 0.618], the median inpatient days per capita was 0.368 [IQR: 0.176, 0.826], and the median hospital expenditures per capita was $2104 [IQR: $1299.93, 3362.97]. The coefficient of determination (R^2), calculated on the test set, ranged between 0.267 and 0.447. Demographic and community characteristics were among the important predictors for all 3 RIHC outcome metrics.

Conclusions: Integrating diverse population characteristics from numerous governmental sources, we predicted 3-outcome metrics of RIHC among U.S. counties with good performance, offering a novel and actionable tool for identifying super-utilizer segments in the population. Wider integration of routinely collected data can be used to develop alternative methods for predicting RIHC among population units.

Keywords

Resource intensive healthcare utilization; Machine learning; Population prediction models; Super-utilizers

1. Introduction

In 2017, nearly one-third of total healthcare expenditures in the United States (U.S.) were dedicated to hospital services [1]. About 5% of the U.S. population, often called “super-utilizers,” are responsible for over 50% of healthcare expenditures, a significant portion of which is driven by acute care utilization [2–5]. For these reasons, there is tremendous interest in analytical tools capable of predicting resource intensive healthcare (RIHC) [2,6].

Prior modelling work relies on data from individual patients to predict future super-utilization of RIHC [3,7–10]. An important challenge with this approach is the episodic rather than persistent nature of super-utilization, such that a patient’s risk in one year may be high but return to average risk the following year [5,6] One potential alternative is a population-centric approach where data from population-units (e.g. state, county) are used to predict RIHC. Predicting RIHC among population-units, rather than individuals, offers several key advantages. First, it leverages established associations between population characteristics and RIHC [5,9–11]. Moreover, population characteristics of super-utilizers tend to be stable overtime, even if the individuals are not [5]. Finally, this approach may be less sensitive to variation in utilization observed among models relying on individual-level data. While this approach represents a departure from existing predictive modelling work, similar population-centric approaches are used to study geographic variation in utilization and spending [12–16]. Adopting a population-centric approach to predicting RIHC may help identify populations consuming the greatest share of RIHC. Moreover, aggregate utilization trends can help clinicians and health systems better understand local practice patterns, which may help inform the delivery of healthcare. Finally, such models may identify population characteristics associated with RIHC, which can provide modifiable targets for interventions.

Generating prediction models among population-units also allows the opportunity to leverage routinely collected governmental data. The U.S. government collects vast and diverse information on population units, and when used for research, this data can support robust modelling. Moreover, this data provides models with traditional risk factors of RIHC (e.g., demographics, disease/health status) and can also provide other data sources not typically used to study healthcare utilization. For example, consumer expenditures are data on buying habits, which may provide information on risk for RIHC because individuals in economically developed societies are said to consume goods and services more than any other activity [17–19]. In addition, consumer expenditures may serve as proxies for unobservable or difficult to measure variables or they may reflect goods or behaviors associated with health or healthcare utilization.

The objective of this study was to leverage the power of machine learning with diverse governmental data, including consumer expenditures, to predict RIHC among U.S. counties. Predicting RIHC among U.S. counties provides a novel approach for understanding the delivery of RIHC and models predicting RIHC rather than super-utilization status may provide additional insight across the full distribution of utilization.

2. Methods

This study predicted RIHC and associated spending among eligible U. S. counties using cross-sectional data and an ecological study design. Counties were eligible if they had a hospital or ER facility in 2017. Data on eligible counties came from 5 sources: (1) American Hospital Association (AHA), (2) the U.S. Census Bureau (USCB), (3) the Centers for Disease Control and Prevention (CDC), (4) the American Community Survey (ACS), and (5) the Bureau of Labor Statistics (BLS). AHA data were obtained from publicly available files from the Health Resources Service Administration (HRSA) while the remaining data were accessed from Data Planet©, a web-based research tool supported by SAGE publishing and licensed through Dartmouth College [20]. Data Planet© uses software engineering and statistical modelling to aggregate licensed and public domain data, providing a convenient tool for extracting data from multiple sources [20]. Fig. 1 illustrates key data components and model development processes utilized in this study. All data used in this study were from 2017, which provided the most current data across all sources at the onset of the project in 2019. All analytical work was performed in R version 3.6.0 (R Foundation). This studied was exempt from the Dartmouth College institutional review board and adhered to STROBE reporting guidelines [22]. Information on data availability is in appendix A.

2.1. Outcome

Three outcome metrics defined RIHC and represent utilization or associated spending from high resource healthcare services. All 3 outcome metrics were extracted from the AHA data obtained from the publicly available 2017 HRSA file and included: (1) total number of *ER visits*, defined as the number of emergency department visits at short term general, short term non-general or long-term hospitals, (2) total number of *inpatient days* defined as the number of adult and pediatric days of care occurring at any hospital type, excluding newborn days or cases, and (3) *hospital expenditures,* defined as total hospital expenditures from short term general, short term non-general, long-term, and Veterans Affairs (VA) hospitals [23]. Due to positive skew, all 3 outcome metrics were log transformed for analyses and expressed as per capita to account for population size of each county [24]. Shapefiles from USCB were used in Tableau to illustrate geographic variation in each per capita outcome metric [25].

2.2. Target-prediction features from U.S. Counties

Four target-prediction feature groups were used: (1) demographics from USCB, (2) adult and child health characteristics from CDC, (3) community characteristics from ACS, and (4) consumer expenditures from BLS. All target-prediction features were extracted from Data Planet© [20]. Demographics were based on 2010 Census, projected to 2017, and included information on age, race, income, and education [20,26]. Adult and child health characteristics were based on the CDC's vital and health statistics along with national health interview survey and included information on physical and mental health status along with access to healthcare [20,27–29]. Community factors were based on the 2017 ACS and included information on employment and housing [20,30,31]. Consumer expenditure data were based on the nationwide Consumer Expenditure Survey (CEX) and provided information on household expenditures for foods, home goods and miscellaneous items [20,27,32,33]. Data Planet© utilizes statistical modelling tools along with demographic information to project estimates from all target-prediction feature groups to U.S. counties in 2017 [34]. All target-prediction features, except for associated county and state name were continuous.

2.3. Pipeline steps: Preprocessing, interactions, feature selection

Variables from the 4 target-prediction feature groups were processed through a data pipeline to clean, transform, and harmonize all variables. To account for population size, variables were expressed per capita (expenditures, income, or population density) or per 100 persons. Moreover, variables were expressed per capita to reflect the appropriate per capita denominator based on specific age-groupings. For example: (1) adults were expressed as 18 years and up, (2) children were expressed as less and 18 years, (3) school aged children were expressed as 3–17 years, and (4) working age population were expressed as 16 years and up. Data were merged across target-prediction feature groups and then to each outcome metric separately, using county and state name.

Data were split into train and test-sets using an 80/20 split, respectively. The train-set was used to identify interactions, perform feature selection, and train the models. The test-set was left-out until model evaluation. Second-order terms were identified using the iml package, which uses Random Forest and the H-statistic to identify and create pair-wise interaction terms [35]. The H-statistic measures variation in the predicted outcome based on the presence of interacted features [35,36]. Interaction terms in the 90th percentile for their corresponding H-statistic were retained. Two separate feature selection techniques were employed, including Least Absolute Shrinkage Operator (LASSO) and Random Forest [37,38]. Retained interaction terms were used when implementing LASSO for feature selection. The LASSO algorithm selects important features and shrinks all other coefficient features to zero [39]. Random Forest can inherently account for interactions within its model [37,39]. With random forest, features in the 90th percentile of feature importance were retained.

2.4. Machine learning model development

Variables identified in each feature selection technique were used in the construction of models using 4 distinct machine learning algorithms, which included linear regression, l1 regularized linear regression, random forest regression, and gradient boosting regression. A systematic approach was undertaken, where the same 4 machine learning algorithms were used with different data groupings (i.e. target-prediction features) to generate distinct models for the 3 RIHC outcome metrics (Fig. 2) [40]. This method was undertaken to (1) identify the best performing model for each outcome metric based on a diverse set of possible data groupings using 4 distinct algorithms and (2) evaluate the information contained in the data groupings. A machine learning pipeline was generated using the sl3 package, which allowed parallel execution and evaluation of the 4 machine learning algorithms on unique combinations of data groupings [41]. The number of unique combinations of data groupings differed across 4 iterations, ultimately generating 4*N models, where N is the number of unique combinations of data groupings (Fig. 2).

Models were developed with 10-fold cross validation using training data [39]. Default hyperparameters were used. Details on model specification are available in appendix B. Mean squared error (MSE), calculated on the test-set using cross-validation, was the metric used to compare models and identify the best performing model. In general, MSE is considered an appropriate metric for evaluating regression models and is widely used in numerous research fields to compare prediction model performance [39,42–44]. A 3-step approach was used to identify the best overall model for each outcome metric (Fig. 2). First, among the 4 algorithms run for each data grouping (within each iteration), the model with the lowest MSE was retained. This process was repeated for each iteration, such that each data grouping, in each iteration, had a best-performing model. Next, models with the lowest MSE across data groupings but within each iteration were retained. Finally, models were compared across 4 iterations, and the model with the lowest MSE was considered the best performing model for the assigned feature selection technique. This process was completed separately for models using variables selected from LASSO and variables retained from random forest feature selection techniques. Finally, the model with the lowest MSE across feature selection techniques was considered the best performing prediction model overall. This process was repeated for each outcome metric, yielding 3 best performing prediction models.

Once the best performing model for each outcome metric was identified, Shapley Additive exPlanations (SHAP) analysis was performed [45,46]. SHAP analysis is a method for improving prediction model interpretability, and its applications are well described in the literature [46,47] SHAP analysis deconstructs each predicted value into the sum of contributions made from each input feature [45–47]. As such, each predicted value (e.g., predicted ER visits, IP days or hospital expenditures) is the sum of the SHAP values plus a base value (for regression models, the base value is the mean of the outcome) [45,47]. The SHAP values illustrate how model inputs (e.g., features/variables) influence the model outputs (e.g., predicted values), providing the ability to comment on both local and global outputs from a given model [46,47]. For this analysis, SHAP summary plots were created for the best performing model, for each outcome metric. The SHAP summary plot illustrates the relationship between the SHAP value for a particular feature and the predicted value [47]. For each feature, a corresponding SHAP value is plotted on the x-axis. The point's color reflects the size of the corresponding feature value (high = red, low = blue) [47]. This allows the opportunity to comment on the impact of the feature value on the predicted value [47]. For example, counties with higher income per capita have higher SHAP values for inpatient day and hospital expenditure outcome metrics, which means they have higher predicted values for these models. In addition, the distribution of SHAP values also provides useful information for model interpretation. For example, counties with very high income per capita appear to have a strong positive impact on SHAP values, corresponding to higher predicted values for inpatient day and hospital expenditure outcome metrics. To conduct SHAP analysis, the reticulate package in R was utilized to interface with the shap package in Python.

3. Results

There were 2475 counties with ER services and 2491 with hospitals services, representing 78.8% and 79.3% of all U.S. counties, respectively. Median per capita values were as follows: 0.450 ER visits [IQR:0.318, 0.618], 0.368 inpatient days [IQR: 0.176, 0.826], and $2104 hospital expenditures [IQR: $1299.93, 3362.97]. Across U.S. counties, geographic variation in all 3 per capita outcome metrics was identified (Fig. 3). General information on target-prediction features are found in Table 1. All univariate descriptive statistics are available in appendix tables C–F. (main effects) and appendix C.1–F.1 (second-order terms). For brevity, univariate statistics for select variables are described here. The median age per county was 41 years, the per capita median income per county was approximately $59,000, and on average, 81% of county residents were white. Approximately 10% of adults per county had diabetes, 15.4% were current smokers, and about 12% had heart disease. Among children less than 18 years per county, almost 27% missed no school in the preceding year and 7.3% had a documented learning disability. Among employed adults per county, almost half used their own vehicle to commute and almost 20% reported a commute time of less than 15 min. Per capita annual expenditures for food consumed in the home, housekeeping supplies, and power tools per county were $2,900, $265, and $14, respectively.

Best performing prediction models, for each outcome metric, used random forest as the feature selection technique, included variables from all 4 target-prediction feature groups, and used a non-parametric machine learning algorithm (Table 2). Models for each outcome metric included approximately 100 candidate predictors identified from random forest feature selection. Observed vs. expected plots for log hospital expenditures per capita (referred to 'hospital expenditures') illustrated good fit across the range of predicted probabilities, however, similar plots for log ER visits per capita (referred to as 'ER visits') and log inpatient days per capita (referred to as 'inpatient day') illustrated modest fit at the extremes of utilization (Fig. 4). The top 5% of predicted counties for each outcome metric were overwhelmingly located in Southern or Midwestern areas of the U.S (appendix Figure A).

Variable importance or relative influence values from the best performing prediction models are available in appendix G–I. The top 20 most important features were predominantly demographic and community characteristics. Consumer expenditures were among the top 20 most important features for the inpatient day and hospital expenditure outcome metrics. Fig. 5 A–C presents SHAP summary plots from the best performing models. Each sub-plot of Fig. 5 ranks the absolute value of the SHAP value for the top 20 most important features for each outcome metric. For each feature, the SHAP value is plotted on the x-axis and its color reflects the actual value of the associated feature (red = high, blue = low) [47]. Together, this illustrates how features impact associated prediction values [47]. For example, Fig. 5 B & C show that higher values for the per capita number of people employed in healthcare and/or social assistance fields have higher SHAP values, corresponding to higher predicted values, for inpatient day and hospital expenditure outcome metrics. Moreover, the distribution of this feature suggests counties with large per capita healthcare and/or social assistance employees have strong positive impacts on predicted values for both outcome metrics.

4. Discussion

Using diverse governmental data extracted, transformed, and merged across multiple sources, this study implemented a machine learning pipeline to generate prediction models for 3 measures of annual RIHC among U.S. counties in 2017. Results have applications across healthcare delivery and research, along with population health. First, models predicted RIHC among U.S. counties with performance comparable to existing models [48–50]. This provides a tool for clinicians, health systems, or local agencies to monitor trends in RIHC, offering a population-centric approach for identifying populations utilizing a disproportionate amount of RIHC. Moreover, these results also showcase the utility of using routinely collected government data to study healthcare utilization. Second, demographic and community characteristics were among the most important predictors for all 3 RIHC outcome metrics. Many community characteristics are modifiable, which can help inform population health interventions, while the demographic characteristics can help identify groups with a higher risk for RIHC. Third, several consumer expenditures contained predictive value for inpatient day and hospital expenditure outcome metrics. This suggests consumer expenditures offer some value for predicting RIHC, lending some support for their use in future research.

Using populations as the unit of analysis may help address limitations observed with existing models, as RIHC outcomes, when aggregated, may be less sensitive to variation in

RIHC attributed to individuals. This approach leverages known and consistent associations between population factors and RIHC, captured from routinely collected governmental data. Rather than targeting individual super-utilizers, this tool can identify population-units (e.g., counties) using a disproportionate amount of RIHC, highlighting counties in need of potential intervention, and providing a general monitoring tool. Across all 3-outcome metrics, most of the top 5% of predicted counties were concentrated among Southern and Midwestern states including Virginia, Louisiana, Kansas, the Dakotas, and North Carolina. These results are consistent for Southern states, which historically had the greatest annual Medicare per capita spending, when compared to other regions in the U. S [51,52]. Midwestern state Medicare spending per capita has historically been more heterogenous across states, however, Plains states experienced higher than average healthcare spending per capita, consistent with findings from our study [51,53]. Lastly, when compared to previous research conducted among states and Census regions, the county-level estimates from this study offer more granular information on the utilization and spending of RIHC [53]. Clinicians and health systems may find these utilization metrics valuable in better understanding their own local practice patterns. County-level estimates may offer a more suitable benchmark for clinicians and health systems to use as comparators to their own utilization or spending trends, as higher levels of aggregation (e.g., state or Census regions) may mask local heterogeneity.

Consistent with previous research, demographics and community characteristics were identified as important predictors for all 3-outcome metrics. Specifically, employee travel time between 30 and 59 min and employment in healthcare or social assistance fields were two variables identified as important predictors across all 3-outcome metrics. Counties where the round-trip travel time for employees was between 30 and 59 min corresponded to lower predicted outcome values for all 3-outcome metrics. A total travel time of 30–59 min is at or below the national average of approximately 60 min [54]. While employee travel time and utilization of RIHC remains largely unstudied, prior research reports adverse associations between employee commute time and physical activity, adiposity, sleep disturbances, and metabolic risk factors [55–58]. As such, a shorter commute time may affect risk for RIHC by modifying upstream risk factors for utilization of RIHC, including health behaviors (e.g., physical activity, diet) or outcomes (e.g., disease or disease exacerbations) [3,10]. These results contribute to a growing body of literature suggesting reduced employee travel time may achieve health benefits [58]. In addition, counties with high concentrations of per capita healthcare or social assistance employees corresponded to higher predictions for all 3-outcome metrics. One explanation is that it reflects areas with greater supplies of healthcare providers, which has shown to affect geographic variation in utilization and spending [16,59]. This represents an important area for future research, especially since healthcare or social assistance employees was not an important predictor for ER visits. Future research should consider if associations between provider supply and variation in utilization are modified by the type of healthcare. Importantly, associations identified in this study were statistically driven and do not represent causal findings. However, results from this study do demonstrate the importance of county-level characteristics in predicting RIHC.

The contribution of consumer expenditures to this study suggests some utility in predicting RIHC. Consumer expenditures were seen as important predictors for the inpatient day and hospital expenditure outcome metrics. Consumer expenditures may be a proxy for income, which could explain higher inpatient days and hospital expenditures as income in the U.S. is often positively associated with healthcare utilization, including RIHC [11]. While this study cannot address causality between consumer expenditures and RIHC, findings suggest county-level consumer expenditures may help predict hospital utilization or spending. Future studies are needed to better understand the associations between consumer expenditures and RIHC.

5. Limitations

While this work offers novel insights into the power of leveraging diverse governmental data to predict RIHC, the work is not without limitations. First, data are cross-sectional and cannot comment on longitudinal trends. Sensitivity analysis found small variation in the outcome metrics over a 3-year period (2017–2019), even among top county-utilizers, suggesting the cross-sectional estimates presented in this study may be robust to periodicity (appendix J, K, L). Second, results from this study are vulnerable to external changes such as large-scale events (e.g., COVID-19 pandemic) or policy changes (e.g., healthcare expansion). Third, study findings cannot address individual-level factors associated with RIHC. Fourth, model results may be biased based on the underlying data used in their construction, however, these data are routinely collected from reputable agencies and are considered valid measures for informing products (e.g., consumer price index) or designations (e.g., county health rankings). Lastly, U.S. counties without ER services or hospitals were excluded from this analysis (N~700). People living in counties without ER services or hospitals will likely use facilities in a neighboring county, thereby contributing to the utilization outcome without also contributing their counties exposure data to the possible target-prediction features. To explore the magnitude of this limitation, future research will explore the overlap between counties and Hospital Service Areas, which are units defining markets for hospital care.

6. Conclusion

Using diverse governmental data and a systematic machine learning pipeline, models predicted 3 RIHC outcome metrics among U.S. counties. This research provides a novel method and framework for predicting RIHC that capitalizes on diverse and routinely collected data. Trends in RIHC were identified at the county-level, offering clinicians and health systems information on local practice patterns and local agencies with potential target-areas for future interventions. Consistent with prior literature, community characteristics were important predictors of RIHC.

7. Funding support

This work was supported by the National Library of Medicine of the National Institutes of Health T32 training grant (T32 LM012204); and the National Heart, Lung, and Blood Institute (NHLBI R01HL130828).

National Library of Medicine of the National Institutes of Health and the National Heart, Lung, and Blood Institute had no roles in the design, analysis or writing of this article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

[1]. Centers for Medicare & Medicaid Services. Health Spending by Type of Service or Product. In: National Health Expenditures 2017 Highlights. Centers for Medicare & Medicaid Services; 2018. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/downloads/highlights.pdf. [Accessed 1 June 2020].

[2]. Iovan S, Lantz PM, Allan K, Abir M. Interventions to decrease use in prehospital and emergency care settings among super-utilizers in the United States: a systematic review. Med Care Res Rev 2019:1077558719845722.

[3]. Ng SH, Rahman N, Ang IYH, et al. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. BMC Health Serv Res 2019;19(1):452. [PubMed: 31277649]

[4]. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed Eng Online 2018;17(Suppl 1):131. [PubMed: 30458798]

[5]. Johnson TL, Rinehart DJ, Durfee J, et al. For many patients who use large amounts of health care services, the need is intense yet temporary. Health Aff 2015;34(8): 1312–9.

[6]. Finkelstein A, Zhou A, Taubman S, Doyle J. Health care hotspotting - a randomized, controlled trial. N Engl J Med 2020;382(2):152–62. [PubMed: 31914242]

[7]. Andriotti T, Dalton MK, Jarman MP, et al. Super-utilization of the emergency department in a universally insured population. Mil Med 2021;186(7–8):e819–25. [PubMed: 33247301]

[8]. Ziring J, Gogia S, Newton-Dame R, Singer J, Chokshi DA. An all-payer risk model for super-utilization in a large safety net system. J Gen Intern Med 2018;33(5): 596–8. [PubMed: 29464478]

[9]. Fitzpatrick T, Rosella LC, Calzavara A, et al. Looking beyond income and education: socioeconomic status gradients among future high-cost users of health care. Am J Prev Med 2015;49(2):161–71. [PubMed: 25960393]

[10]. Alberga A, Holder L, Kornas K, Bornbaum C, Rosella L. Effects of behavioural risk factors on high-cost users of healthcare: a population-based study. Can J Public Health 2018;109(4):441–50. [PubMed: 30232715]

[11]. Wammes JJG, van der Wees PJ, Tanke MAC, Westert GP, Jeurissen PPT. Systematic review of high-cost patients' characteristics and healthcare utilisation. BMJ Open 2018;8(9):e023113.

[12]. Newhouse JP, Garber AM. Geographic variation in Medicare services. N Engl J Med 2013;368:1465–8. [PubMed: 23520983]

[13]. Welch HG, Sharp SM, Gottlieb DJ, Skinner JS, Wennberg JE. Geographic variation in diagnosis frequency and risk of death among Medicare beneficiaries. JAMA 2011;305(11):1113–8. [PubMed: 21406648]

[14]. Yongkang JZL. Geographic variation in Medicare per capita spending narrowed from 2007 to 2017. Health Aff 2020;39(11).

[15]. Zhang Y, Baik SH, Fendrick AM, Baicker K. Comparing local and regional variation in health care spending. N Engl J Med 2012;367(18):1724–31. [PubMed: 23113483]

[16]. Zhang Y, Li J, Yu J, Braun RT, Casalino LP. Social determinants of health and geographic variation in Medicare per beneficiary spending. JAMA Netw Open 2021;4(6):e2113212. [PubMed: 34110394]

[17]. Ricket IM, Khayal I, Brown JR. Consumer data and risk stratification for conronary heart disease Northeast Regional IDeA. In: Conference; 2019.

[18]. Mothersbaugh D, Hawkins D. Consumer behavior: building marketing strategy. New York, NY: McGraw-Hill Education; 2015.

[19]. Rani P Factors influencing consumer behaviour. Int J Curr Res Aca Rev 2014;2(9): 52–61.

[20]. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. Accessed (15 July 2020).

[22]. Cruschieri S The STROBE guidelines. Saudi J Anaesth 2019;13:S31–4. [PubMed: 30930717]

[23]. Health Resources Services Administration. Technical Documentation. In: Area Health Resources Files County-Level Data 2017–2018. Health Resources Services Administration; 2018. https://data.hrsa.gov/data/download. [Accessed 1 July 2020].

[24]. Mihaylova B, Briggs A, O'Hagan A, Thompson SG. Review of statistical methods for analysing healthcare resources and costs. Health Econ 2011;20(8):897–916. [PubMed: 20799344]

[25]. United States Census Bureau. Cartographic boundary files-shapefile. United States Census Bureau; 2017. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html. [Accessed 1 March 2021].

[26]. EASI Market Planner-Demographics. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[27]. EASI Market Planner-Consumer Food Expenditures. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[28]. EASI Market Planner-Health-Adults. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[29]. EASI Market Planner-Health-Children. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[30]. EASI Market Planner-Housing Units. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[31]. EASI Market Planner-Employment. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[32]. EASI Market Planner-Consumer Miscellaneous Expenditures. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[33]. EASI Market Planner-Consumer Home Expenditures. Data Planet™. New York: SAGE Publishing; 2017. https://dataplanet.sagepub.com. [Accessed 10 April 2020].

[34]. Data Planet™. Data Planet statistical datatsets: About Data Planet statistical datasets. New York: SAGE Publishing; 2017. https://data-planet.libguides.com/c.php?g=398594&p=2710212. [Accessed 1 March 2021].

[35]. iml: An R package for interpretable machine learning, 3.26. J Open Source Software; 2018. p. 786.

[36]. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat 2008;2(3).

[37]. Liaw AWM. Classifcation and regression by randomForest. R News 2002;2/3: 18–22.

[38]. Hastie T, Qian J, Tay K. An Introduction to glmnet. In: R CRAN; 2020. https://cran.r-project.org/web/packages/iml/iml.pdf. [Accessed 15 November 2021].

[39]. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York, NY: Springer; 2013.

[40]. Xin DML, Song S, Parameswaran A. How developers iterate on machine learning workflows. arXiv; 2018.

[41]. Coyle JR, Hejazi NS, Malenica I, Phillips RV, Sofrygin O. Modern pipelines for machine learning (Super Learning). In: R CRAN; 2021. https://github.com/tlverse/sl3. [Accessed 15 November 2021].

[42]. Chicco D, Warrens MJ, Jurman G.The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 2021;7:e623.

[43]. Shi HY, Lee HH, Tsai JT, et al. Comparisons of prediction models of quality of life after laparoscopic cholecystectomy: a longitudinal prospective study. PLoS One 2012;7(12):e51285. [PubMed: 23284677]

[44]. Acharya MS, Armaan A, Antony AS. A comparison of regression models for prediction of graduate admissions. Second International Conference on Computational Intelligence in Data Science; 2019.

[45]. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 31st Conference on neural information processing systems. 2017.

[46]. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explainations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67. [PubMed: 32607472]

[47]. Molnar C Interpretable machine learning. A Guide for making black box models explainable. Munchen, Germany. Mucbook Clubhouse; 2019.

[48]. Caballer-Tarazona V, Guadalajara-Olmeda N, Vivas-Consuelo D. Predicting healthcare expenditure by multimorbidity groups. Health Pol 2019;123(4):427–34.

[49]. Kuo RN, Lai MS. Comparison of Rx-defined morbidity groups and diagnosis-based risk adjusters for predicting healthcare costs in Taiwan. BMC Health Serv Res 2010; 10(26).

[50]. Huang XPS, Lavergne R, Ahuja M, McGraul K. Predicting the cost of health care services: a comparison of case-mix systems and comorbidity indices that use administrative data. Med Care 2020:58.

[51]. Skinner JCauses and consequences of regional variations in health care, 2. Handbook of Health Economics; 2011. p. 45–93.

[52]. Institute of Medicine (US) Committee on Guidance for Designing a National Healthcare Disparities Report. In: Swift EK, editor. Guidance for the National Healthcare Disparities Report. Washington (DC): National Academies Press (US); 2002.

[53]. Lassman D, Sisko AM, Catlin A, et al. Health spending by state 1991–2014: measuring per capita spending by payers and programs. Health Aff 2017;36(7): 1318–27.

[54]. Burd C, Burrows M, McKenzie B. Travel time to work in the United States: 2019. In: American Community Survey Reports. United States Cenesus Bureau; 2021. https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs-47.pdf. [Accessed 1 June 2021].

[55]. Urhonen T, Lie A, Aamodt G. Associations between long commutes and subjective health complaints among railway workers in Norway. Prev Med Rep 2016;4:490–5. [PubMed: 27660744]

[56]. Kunn-Nelen A Does commuting affect health? Health Econ 2016;25(8):984–1004. [PubMed: 26010157]

[57]. Hoehner CM, Barlow CE, Allen P, Schootman M. Commuting distance, cardiorespiratory fitness, and metabolic risk. Am J Prev Med 2012;42(6):571–8. [PubMed: 22608372]

[58]. Raza A, Pulakka A, Magnusson Hanson LL, Westerlund H, Halonen JI. Commuting distance and behavior-related health: a longitudinal study. Prev Med 2021;150: 106665. [PubMed: 34081935]

[59]. Ricketts TC, Belsky DW. Medicare costs and surgeon supply in hospital service areas. Ann Surg 2012;255(3):474–7. [PubMed: 21975316]

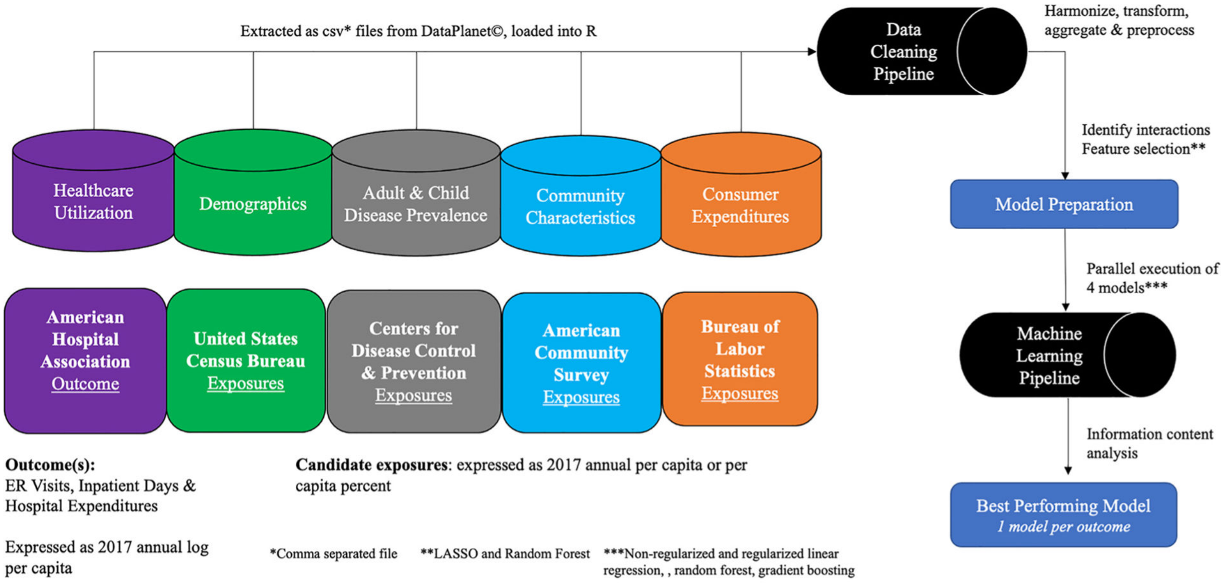
Extracted as csv* files from DataPlanet©, loaded into R
Data Cleaning Pipeline
Harmonize, transform, aggregate & preprocess
Healthcare Utilization
Demographics
Adult & Child Disease Prevalence
Community Characteristics
Consumer Expenditures
Identify interactions
Feature selection**
Model Preparation
Parallel execution of 4 models***
American Hospital Association Outcome
United States Census Bureau Exposures
Centers for Disease Control & Prevention Exposures
American Community Survey Exposures
Bureau of Labor Statistics Exposures
Machine Learning Pipeline
Information content analysis
Outcome(s):
ER Visits, Inpatient Days & Hospital Expenditures
Expressed as 2017 annual log per capita
Candidate exposures: expressed as 2017 annual per capita or per capita percent
Best Performing Model
1 model per outcome
*Comma separated file
**LASSO and Random Forest
***Non-regularized and regularized linear regression, , random forest, gradient boosting

Fig. 1. Data & Study Overview.

Data from 5 governmental sources were extracted and cleaned in a pipeline before undergoing model preparation and model development using a machine learning pipeline.

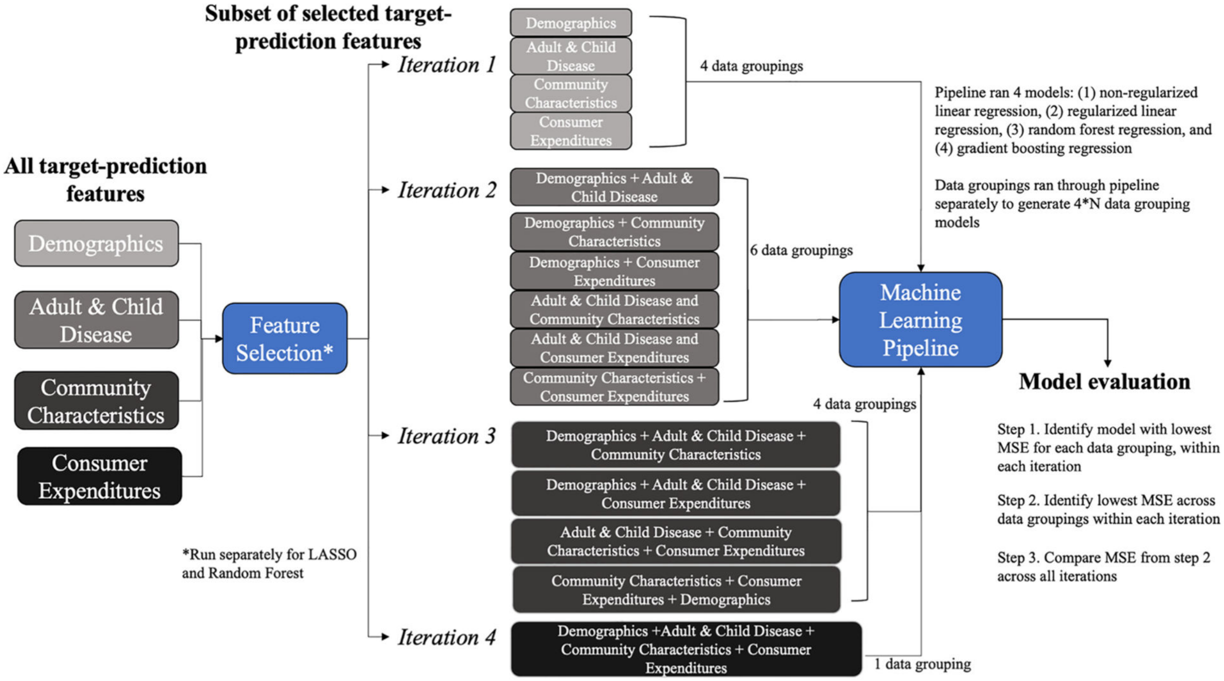
Subset of selected target-prediction features
All target-prediction features
Demographics
Adult & Child Disease
Community Characteristics
Consumer Expenditures
Feature Selection*
*Run separately for LASSO and Random Forest
Iteration 1
Demographics
Adult & Child Disease
Community Characteristics
Consumer Expenditures
4 data groupings
Iteration 2
Demographics + Adult & Child Disease
Demographics + Community Characteristics
Demographics + Consumer Expenditures
Adult & Child Disease and Community Characteristics
Adult & Child Disease and Consumer Expenditures
Community Characteristics + Consumer Expenditures
6 data groupings
Iteration 3
Demographics + Adult & Child Disease + Community Characteristics
Demographics + Adult & Child Disease + Consumer Expenditures
Adult & Child Disease + Community Characteristics + Consumer Expenditures
Community Characteristics + Consumer Expenditures + Demographics
4 data groupings
Iteration 4
Demographics +Adult & Child Disease + Community Characteristics + Consumer Expenditures
1 data grouping
Machine Learning Pipeline
Pipeline ran 4 models: (1) non-regularized linear regression, (2) regularized linear regression, (3) random forest regression, and (4) gradient boosting regression
Data groupings ran through pipeline separately to generate 4*N data grouping models
Model evaluation
Step 1. Identify model with lowest MSE for each data grouping, within each iteration
Step 2. Identify lowest MSE across data groupings within each iteration
Step 3. Compare MSE from step 2 across all iterations

Fig. 2. Systematic machine learning approach.

Illustrates information content assessment of target-prediction features from 4 unique data sources in predicting 3-outcome metrics of resource intensive healthcare. Four distinct machine learning algorithms were used with unique combinations of data groupings (i.e., target-prediction features) across 4 iterations. This process was repeated for the 3-outcome metrics, generating a total of 90 unique data groupings and 360 distinct models across both feature selection techniques for all 3-outcome metrics.

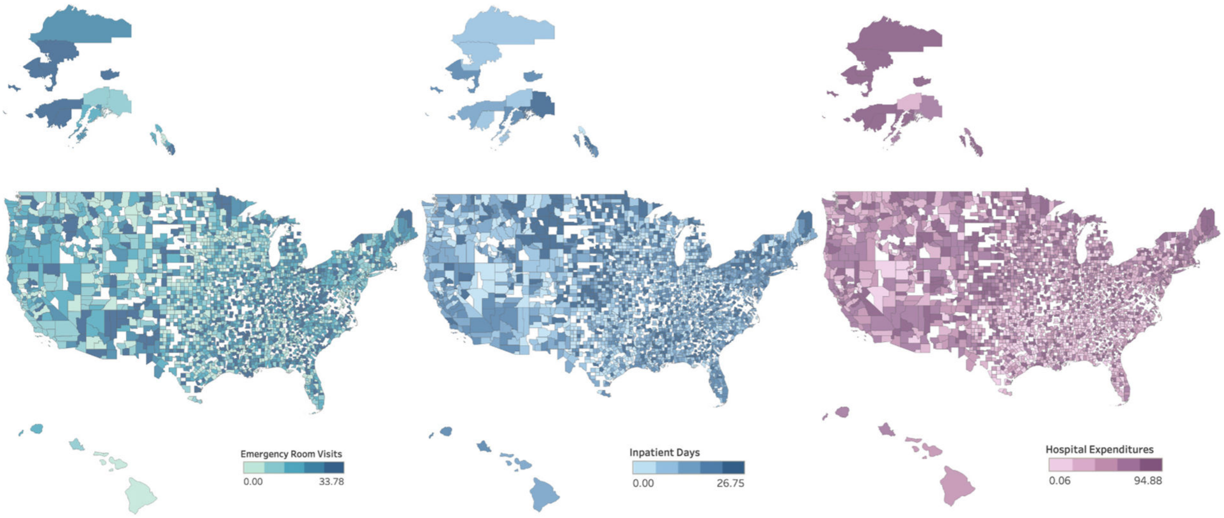

Fig. 3. Per Capita Emergency Room Visits, Inpatient Days, and Hospital Expenditures among U.S. Counties in 2017.

Heat map of annual per capita emergency room visits, inpatient days, and hospital expenditures from U.S. counties in 2017, broken into quintiles.

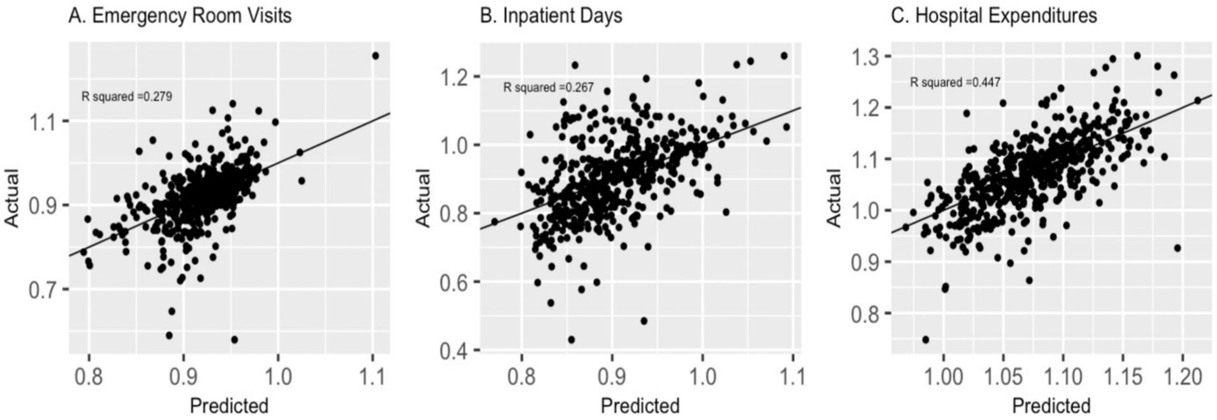

Fig. 4. Observed v. Expected Plots for Best Performing Machine Learning Prediction Models for Emergency Room Visits, Inpatient Days, and Hospital Expenditures among U.S. Counties in 2017.

A.Log Emergency Rooms visits per capita |R^2 0.279 | MSE: 0.003 B. Log Inpatient Days per capita |R^2 0.267 | MSE: 0.009 C. Log Hospital Expenditures per capita |R^2 0.447 | MSE: 0.003.

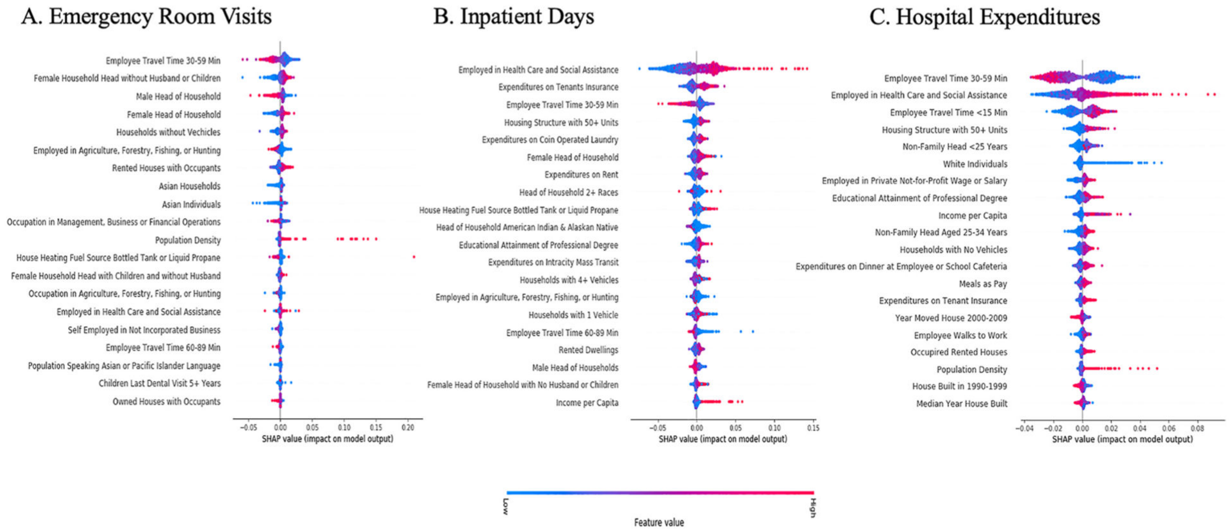

Fig. 5. SHAP Value Analysis from for the top 20 features from Best Performing Machine Learning Prediction Models for Emergency Room Visits (A), Inpatient Days(B), and Hospital Expenditures (C) among U.S. Counties in 2017.

Each county in the test dataset appears as its own point for all illustrated variables, where the point represents the absolute value of the associated SHAP value. The color corresponds to the raw values of the associated feature, for each point on the graph (high = red, low = blue).

Table 1

Type and description of available target-prediction feature groups and variables for machine learning prediction models.

Target-prediction feature group	Total variables available for prediction models[a]	Data source[b]	Data description	Variable examples
Demographics	149 main effects 15 second order terms	United States Census Bureau	Per capita information on race, gender, family size, education, income etc.	Per capita percent of:(1) adults 25 years+ with college degree (2) Black household heads
Adult & Child Health Characteristics	146 main effects 15 second order terms	Centers for Disease Control & Prevention	Per capita prevalence of select diseases, health behaviors etc.	Per capita percent of: (1) children (less than 18) with asthma (2) adults (18 years+) with overweight BMI
Community Characteristics	151 main effects 15 second order terms	American Community Survey	Per capita employment information, housing characteristics etc.	Per capita percent of:(1) travel time between 30 and 59 min (2) median age and size of home
Consumer Expenditures	571 main effects 57 second order terms	Bureau of Labor Statistics	Per capita expenditures on food, household goods, and miscellaneous items	Per capita expenditures on: (1) beef or red meat (2) infant snowsuits or jackets

[a] A total of 1119 target-prediction features from 1 of 4 groups was made available to the machine learning prediction model pipeline. Main effects and second order terms were used in parametric prediction models while second order terms were excluded from nonparametric prediction models. Variables from each of the 4 groups were available for all eligible U.S. counties.

[b] Data was extracted from DataPlanet©, which aggregates public domain and licensed data. The original governmental data source is listed in the table.

Table 2

Best performing models for 2017 emergency room visits, inpatient days, and hospital expenditures among U.S. Counties.

Outcome (log per capita)	ER Visits (N = 2475)	Inpatient Days (N = 2491)	Hospital Expenditures (N = 2491)
Target-prediction feature groups included [a]	4	4	4
Feature Selection	Random Forest	Random Forest	Random Forest
Model Type	Random Forest	Gradient Boosting	Random Forest
MSE [b]	0.003	0.009	0.003
R2 [c]	0.279	0.267	0.447

[a] Target-prediction feature groups: Demographics, Adult & Child Health Characteristics, Community Characteristics, and Consumer Expenditure Variables.

[b] MSE = mean squared error, calculated on test-set.

[c] Coefficient of determination, calculated on test-set.