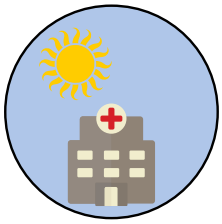

HCO THRIVE

Local-Area Healthcare Utilization Projections (LAHUP)

Technical Documentation: Hospital Service Lines

Date: April 2, 2025

Version: 0.1

Contents

I. Overview

The purpose of this document is to provide an in-depth description of HCOThrive's Local Area Utilization Projections (LAHUP) for Hospital Service Lines (including inpatient admissions and ER visits). The document outlines the overall process along with key considerations and decisions in producing the projections.

Document Updates

Note that our projections are updated on a regular basis along with additional information as well as new projections not currently stocked. It is our mission to provide the most accurate projections possible using proven, state-of-the-art methodology. Thus, this technical documentation is updated regularly to reflect any and all changes so please check for the latest version.

Document Structure

The rest of this technical documentation is structured as follows: *Section II* describes the data sources that serve as input to our models. *Section III* details the modeling and estimation process as well as the resultant output. Finally, *Section IV* covers assumptions and limitations.

II. Input

This section provides a list of data sources used in the modeling and estimation process as well as a brief description of each source and rationale behind its use. Readers interested in greater detail regarding data inputs are encouraged to visit the data providers website or review the source material for themselves.

Data Sources

Model Data: Medical Expenditure Panel Survey (MEPS)

LAHUP uses the Medical Expenditure Panel Survey (MEPS) to derive statistical models for provider visits based on demographics and the presence of chronic health conditions. Broadly, MEPS contains information on,

- Healthcare utilization and expenditures
- Demographic characteristics
- Insurance status
- Chronic health conditions
- Perceived health status
- A variety of other individual-level measures

In addition, MEPS also contains detailed individual health conditions, visit characteristics, and employment data as well. There is also a provider survey linked to a a subset of the MEPS-HC.

Geographic Population Data (Adult): Behavioral Risk Factor Surveillance Survey (BRFSS)

Behavioral Risk Factor Surveillance Survey (BRFSS) data was used to model adult utilization at the state-level using the model derived from MEPS. The BRFSS contains a variety of both demographic and chronic health condition variables similar to MEPS collected from the adult (18+) population in each U.S. state. For LAHUP, BRFSS is used to estimate total predicted utilization for adults within each state. As described in the Section III., the BRFSS is useful in this regard as it contains the full, joint distribution of demographics and chronic conditions unlike other data sources, such as the ACS.

Beyond the health-related factors listed above, the BRFSS also contains demographic questions regarding sex, age, race/ethnicity, martial status, education, and income among others. And, as noted previously, BRFSS contains state-level geographic information along with limited information on urban/rural status and/or MSA status.

Geographic Population Data (Child): American Community Survey (ACS)

The American Community Survey (ACS) was used to model child (0-17) utilization at the state-level as the BRFSS is only collected from adults. The ACS represents a 1% sample of the U.S. and Puerto Rico population (approximately 3.54 million people surveyed each year). Although small-area (e.g, county, census tract, block group) data is available only in summary format, individual-level results are available to the public aggregated at the Public Use Microdata Area (PUMA) which is an area of 100,000+ people or more. In essence, the ACS represents a replacement for the long-form decennial census and contains detailed information on,

- Household characteristics
- Demographics
- Geography
- Employment
- Health insurance status
- Disability status

While ACS-PUMS lacks health information found in MEPS and BRFSS such as chronic conditions, it does contain similar demographic variables.

Population Projections

Population projections for all U.S. counties were derived using an adaptation of the methodology described in Hauer (2019). Briefly, projections were produced with an autoregressive integrated moving average (ARIMA) using cohort methodology from the demographic forecasting literature. Specifically, in order skirt data limitation issues that would prevent the use of traditional cohort-component methods, the author utilized a combination of cohort-change-ratios (CCRs) and cohort-change-differences (CCDs). Beyond permitting projections using extant county-level data, this dual approach also avoid the pitfall of unreasonably high (exponenetial) growth in small population areas (a property of CCRs) and potential negative values (a property of CCDs) for areas where population is expected to decrease.

Projections range from 2020 to 2100 and are given by the following categories,

- **Age** (0-4,5-9, …, 80-84, 85+)
- **Sex** (Male, Female)
- **Race** (White, Black, Hispanic, Other)

Table 2.1: Data Sources

Data Set	Publisher
Utilization Data	
Medical Expenditure Panel Survey (MEPS)	Agency for Healthcare Research and Quality (AHRQ)
Geographic Population Data	
American Community Survey (ACS)	United States Census Bureau
Behavioral Risk Factor Surveillance System (BRFSS)	Center for Disease Control (CDC)
Population Projection Data	
Hauer (2019)[1]	Scientific Data

[1] Hauer, M. Population projections for U.S. counties by age, sex, and race controlled to shared socioeconomic pathway. *Sci Data* 6, 190005 (2019).

III. Process

Harmonization & Aggregation

Due to the fact multiple demographic variables were measured using different response scales/categories across MEPS, BRFSS, and ACS-PUMS, it was necessary to transform variables within both surveys so that all categories were aligned. Generally, this involved aggregating categories to a shared common denominator in the case where one survey originally had more categories (e.g, differing levels of education).

It was also occasionally necessary to harmonize variables across different survey years due to changes in questions over time. As noted elsewhere, survey's from multiple years were combined in order to create a larger and more stable sample-size in order to derive the utilization model.

The most recent five (5) years of MEPS survey data was combined to create the model data set. Predictor variables included demographics (age, gender, race/ethnicity, income, etc.) as well as chronic conditions/major health incidents (diabetes, history of stroke, etc.). Additional variables were census region and a year indicator variable.

Similarly, the five (5) most recent years of the BRFSS were also combined to increase effective sample size for the adult estimation data set. Given the large annual sample size of ACS-PUMS, only the most recent year was used for estimating child utilization.

Model Variables

This section describes the variables used in the LAHUP models. Outcomes include utilization in terms of visits (e.g., ER and service line specialties) as well as total length of stay for inpatient admissions. Explanatory variables primarily consist of demographics and chronic health conditions consistent with research on the determinants of healthcare utilization behavior. Additionally, predictors were also selected based on their presence in both MEPS and BRFSS.

Outcome Variables

Service Line Utilization. Service line organization, long utilized in other industries is becoming increasingly prevalent in healthcare Sciulli & Missien (2015). In essence, an organization (e.g., hospital, system, etc.) that adopts a service line approach creates divisions based on clusters of patient diagnostics (DRGs) each with their own specific support staff and resources (Sciulli & Missien, 2015). Not only does this have the potential ensure higher quality of patient care due to better coordination within a particular line, but also greatly aids organizational management in terms of profitability analysis and strategic planning Roberts (2021).

Service line utilization rates were determined by setting (e.g., outpatient, office, etc.) and physician specialty (e.g., Pediatrician, Cardiovascular, etc.) using the MEPS Household Component files. This detailed visit data was then linked and merged with the overall annual MEPs file containing information on demographics and health status.

For each individual in the data set total visits by setting and provider specialty were summed within a calendar year. Those individuals who did not have any record of visiting a specific location and specialty (e.g., office and primary care physician) were assigned a value of zero. Variables comprising service lines were summed across office and outpatient settings.

These constructed utilization variables served as the basis for the modeling described in the following section.

Table 3.1: Service Lines

Service Line	Service Line Cont.
Cardiology Visits	Internal Medicine Visits
Dermatology Visits	Nephrology Visits
Emergency Room Visits	Neurology Visits
Endocrinology Visits	Oncology Visits
ENT (Otorhinolaryngology) Visits	Ophthalmology Visits
Family Practice Visits	Orthopedics Visits
Gastroenterology Visits	Pediatrics Visits
General Practice Visits	Primary Care Visits
Gynecology / Obstetrics Visits	Psychiatry Visits
Inpatient Admissions	Urology Visits
Inpatient Nights	

Explanatory Variables

Adult Model Variables. Demographic, health, and disability model variables are summarized in Table 3.2, Table 3.3, and Table 3.4 respectively below. Note that the inclusion of specific variable within a model depends on a) the availability of the variable across data sets and b) the contribution of the variable to the overall model fit (AIC/BIC).

Child Model Variables. The child model only included a subset of the demographic models as these were the only common variables across MEPS and ACS.

Table 3.2: Demographic Variables

Variables	Categories
Age	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, ..., 80-84, 85+
Sex	Male, Female
Race	White, Black, Hispanic, Other
Marital Status	Married, Not-Married
Household Income*	<$10,000, $10,000 to <$15,000, $15,000 to < $20,000, ..., $75,000+
Education Level*	No Degree, High School, At Least Some College
Insurance Coverage Status	Yes, No
Census Region	Northeast, South, Midwest, West

* indicates variable was only used in the adult model.

Table 3.3: Health-Condition Variables

Variables	Categories
Smoking Status	Yes, No, Missing
Stroke History	Yes, No
Emphysema	Yes, No
Cancer History	Yes, No
Diabetes History	Yes, No
Myocardial Infarction History	Yes, No
Heart Disease History	Yes, No
Asthma History	Yes, No
Arthritis History	Yes, No

Note:
Variables were used in adult model only.

Table 3.4: Disability/Difficulty Variables

Variables	Categories
Difficulty Dressing	Yes, No
Difficulty Doing Errands	Yes, No
Difficulty Hearing	Yes, No
Difficulty Seeing	Yes, No
Difficulty Walking	Yes, No

Note:
Variables were used in adult model only.

Model Building

Variable Selection

MEPS utilization variables are almost always operationalized as the sum total of visits to a specific provider within an annual period. An exception to this is total inpatient nights (LOS). For home health utilization, please see our LAHUP-HHA Technical Documentation. Considering healthcare utilization variables are often over-dispersed, models were primarily run using negative binomial regressions [NB2; Cameron & Trivedi (2013),Hilbe (2011)]. In the event that a model failed to converge, showed poor fit and/or predictive power, or when basic count distribution norms were severely violated, an alternative model was used. This may involve simply using a standard GLM Poisson model, a quasi-Poisson model, or in the case of an atypical distribution, using a Poisson/NB2 hurdle model or generalized additive model [GAM; Hilbe (2014),Wood (2017)].

As noted in the previous section, adult models were also run hierarchically based on three variable groups:

1. Demographic Variables
2. Health Conditions
3. Disabilities/Difficulties

Across outcomes, the model fit - assessed via AIC, dispersion, and residual analysis (Hilbe, 2014) - was consistently best when all three groups of variables were included and thus the a model that included all explanatory variables was ultimately used.

Model Estimation

Utilization models were developed and estimated using the aggregated MEPS data from the prior five (5) years. Two separate models were run, one for adults and one for children. Once a final model was selected, using the criteria described in the previous section, the coefficients were saved for use with either BRFSS (adults) or ACS-PUMS (children).

Again, it is important to reiterate that the models developed in MEPS are estimated in geographically detailed data sets for the purpose of capturing configurations of population demographics and health conditions within a geographical region, but not area-specific (e.g., county) variation which is not available in MEPS.

After running the adult and child models using BRFSS and ACS-PUMS, respectively, the predicted utilization rates were saved for use with the county-level demographic projections. Thus, the estimation process can be summarized in the following steps,

1. Estimate adult/child models with MEPs
    - Save final model coefficients
2. Run models with BRFSS/PUMS
    - Save predicted utilization rates

County-Level Estimates

In order to obtain future utilization estimates by county, the predicted utilization rates were aggregated by the broad demographic characteristics present in the county-level population projection file and then multiplied and then multiplied with the projections. The aggregation process (weighted mean of predicted utilization) was done at the state-level for those states with population greater or equal to the median population of all states. Whereas states with population lower than the median U.S. state population were aggregated at the divisional (census defined). So, for example, states such as Wyoming or Montana were aggregated at the "Mountain-Division"-level whereas a state like Texas was aggregated at the state-level.

As noted previously, model predictors (age, asthma history, etc.) are all comprised of k unique categories. Within a specified region, these k categories were averaged (via survey weighted means) together at the age, sex, and race categories. In other words, the expected annual utilization rate of a male, 37 year-old Hispanic with a household income of $72,000 and hypertension - to list a sample of detailed characteristics - is then weighted and averaged with all other Hispanic males in the age range of 35-39 within a given county. Thus, assuming present trends remain stable (assumptions discussed in the following section), the resulting utilization rate serves as a constant multiplicative factor for the county-level population going forward into the future.

IV. Assumptions & Limitations

In the most basic terms, the accuracy and validity of our utilization projections can be summarized by the following equation,

$$\begin{aligned}&(\text{Utilization Model Accuracy}) \times (\text{Pop. Projection Model Accuracy}) \times (\text{Stability of Current Trends})\\ &= \text{Utilization Projection Accuracy}\end{aligned}$$

Accordingly, the following subsections provide detail on factors affecting the components listed in the above equation.

Assumptions

As with any projection methodology, LAHUP both operates based on a set of assumptions and includes a number of limitations. The first and foremost assumption is that,

- Present trends continue into the future

Considering healthcare in the U.S. is constantly the subject of new, potential legislation as well as more general industry upheaval, this assumption is unlikely to hold in the long-run. Other projections such as HRSA have previously attempted to incorporate sensitivity analysis for potential future changes such as ACA healthcare reform. This approach is far more feasible when potential policy implications are known or can be anticipated to some degree. However, the impact of other, particularly novel, events is much more difficult to forecast.

For example, at the onset of the COVID-19 pandemic, there was widespread concern over emergency departments becoming overwhelmed with cases, but aside from a major influx in several urban hot-spots, this did not come to pass. Instead, the result was that overall healthcare utilization was reduced due to restrictions and provider reallocation of resources, but then rebounded with "bounce-back" effect once the emergency was over. Thus, the exact effects of unexpected events can only be assessed several years down the road once the data becomes available. This example demonstrates the difficulty of anticipating and modeling major healthcare-related events in the future.

That said, the fact that LAHUP is updated and produced on an annual basis helps ensure projections are based on the most recent and accurate historical data available. Thus, models and resulting projections reflect the impact of healthcare changes as soon as possible.

Limitations

Geographic Specificity

There is a wide variety of methodology available for making local or small-area estimates including standard unit and area estimation techniques (Rao & Molina, 2015) from the statistics literature as well as multilevel regression and post-stratification [MRP; Zhang et al. (2014),Zhang et al. (2015)] techniques from epidemiology and public health. However, these all require at least some data linked to the small areas of interest. Considering the fact that lowest unit of geography in MEPS is census region, small area-estimation (SAE) is not a feasible approach within LAHUP. Thus, estimates and resultant future projections are generated through a "bottoms-up" approach based on individual, but *not* geographic characteristics. As discussed previously in the modeling section, this means that geographic-specific variation is not specifically incorporated in LAHUP models.

Data Dependency

All projections are dependent on the model inputs. Specifically, they primarily depend on the representativeness/accuracy of MEPs, BRFSS, and ACS-PUMS as well as the demographic population projections.

Although U.S. government data products have been, as a whole, shown to be highly representative and accurate, even for sensitive behaviors [e.g., drug use; Pierannunzi et al. (2013)], there are a number of potential limitations with respect to use in LAHUP.

In particular, the identification and classification of certain types of utilization (e.g., non-physician primary care visits) can be challenging given the available variables in the data. Thus, in many cases, results should be interpreted within the context of specific operationalizations described in this document as well as the government's data documentation.

References

Bureau, U. S. C. (2020). *American community survey*. https://www.census.gov/programs-surveys/acs

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

Disease Control, C. for. (2020). *Behavioral risk factor surveillance system*. https://www.cdc.gov/brfss/index.html

Hauer, M. E. (2019). Population projections for u.s. Counties by age, sex, and race controlled to shared socioeconomic pathway. *Scientific Data*, *6*. https://doi.org/10.1038/sdata.2019.5

Healthcare Research, A. for, & Quality. (2020). *Medical expenditure panel survey*. https://www.meps.ahrq.gov/mepsweb/

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Nevers, R. L. (2002). A financial argument for service-line management.(business). *Healthcare Financial Management*, *56*(12), 38–43.

Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (BRFSS), 2004–2011. *BMC Medical Research Methodology*, *13*(1), 1–14.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). John Wiley & Sons, Inc.

Roberts, S. (2021). Service line development serves to support the entire system. *Frontiers of Health Services Management*, *37*(3), 29–34.

Sciulli, L. M., & Missien, T. L. (2015). Hospital service-line positioning and brand image: Influences on service quality, patient satisfaction, and desired performance. *Innovative Marketing*, *11*(2), 20–29.

Wood, S. N. (2017). *Generalized additive models* (2nd ed.). Taylor & Francis Group, LLC.

Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., & Croft, J. B. (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: A case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American Journal of Epidemiology*, *179*(8), 1025–1033.

Zhang, X., Holt, J. B., Yun, S., Lu, H., Greenlund, K. J., & Croft, J. B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology*, *182*(2), 127–137.