Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems

Benjamin Knisley, MS, Department of Behavioral Health Research and Evaluation, San Bernardino, California
Joshua Morgan, PsyD, Department of Behavioral Health Research and Evaluation, San Bernardino, California
Linn Carothers, PhD, California Baptist University, Riverside, California
Keith Haigh, MPA, Department of Behavioral Health Research and Evaluation, San Bernardino, California

ABSTRACT

The California Medi-Cal 2020 waiver includes a five-year pilot program known as Whole Person Care (WPC), which focuses on health interventions that coordinate physical health, behavioral health, and social service needs of beneficiaries who are high users of multiple county entities. San Bernardino County developed an analytic approach combining and matching health and social services data from multiple County departments utilizing disparate systems and identifiers. This required collaboration of data experts in multiple County departments to produce the most accurate matching approach. In many high utilizer projects, potential service recipients are identified through costs or basic utilization. However, both of these have limitations and may miss individuals who are not appropriately engaged in their care or their care is not appropriately coordinated. Through an iterative process, a scoring methodology was developed to stratify utilizers of County health services to identify individuals who are most likely to need care coordination and health engagement services. Additionally, in order to begin building and testing a predictive model, a retrospective cohort was evaluated and scored and multiple logistic regression was implemented to demonstrate the factors and service utilization patterns that most contribute to high utilizer. Over time, and with new incoming data, the model will be refined to better assess the combinations of factors, services, and score methodology that predict those who most need care coordination services to improve the quality of their care and access to outpatient services for better health outcomes.

INTRODUCTION

In San Bernardino County, data analytics is essential to truly understanding the health of our population and telling the story of our consumers and services. The deep connections between health and non-health data have very real meaning and potentially profound effects on the lives of our residents. Utilizing analytics enables our partners and us to better serve our residents by more effectively advocating for stigma reduction and needed care and promoting awareness, wellness, resilience, and recovery in our community (Lowman, 2017).
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

WHOLE PERSON CARE WAIVER & INTENT

In keeping with the vision for a healthier community, San Bernardino County has invested in the resources to perform big data analytics to transform the continuum of care.

This goal is supported by legislation and federal waiver projects that solidify the commitment between the state and the counties to improve the quality of care in the outpatient system while reducing the cost of care for acute, higher cost services, such as emergency room visits and hospital stays in order to strengthen the performance and quality of California's health care delivery system (DHCS, 2015).

Statewide, California’s latest 1115 Medicaid Waiver Renewal is entitled Medi-Cal 2020 and includes multiple projects, including the Whole Person Care pilots. Medi-Cal 2020 is facilitating investments of $15 to $20 billion dollars in federal funds to promote system transformation, which makes possible deeper integration through complex care coordination across the county system of care by including physical health, behavioral health, and long-term care providers to improve health outcomes and quality of life overall. Furthermore, the rapid increase in Medi-Cal enrollment (from 7.6 to 13.5 million - nearly 78% or roughly 5.9 million people) due to Medicaid expansion and the advancement of Medi-Cal managed care throughout the state and across populations are important achievements and pave the way for new opportunities for California to serve its population. More specifically, the Whole Person Care 1115 waiver is designed to improve the quality of care and ultimately the health of Medi-Cal members with multiple diagnoses and complex care needs who qualify for intensive care coordination support under the program to improve quality and health care outcomes across multiple settings of care for the most complex members (DHCS, 2015 & 2017). In summary, to promote and assess the ongoing developments of this programs initiative the following core objectives will be fully addressed and constitute the aim of San Bernardino County to address these needs:

1. Improve health care quality and outcomes for the Medi-Cal population
2. Improve access to primary and specialty care services
3. Address social determinants of health and improve health care equity

To help meet these objectives a multiple-period clinical decision model has been developed in SAS® to aid clinical and hospital administrators with program needs and multiple-user specifications. In particular, with those clients who are deemed to be high-utilizers, we propose a model and methodology to identify early on those clients who are at high-risk during their continuum of care. In effect, the chief goal of the department is to help improve the quality of life for consumers by providing the most appropriate behavioral health services, in the least restrictive manner, at the earliest stage possible. This strategy benefits the consumer, as well as, saves taxpayer dollars.

COUNTY HEALTH DATA SYSTEM STRUCTURE

The County of San Bernardino Department of Behavioral Health (DBH) offers a wide range of mental health and substance use disorder services for children, transitional age youth, adults, and older adults. Moreover, San Bernardino in figure 1 is geographically the largest county in the
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

contiguous United States, covering 20,052 miles from Los Angeles on the west to Arizona on the east, which covers more territory than Connecticut, Massachusetts, and Rhode Island combined. Therefore, the creation of the SAS data warehouse to integrate approximately a dozen different systems, many of which do not interface directly with one another is a strategy to monitor multiple performance outcomes across the county and to improve clinical and administrative decision-making. Due to the vital role that data and analytics plays in population health initiatives, DBH collects data from a variety of disparate data sources to form a more holistic view of the residents it serves. While the primary data sources are behavioral health in nature, the department realizes the impactful relationship that other data like physical health, criminal justice, and other social services contain and has included these data and their owners in its analytic endeavors (Lowman, 2017).

![Map of the contiguous United States with states highlighted]

**Figure 1. San Bernardino County, California – the Largest County in the United States.**

**DATA SHARING CHALLENGES**

Health information is regulated by multiple federal and state laws. Within the work of Whole Person Care, there are multiple health and social services laws and regulations that must be navigated, creating differences and conflicts among what data can be shared with the various involved organizations. In most cases, behavioral health information sharing laws are the most restrictive. Other health and social services systems are more easily able to share data with behavioral health systems than the reverse. Therefore, data management and analytics for this Whole Person Care project is being led by the Department of Behavioral Health.
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

Diverse data collection and repository systems and approaches, not all of which are electronic, lead to practical barriers to integrating and matching data. Under the current approach, due to both the practical system limitations as well as legal privacy and security requirements, each Department must export specific data sets from their systems and securely transmit them to Behavioral Health for integration and processing. Integration and matching is facilitated by DBH’s data warehouse, which provides the flexibility and technical power needed to more efficiently and effectively conduct this processing. Due to the manual exports required at this time, data is always delayed and so is a point in time snapshot.

DATA FLOW PROCESS

Integrating health information in the county requires robust systems to generate matched datasets, scoring method implementation, and predictive analytics. With this triple aim SAS® Enterprise Guide™ was used successfully to implement three key phases of data integration; namely, extraction, transformation, and loading (ETL) methodologies. First, extraction of hospital/inpatient encounters, outpatient services, and housing status information is gathered from all participating entities – afterwards, transformation of data is performed to cleanse, reformat, standardize, aggregate, and apply standard business rules and practices to our incoming datasets. In the end, resulting datasets are converted into a specified target file format to be loaded and assessed for final statistical analysis.

MATCHING METHODOLOGY

The current version of the matching process is mainly deterministic. Identity matching starts with a direct one-to-one match of social security number (SSN), where they exist across datasets, which is augmented by a one-to-many fuzzy match of first name, last name, and date of birth (DOB) using COMPGED. Also, other functions such as COMPLEV and SPEDIS were contemplated but not used when considering the wide-flexibility and precision COMPGED achieves with our datasets. Moreover, performance issues were not primarily our focus due to our need to obtain accurate matches. Resultantly, DOB is broken out into discreet month, day, and year values, and those are compared individually, and the combined score must be less than 100. For last name COMPGED score, the selected acceptable value is less than 150, which allows for slight variation in spelling as well as common hyphenated last names. The first name score is recorded and available as a filter, although this has not been necessary in this step. These matched records are then removed and the remaining records are evaluated on first name, last name, and DOB combined. In this round, COMPGED is again used, but for evaluating the distance between SSN records, which proved to be easier to develop in Enterprise Guide (EG). Where records are not available across data sets, the value is 1800. When they are very far apart, the value is up to 1000. When they are close, as in a typo, or dropped leading zero, the value is 200 or less. For this round, SSN score with COMPGED must not be between 201 and 1000. These two methods provide the bulk of matched records.

At this point, the remaining unmatched records are evaluated where last name matches and first name, DOB, and SSN are close matches using COMPGED. DOB again is scored at <=100 to allow for typos, and SSN must not be between 201 and 1000. At 200 several first name mismatches were introduced so first name is less than 200. These are sorted in descending order
by first name score and remaining questionable matches are filtered out. The remaining unmatched records are tested for first name matches with COMPGED scores for last name of <100, DOB <=100, and SSN not between 201 and 1000, allowing for close matches of hyphenated last names. The final test is for exact match by DOB with first name match <200 and hyphenated last names separated and exact matches of separated last names required. Future matching will include gender as well as lessons from comparisons to SEER (Surveillance, Epidemiology and End Results) for ongoing validation (Dusetzina, Tyree, Meyer, Meyer, Green, & Carpenter, 2014).

**SCORING METHODOLOGY**

Potential enrollees are identified via a quantitative stratification based on scoring county health care utilization as well as housing status. Stratification of the population requires a scoring methodology that appropriately weights interactions according to severity, frequency, and duration. Scoring methodologies are a common practice in healthcare analytics in order to tailor program planning and interventions to various levels of at-risk populations, according to acuity and severity of the need within a high-need population.

Inspired by Dr. Jeffrey Brenner’s work in Camden, NJ, and a scoring methodology developed by the Santa Clara Valley Health and Hospital System, the County conducted a hot spotting pilot analysis to assess its local applicability as well as local needs. This pilot analysis included data sets representing beneficiaries served by at least one of five County health or social services Departments. This approach interprets utilization of primarily routine outpatient encounters as an indication of appropriately managed care, whereas higher numbers of encounters in urgent or crisis services, emergency department visits, or hospitalizations (including medical and behavioral health) imply a lack of effective case coordination. For this reason, most routine outpatient services, like primary and specialty care encounters, were not to be initially included. Public health primary care clinic encounters were included in the analysis to see the frequency of contact across County systems, and to include some potential individuals who heavily (and likely inappropriately) use public health services. The quantitative analysis is based on the frequency and duration of patient encounters with multiple systems of care and required matching patients across these systems by unique identifiers, including name, social security number, date of birth, and record numbers.

Preliminary testing of this methodology has provided effective stratification of beneficiaries served across County Departments. This methodology is used to identify the top scoring beneficiaries (redone intermittently to identify new potential participants and expand the population, as previously defined), who the County attempts to contact to enroll in the WPC project. The County has continued to refine the methodology in order to appropriately weight various utilization and conditions. For instance, a particular area of score testing was for housing status. An individual with a risk of homelessness should be prioritized and therefore higher in the stratification. However, this risk alone should not place a potential enrollee in a top tier of scoring. Therefore, multiple score iterations were tested to find a weighting that could represent an appropriate balance of not over- or under-weighting this social determinant.
LESSONS LEARNED

Major lessons learned from this process are two-fold. First, the challenges from our matching methodology and data flow process led to the need to fully develop preemptive methods that would be embedded in the process architecture as integrity checks. These checks are meant to maximize data integrity by providing internal review of both the system datasets and the final data processing prior to analysis. Analytics programming in combination with expert eyes on the data helps to doubly ensure the quality of our datasets throughout every step of this process. Further, using advanced analytics software enables San Bernardino County to streamline and process results efficiently. However, retrospective methods are required in order to diagnose potential errors within the data, specifically, a cleaning focus must be adopted that removes duplicates, eliminates unwanted data elements, and scans for mismatches with programmed autocorrecting capabilities. Overall, when both methods are applied correctly the quality and quantity of records that we obtain are optimal for analysis.

MODELING FOR FUTURE PREDICTIONS OF HIGH UTILIZERS

One of the projected problems of this study is the likely phenomenon known as the regression to the mean. That is, in several studies extreme observations that warrant immediate attention will tend to move closer to the mean only to disappear from the system before any practical investigation or intervention can be performed in a timely manner. Moreover, among our community partners there is strong evidence to suggest that new sets of observations will emerge to indicate new high utilizer populations, making the tracking process much more difficult and expensive for any prospective endeavors. Resultantly, the time needed to track the occurrence of high utilization prospectively is both difficult and infrequent which in turn encourages our investigation to consider and evaluate retrospective studies in order to better track the occurrence of high utilization in our population. Hence, as an alternative to prospective cohort studies we turn our attention to case-control or retrospective studies in which individuals are known either to have the occurrence of a disease or not. Furthermore, our clients are then tracked backwards in time in order to decide whether or not the factor of interest in our prediction equations has been present (Everitt, 1992). As an example, we consider the 1976 Prentice evaluation and study of logistic regression in light of retrospective studies: Consider a dichotomous exposure variable, F, and a disease indicator variable, D, such that values of 1 and 0, for both F and D, refer to presence and absence, respectively. The association between F and D may be studied retrospectively by looking back from diseased (case) and disease free (control) individuals to ascertain exposure (Prentice, 1976). Furthermore, the analysis of retrospective studies is complicated by the need to accommodate factors that are related to exposure and discriminate between cases and controls (Prentice, 1976). More specifically, confounding and effect modification are essential components in the final evaluation checking for both parsimony and validity within the model specifications.

In later sections we will further probe the various contributing variables that are responsible as effect modifying in addition to those that are confounding our final results. This careful consideration will enhance model discrimination and provide a more robust understanding of the model's predictive capacity. For example, in research directed to the effect of exposure on the risk of developing a particular illness, a central problem is the need to consider extraneous
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

factors that might be explanatory, partially or totally, of the magnitude of the estimate of the effect. An understanding of the nature of such factors, referred to as “confounders” or “confounding factors,” is thus essential to study design and data analysis – and finally to the interpretation of the resulting estimates (Meittinen & Cook, 1981). Moreover, in the detection of confounding it is commonplace to use the data-based criterion that control of the extraneous factor changes the estimate of the effect; that is, after running our regression do the associated coefficients with each designated predictor variable change by more than ten percent (Meittinen & Cook, 1981). Also, an alternative to this is that the extraneous factor be associated, in the data, with both the exposure and the illness, for example, suppose drinking tea is associated with smoking (perhaps people who drink tea also tend to smoke) the smoking forms a confounding variable that can make it falsely appear as though tea drinking leads to heart disease (Meittinen & Cook 1981). In effect, the decision of specifying which variables are potential confounders should not just be based on quantitative methods; we must also consider the possible causal relationships between the exposure, outcome, potential confounders, and other relevant variables in order to eliminate bias (Kleinbaum & Klein, 2010).

This possible confounding interaction in addition to effect modifications will fully be explored in upcoming sections, interestingly, it is important to note that modeling the retrospective probabilities of exposure by means of a binary logistic regression model leads to a direct estimation of the odds ratio associated with the exposure and of the dependence of the odds ratio on other explanatory variables (Prentice, 1976). Thereby, this realization leads us to a very important result in the upcoming section where we test for effect modification using a likelihood ratio test. Hence, after checking our model for any potential confounders or effect modifiers we have effectively calibrated our model to now be assessed for its potential to discriminate between clients that are and are not high utilizers, and then test for model validation using several sets of data spanning different time periods. Lastly, in order to avoid ambiguity in this analysis the term “discrimination” refers to the ability of the model to distinguish between those clients having the high utilization outcome and those not having this condition. In the following section, model discrimination will be meticulously analyzed with the area under the receiver operating characteristic curve (ROC) and its appropriate c-index. In addition, “validation” or “validity” may be defined as the demonstration that the predictive accuracy of the model is similar when it is applied to a different group of clients than those used in the construction of the model. Doing so will then allow us to more fully address our populations needs over time which will create comparisons that more fully explain the predictive power (i.e., estimating efficiency) of the model currently under review in this retrospective study (Prentice, 1976; Anderson, Jin, & Grunkemeier, 2003).

THE MODEL

Multiple logistic regression is the statistical technique of choice when we wish to estimate the probability of a dichotomous outcome (i.e., binary variable) such as the presence of an exposure like disease in the midst of multiple predictors (Anderson, Jin, & Grunkemeier, 2003). In particular, due to the size of our data the training efficiency for logistic models are quite good despite the curse of dimensionality. That is, as the number of features in our data increase, the distance between points grows and observations are more likely to be linearly separable
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

(Wujeck, Hall, & Günes, 2016). Furthermore, a logistic regression model is a simple and clear-cut model for deployment along with the fact that it will be less prone to overfitting the training data since no high output algorithmic fitting is employed (Wujeck, Hall, & Günes, 2016). In other words, the logistic regression model provides a great opportunity to generalize our high-volume behavioral health data for future findings and predictions. Furthermore, figure 2 provides the overall layout of this statistical process and the fundamental steps needed to construct our final model.

\[ \text{Mathematically, we can describe the model with the following expression:} \]

\[ \text{Figure 2. Model Construction and Statistical Process Flowchart.} \]

Before proceeding, several assumptions are required for this logistic regression model to fit well, in particular, it is assumed that the predictors are uncorrelated with one another, that they are significantly related to the response, and that the observations or data elements of a model are also uncorrelated (Hilbe, 2014). Moreover, the covariates must exhibit a linear relationship with the outcome of interest. With this technique, the maximum likelihood method is used to determine objectively the statistical weights to be assigned to each variable (Lemeshow, 1988). In this section, we will describe the essential components of the logistic regression model and its general relationship with the data itself. Mathematically, we can describe the model with the following expression:
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

\[
\log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,
\]

Where the logit form of the logistic model, \( \logit(P(X)/1-P(X)) \), is defined as the natural log of the odds for developing a condition for a person with a set of independent variables specified by \( X \) (Kleinbaum & Klein, 2010). Moreover, each \( \beta \)-coefficient is estimated for all corresponding explanatory variables and intercept (i.e., \( \beta_p, p = 0, 1, \ldots, p \)). In effect, the logistic regression equation uses past experience of a group of clients to estimate the odds of an outcome by mathematically modeling or simulating that experience and describing it by means of a regression equation (Anderson, Jin, & Grunkemeier, 2003).

In this analysis the response variable represents a condition that we call high utilization. To determine high utilization, an in-house scoring mechanism was developed to generate appropriate groups. The scoring mechanism combined the scores across multiple service departments in order to provide a measure that classified which consumers were high utilizers. This final score, called the WPC score, was generated with the addition of all other service scores obtained from the breadth of departments used in this analysis. All other explanatory variables will be multiple service department scores that are used in conjunction with classification. Consequently, each explanatory variable represents a particular service department that participated in the production of the total WPC score. Clients who fell into the ninety-fifth percentile of this total score were retained for further analysis, as these clients represent the highest utilizers in the county. Our objective with the logistic regression model is then to predict which combination of service departments among other demographic and service scores are most likely to predict the outcome of a high utilizer. In San Bernardino County, this was demonstrated and several methods were implemented to check model calibration, model discrimination, and model validation – these and other assumptions will be fully discussed in the following sections.

**MODEL CALIBRATION**

Model calibration is focused on determining the “best” model when carrying out mathematical modeling using logistic regression. In particular, the strategy for San Bernardino County employs a combination of the traditional Hosmer, Lemeshow, Kleinbaum and Klein paradigms which involves three key stages: (1) variable specification, (2) confounding assessment, and (3) interaction assessment (Hosmer & Lemeshow, 2000; Kleinbaum & Klein, 2010). When all three steps are performed adequately we say that our model has been successfully calibrated for discrimination analyses.

Variable specification heavily integrates and involves the process of data reduction; that is, from all available variables, only those most associated with outcomes are selected for inclusion in the final model (Lemeshow, 1988). That is, for nominal, ordinal, and continuous variables with fewer integers the test of association is performed using a contingency table of outcome \((y = 0, 1)\) versus the \(k\) levels of the independent variable. In contrast, for continuous variables association is assessed by fitting a univariable logistic regression model to obtain the estimated
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

coefficient, the estimated standard error, the likelihood ratio test for the significance of the coefficient, and the univariable Wald statistic. Moreover, the traditional approach to statistical model building involves seeking the most parsimonious model that still explains the data. The rationale for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and is more easily generalized (Hosmer & Lemeshow, 2000).

In our study, high utilization is assessed against a vast backdrop of potential risk factors to evaluate and analyze. With our strategy, we begin with a set of control variables that we primarily suspect to be highly associated with the outcome of high utilization. Afterwards, we specify potential confounding variables based upon the expert domain knowledge of clinicians. Furthermore, confounding variables are chosen carefully bearing in mind past research and statistical problems such as multicollinearity that might result from certain choices that may go unnoticed (Kleinbaum & Klein, 2010). In addition, Pearson correlation coefficients are computed to test for high correlations that indicate issues with co-linearity and confounding. Altogether, the end of this process permits San Bernardino County to define the largest possible model initially to be considered.

Next, with all variables selected, the data reduction methods are performed using forward stepping multiple logistic regression analysis to obtain a major reduction in variable population. Moreover, the remaining variables from the previous analysis and only significant variables are kept in the model (Lemeshow, 1988). With features that were missed we confer with domain specialists to obtain confirmation for their exclusion. Immediately following the fit of this multivariable model, the importance of each variable is then verified with an examination of the Wald statistic at a cut-off alpha (α) level of 0.20, and a comparison of each estimated coefficient with the coefficient derived from the model containing only the variable under review to assess confounding (Hosmer & Lemeshow, 2000). In this analysis we must consider how to control for these confounding variables when causal relationships obfuscate many predictors at once.

Addressing this issue requires us to examine the estimated β-coefficients. That is, do any of the estimated β-coefficients in any of the variables change by more than 10% and does this change appear to be “clinically” significant? If it is, then we should be concerned about their need to remain in the model since this indicates they are confounding variables. Moreover, variables that do not contribute to the model based on these criteria will be eliminated and a new model will be fit (Hosmer & Lemeshow, 2000). This new model will be compared against previous fitted models with the Likelihood ratio test while taking note of any β-coefficients with previous variables in our models that have changed dramatically. Resultantly, the model that remains will be known as the preliminary main effects model (Hosmer & Lemeshow, 2000).

Afterwards, all continuous variables that have entered into the model are subsequently checked in order to see if they are in fact linear in the logit. That is, does the relationship between the predicted logit values and the continuous variable possess a linear relationship? If it does not, we proceed to remove the indicated variable and reevaluate model fit. However, in cases where the continuous variable is necessary for model fit we will transform the variable to take on discrete categories to be assessed and fitted for ongoing model purposes. The conclusion of this second process leaves us with our main effects model (Hosmer & Lemeshow, 2000).
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

Next, we proceed with our interaction assessment by incrementally including those variable interactions that are deemed to be clinically significant. Afterwards, we check that our model is hierarchically well formulated, that is, all higher order interaction terms present in the model include the presence of all corresponding lower order terms in the model (Kleinbaum & Klein, 2010). Next, by a series of likelihood ratio testing comparing different models we identify the significance of terms within our model hierarchy. Specifically, we test for statistical differences between the main effects model with variations of the interaction models we suspect may prove to be more explanatory. Lastly, after confirming that we can no longer improve the model further we perform the Hosmer-Lemshow goodness-of-fit test to examine if the model has sufficient evidence to indicate that lack of fit has occurred (Hosmer & Lemeshow, 2000). If so, then we must reevaluate the model parameters to see which minimal adjustments are needed to calibrate the model to fit – and afterwards, when model fit is achieved we attain the preliminary final model (Hosmer & Lemeshow, 2000).

MODEL DISCRIMINATION

This section is focused on how to assess the discriminatory performance of a binary logistic model. In particular, the preliminary final model provides good discriminatory performance if the covariates in the model help to predict which clients will become high utilizers and which do not (Kleinbaum & Klein, 2010). In this analysis we use the 95th percentile as our cut-point throughout our modeling. Afterwards, making our predictions we generate a diagnostic table/confusion matrix to document the quantity of true as well as false positive and negative classifications. The proportion of true positives among all cases is called sensitivity, and the proportion of true negatives among all noncases is called the specificity. Ideally, perfect discrimination would occur if both are equal to 1. However, a known drawback to measuring discrimination as described above is that sensitivity and specificity that results from a single cut-point may vary with the cut-point chosen (Kleinbaum & Klein, 2010). Consequently, a helpful alternative would be to obtain a summary measure based on a range of cut-points chosen for a given model. Such a measure is available from an ROC curve. ROC stands for receiver operating characteristic and when applied to a logistic model, an ROC is a plot of sensitivity vs. 1 – specificity derived from several cut-points for the predicted value. Equivalently, the ROC is a plot of the true positive rate by the false positive rate. More specifically, an ROC provides an appropriate answer to the question, namely, how often will a randomly chosen (true) case have a higher probability of being predicted to be a case than a randomly chosen true noncase (Kleinbaum & Klein, 2010)?

In SAS, the Logistic procedure provides the “c” statistic, which gives the area under the ROC curve (AUC). This area will be used to assess the discriminatory power of the model under review. For example, a value of 0.50 would be equivalent to the discriminatory power of a fair coin. That is, there is really no discriminatory power with an AUC equal to 0.50 or less. However, with values up and above 0.70, there is reason to suggest that the model is performing fairly in discriminating high utilizers as it should. Moreover, equipped with the AUC we use this measure to determine which model between two competing methods is better in predicting high
integrating health data sources to identify, stratify, and predict high utilizers of public systems, continued

utilizers. That is, those models with a larger AUC are deemed to be improved models granted they have been screened and calibrated beforehand.

MODEL VALIDATION

At this point, given that we have sufficient and appropriate data, massaged the data into a form suitable for modeling, identified key features to include in the model, and established how the model is to be used we will now begin validating the model for production (Wujeck, Hall, & Günes, 2016). To do so we will go into a brief discussion of the model training, assessment, and the final selection process.

First, throughout the modeling process discussed so far, we have been using what is known as validation data. These validation datasets are holdout data that have been used to assess the model during training for the purpose of selecting variables and adjusting model parameters in order to generate a more accurate, generalizable model (Wujeck, Hall, & Günes, 2016). In this next stage, we will now examine to see how well the model performs with a test dataset derived from several time periods that we have prepared beforehand to assess the model’s effectiveness. Indeed, a convenient confusion matrix/diagnostic table from our discrimination analysis was used to capture the training error rate from our preliminary final model. Then, equipped with this error rate we may compare our logistic regression algorithm with more complex strategies such as decision trees, Bayesian CART, quadratic discriminant analysis (QDA), and others for future prediction analyses. More importantly, if clinicians and our department staff are satisfied with the current error rate obtained, we proceed to make use of this now final model to begin early classification of potential high utilizers in the county. In the future, as more data is being collected we would immediately reassess the model’s effectiveness and reorient the model to ensure smaller if not equivalent error rates are maintained with future predictions.

POLYTOMOUS LOGISTIC REGRESSION

Throughout this analysis the County of San Bernardino has been primarily focused on developing a dichotomous logistic regression model to predict high utilizers in the county. However, it is of great interest not only to predict whether or not a client becomes a high utilizer, but what level of severity in mental health we can identify for clients in their continuum of care for tracking and monitoring purposes. Hence, it is our desire to build a regression model that will predict and assign a severity category for each client in our health system. These multiple categories range from low severity to mild, moderate, and severe (i.e., potential high utilizer). To accomplish this task, we note that estimates of the logistic regression coefficients obtained by multinomial methods can be approximated by separately fitting logistic models to each category (Hosmer & Lemeshow, 2000). Afterwards, we specify cut-point percentiles for the various categories with a clinical focus in mind. When the cut-points are decided, we set out to build four separate logistic regression models that evaluate each category individually using the techniques outlined in this paper. Upon completion, the results are combined together to form the final multinomial model that will be used to make clinical predictions and classifications. Not only will performing this analysis enhance the amount of detail we can make with our previous
conclusions, but it enables us to further investigate new unexpected findings and their nuanced interactions.

CONCLUSION

In a typical clinical setting, many decisions are made every day. In fact, in most clinical scenarios the decision making process commonly used to guide clinical care is mostly ad hoc and informal (Rosenberg, Joseph, & Barkun, 2000). In minor cases where few inputs or alternatives are present, these methods are usually sufficient. However, the truth is it becomes exceedingly difficult to think in many dimensions simultaneously. That is, when clinicians are presented with complex problems, poorer decision making becomes more frequent and often requires much assistance. Faced with these difficulties, we desire a systematic approach to decision making under conditions of uncertainty (Rosenberg, Joseph, & Barkun, 2000). Resultantly, San Bernardino County has chosen to partake in this mission and has proposed the usage and implementation of multiple logistic regression to serve our clients, clinicians, and the community to help us better understand, explain, and predict high utilizers within our system of care. More importantly, clinical decision modeling support is the aim for our department and San Bernardino County is actively seeking ways to leverage analytics to better partner with our local community stakeholders as allies to help better serve our clients in their continuum of care.

REFERENCES


Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued


**ACKNOWLEDGEMENTS**

First, we would like to give a very special thanks to Megan Daly for her amazing and professional support both in her project management endeavors and trouble-shooting methodology. Furthermore, we would like to thank Nicole Caldwell both for her reviews and contributions to further a deeper understanding of the Whole Person Care initiative. Last but not least, we would especially like to thank Kirk Paul Lafler for his professional editing and drafting support both as our mentor and exhorter to go above and beyond with our current work.
Integrating Health Data Sources to Identify, Stratify, and Predict High Utilizers of Public Systems, continued

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Benjamin Knisley, MS
Enterprise: San Bernardino County Department of Behavioral Health, Research and Evaluation
Address: 303 E Vanderbilt Way
City, State ZIP: San Bernardino, CA 92408
Work Phone: 909-388-0997
E-mail: BKnisley@dbh.sbcounty.gov

Name: Joshua Morgan, PsyD
Enterprise: San Bernardino County Department of Behavioral Health, Research and Evaluation
Address: 303 E Vanderbilt Way
City, State ZIP: San Bernardino, CA 92408
Work Phone: 909-388-0962
E-mail: Joshua.Morgan@dbh.sbcounty.gov

Name: Linn Carothers, PhD
Enterprise: California Baptist University, Department of Mathematical Sciences
Address: 8432 Magnolia Ave
City, State ZIP: Riverside, CA 92504
Work Phone: 951-343-4961
E-mail: lcarothers@calbaptist.edu

Name: Keith Haigh, MPA
Enterprise: San Bernardino County Department of Behavioral Health, Research and Evaluation
Address: 303 E Vanderbilt Way
City, State ZIP: San Bernardino, CA 92408
Work Phone: (909) 388-0961
E-mail: khaigh@dbh.sbcounty.gov

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.