Clarify

Clarify's Data Science Methodology

The Data, Architecture, and Modeling That Underpin the Clarify Atlas Platform®

OCTOBER 2023

WHITE PAPER

Table of contents

Introduction

Data science has the potential to revolutionize the management of patient health by identifying previously unknown patterns and hidden opportunities within healthcare data. However, for large-scale predictive analytics to be effective, it is essential to have a substantial amount of clean, high-quality data. Unfortunately, the healthcare industry has struggled with disorganized and fragmented data sources, poor data quality, and flawed methodologies. This has hindered the use of data science for healthcare decision-making, which has negatively impacted patient care.

Clarify's proprietary approach to cleaning and enriching vast, disparate sources of healthcare data delivers unmatched market intelligence to health plans, health systems, and life sciences companies. The Clarify Atlas Platform brings together one of the largest, longitudinal, patient- level datasets in the industry, aggregating traditionally siloed claims, prescription, and social and behavioral data from over 300 million lives. Clarify is one of only a handful of for-profit companies with CMS Qualified Entity (QE) status, which feeds the Platform with 100 percent of Medicare fee-for-service (FFS) datasets. These are richer than the more widely used Medicare Limited Data Set (LDS) provided to other vendors. Atlas ingests refreshed datasets and new data sources continuously via an automated data cleaning process that identifies and corrects outliers, distinguishes between unmarked screening and treatment diagnoses, attributes physicians to patients, and sorts claims into appropriate specialty- or disease-related categories.

The Clarify Atlas Platform is unmatched in its ability to elevate the usability of healthcare and social and behavioral data to a standard suitable for data analytics at a scale unseen in healthcare. By linking CMS claims data with commercial claims, prescription, and socioeconomic data, our models are trained on large cohorts and a more comprehensive picture of each patient's longitudinal healthcare journey. The enrichment of traditional healthcare data with data that are not typically accessible or usable optimizes our Platform's case-mix adjustments and allows our predictive models to uncover patterns that would otherwise be hidden from models built with claims data alone. This allows the Platform to generate patient- level predictive models that more precisely benchmark prior performance and predict future outcomes.

With hundreds of terabytes of enriched data at our disposal, our ability to extract intelligence that is meaningful and timely also requires a robust yet flexible Platform architecture that standardizes patient data. Atlas unifies regularly refreshed, patient-level datasets into a proprietary data schema, applying clinically relevant standard units of analysis (informed by Clarify's healthcare experts) to every patient's care journey. These units are the building blocks of Clarify's predictive models, which surface reliable and meaningful business insights in software solutions. With this universal framework, the Platform can ingest any patient-level data source and cut and analyze its data in seemingly endless ways.

The introduction of new care paradigms and restructuring of the healthcare system leaves payers, providers, and life sciences companies on a constant search for new ways to generate value. As clinical pathways and value-based care paradigms evolve, precision healthcare becomes a reality, and life sciences companies innovate at a faster pace, opportunities to improve quality and outcomes and reduce cost are certain to be moving targets. The advantage that the Clarify Atlas Platform affords is that it assesses performance and outcomes on-demand with continuously refreshed data and can answer new business questions as they arise. The Platform helps health systems grow revenue by building more precise growth strategies, accessing price transparency intelligence, and improving provider performance. It helps payers optimize networks, strengthen contract negotiations, enhance value-based contracts, and improve clinical effectiveness. And it helps life sciences companies accelerate clinical trial recruitment and maximize brand growth.

This paper describes the methodologies that Clarify uses to develop predictive models by detailing our end-to-end data science pipeline in five steps: (1) processing, (2) organization, (3) analysis, (4) modeling, and (5) interpretation.

Data Processing

Data Acquisition & Assessment

The Clarify Atlas Platform® links and tokenizes claims, prescription, and social and behavioral data at the patient- and physician-level across 300+ million lives. It includes 15+ billion claim records from CMS and commercial payers, 200+ million total payer complete lives (for which we have all claims for a patient within their enrollment period), 90 percent of all US prescription data, and 400+ social and behavioral determinants of health on every American. The wealth of social and behavioral data comes from sources such as credit agencies, consumer databases, and public and private records, providing patient-level information on housing stability, access to transportation, food security, and many other behavioral and individual-level attributes, allowing Clarify models to incorporate socioeconomic factors most relevant to healthcare predictions. Clarify's status as a member of the CMS Qualified Entity (QE) program feeds the Platform with 100 percent of Medicare FFS Part A (Hospital), Part B (Medical), and Part D (Pharmacy) CMS datasets and data on 60 million Medicaid lives, which are richer datasets than the Limited Data Set (LDS) that is provided to other vendors. The QE program is highly selective, with rigorous information security guidelines related to the handling and use of healthcare data.

According to a survey by Dimensional Insight (https://www.dimins.com/white-papers/survey-data-trust/), more than half of healthcare CIOs lack confidence in their data.

Clarify datasets are validated using a combination of:

- Library of validation rules developed by Clarify over time
- Comparison to national benchmarks
- Comparison to other Clarify datasets
- Statistically relevant feature comparisons
- Ad hoc analysis by teams of data analysts

Data Quality

In order to account for any data anomalies that may arise, there are several approaches may be used. One option is to omit patient records that may contain incomplete or inaccurate information or outliers. Another approach is to impute or predict corrected values for any questionable data points. Additionally, certain features within the dataset may be deemed untrustworthy and, therefore, ignored based on the rules and checks outlined above. By implementing these strategies, we work to ensure that our analytics start with training and inference data most likely to give consistent and defensible results.

Data Organization

Data Standardization

Each new dataset is normalized into a single schema. We then standardize it to our Abstract Claims Schema, and certain data elements are normalized. For example, gender data is normalized to consistent M and F values instead of a variety of other representations such as 0, 1, Male, Female, etc. Claim headers are generated from claim lines to simplify analyses and downstream usage.

Data Enrichment

The dataset is then enriched by a combination of industry & Clarify methodologies and reference datasets such as:

INDUSTRY
CMS software such as MS-DRG, CMS-HCC
Pharmacy reference data such as PQA, Medi-Span
Condition classifications such as CCSR, CCW

CLARIFY
Type of Service (TOS) Classification
Clarify Price Standardization
Specialty Claims and Cost Classification
PCP and Specialist Attribution

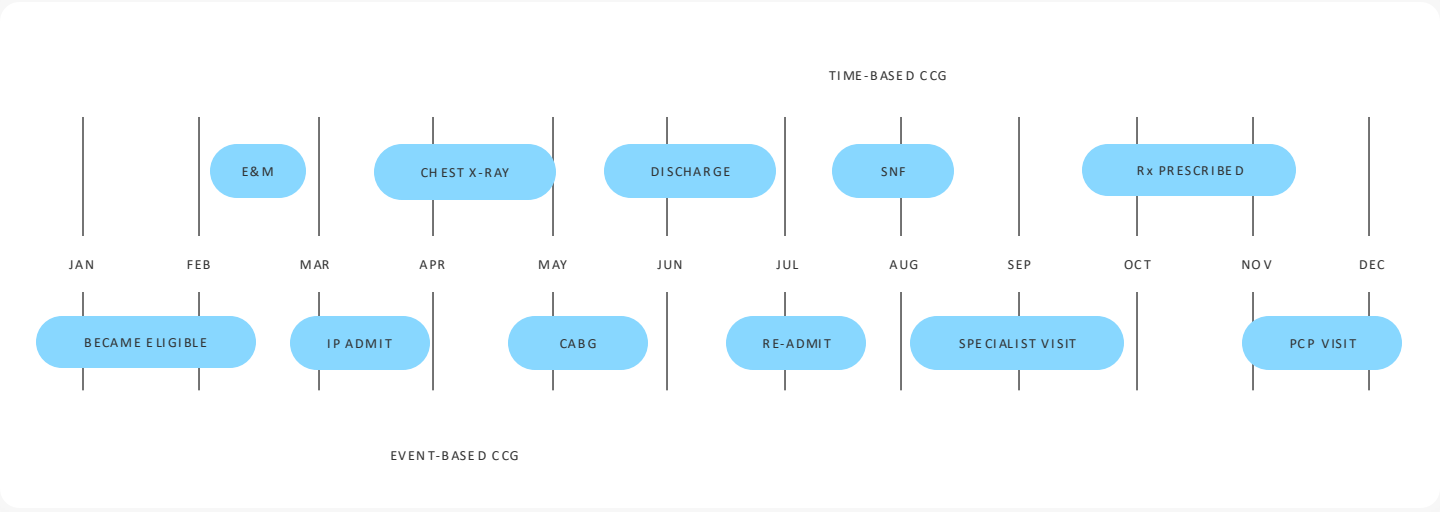
Data Grouping

We organize longitudinal patient data into clinically relevant standard units of analysis called Clarify Care Groupings (CCGs). This standardization allows us to evaluate and benchmark any patient and any type of care journey.

The example below shows a longitudinal patient journey over 12 months. An event-based CCG may start at an IP Admit and end at Discharge, whereas a time-based CCG may start in April and end in December. Other examples of time-based CCGs are a patient record in 2017, a period of 90 days after knee surgery, and so on. Examples of event- based CCGs are diabetes diagnosis, hip replacement, ED admission, and so on.

The CCGs add a layer of flexibility to how we can cut the data, such that we can build fit-for-purpose groupers that answer business questions to the most specific use cases. For example, CCGs can be structured to match specific programs like BPCI-A and Pathways to Success to test how new programs could work.

FIGURE 1. CLARIFY CARE GROUPINGS (CCGs) METHODOLOGY ALLOWS FOR EVENT-BASED AND TIME-BASED ANALYSIS



Data Analysis

Data Quality

The operated-on data at each stage of the data journey goes through rigorous checks to confirm that the data has not been corrupted or manipulated in ways that introduce unexpected behaviors to downstream stages. These data checks are achieved with several tools that allow users across all Platform teams to maintain and add to data tests.

Verification

When iterating over many clinically relevant groups and modeling observations models can quickly tally in the thousands and ensuring that sample sizes are adequate to support the type of configured models, pre-model verification is a vital step in the road to insight. We make sure we have sample sizes to support the configured algorithms and ensure that the algorithms applied represent the mathematical model suited for modeling the data. This happens before models have ever been trained and is also a verification check we make after modeling runs have been completed.

In addition to verifying that sample sizes meet a necessary threshold, we also perform a clinical verification step, which is arguably much more important. Our Data Science and Clinical Informatics teams work together to verify that the combinations of clinical groups, program models, specialties, observations, and features used in our modes make for clinically sound combinations. This not only ensures that we produce useful and responsible models but also that our automated modeling functions smoothly.

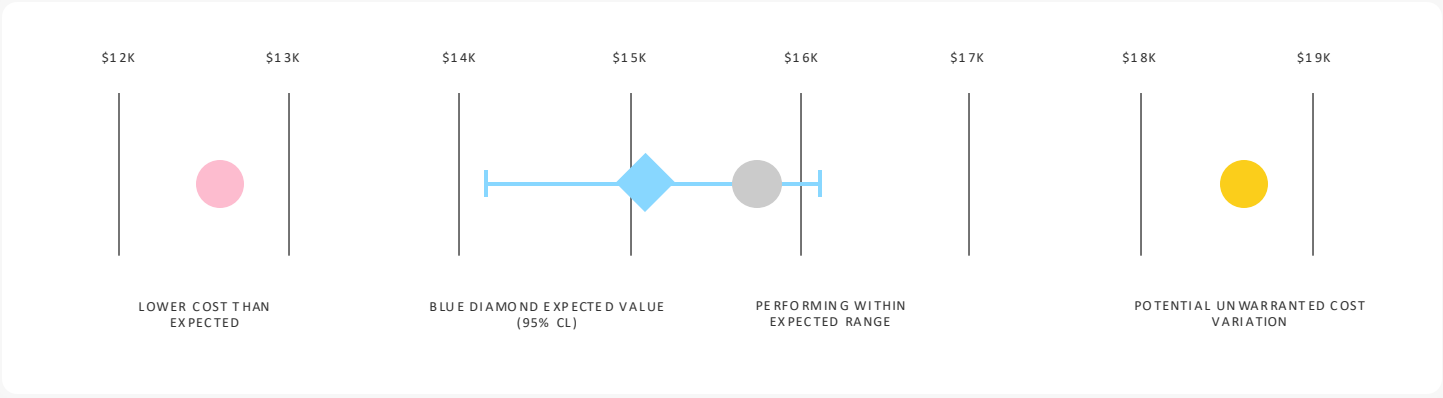
Exploratory Data Analysis

Exploratory data analysis is an essential activity for any cross-collaboration between data scientists, analysts, and clinical informaticists. By delving into the data and uncovering patterns, trends, and anomalies, these professionals can gain insights that guide decision-making and improve patient outcomes. At Clarify, we have developed tools and capabilities for exploratory data analysis that can rapidly transition research using data visualization, statistical analysis, and predictive analytics to automate reporting and product improvements.

Data Modeling

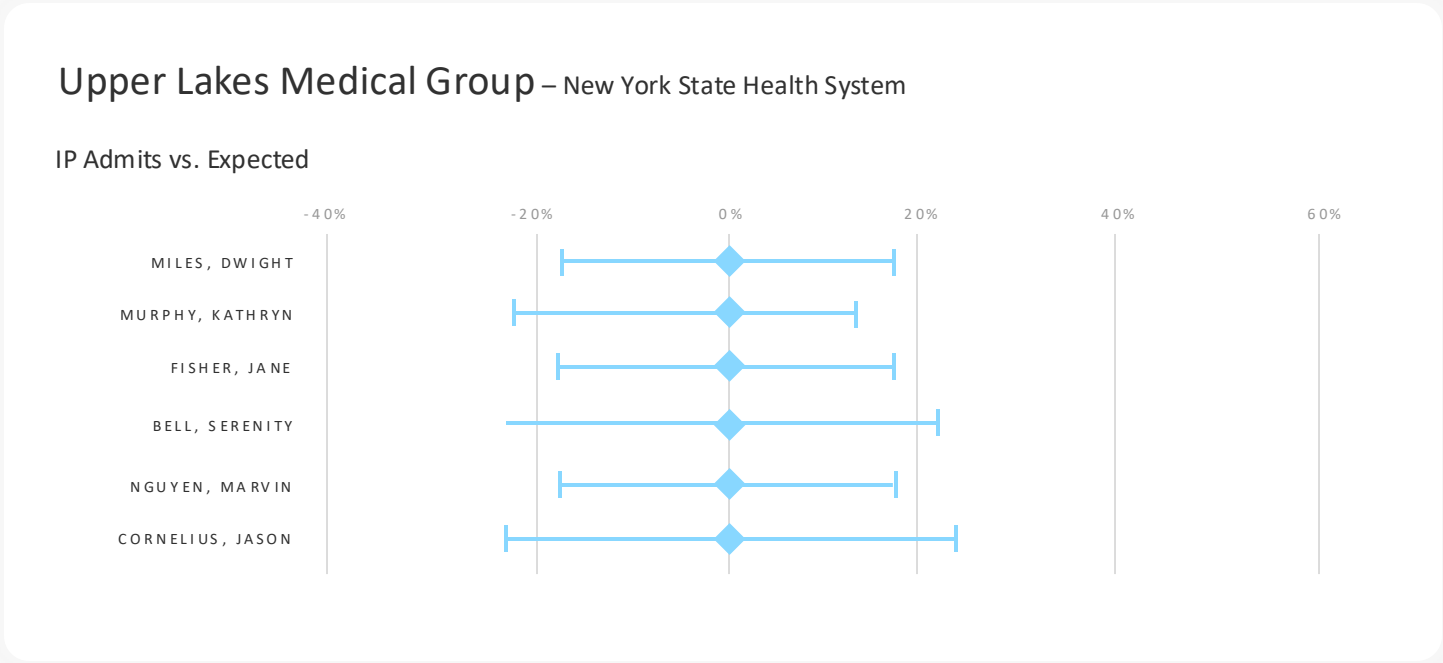
Clarify develops predictive models to (1) retrospectively compare observed patient care to predicted clinical and financial performance, and (2) prospectively predict patient risk and care utilization in real time at the point of care. We apply an efficient data modeling pipeline to generate accurate predictive models that yield benchmarks and values, called Clarify Blue Diamonds.

FIGURE 2. THE PLATFORM APPLIES PREDICTIVE ANALYTICS TO CREATE BLUE DIAMOND EXPECTED VALUES



The predicted value we generate effectively provides the most accurate case-mix adjustment, accounting for provider factors (e.g., hospital size, academic vs. non-academic), patient characteristics (e.g., comorbidities, DRGs, procedure types, age), and regional factors (e.g., rural versus urban) to simulate how an average provider would be expected to perform on that metric on the exact same set of patients. The figure below shows observed to expected performance on average PMPY spend across major PCP groups in California.

FIGURE 3. EXAMPLE OF OBSERVED AND EXPECTED AVERAGES, PROVIDING AN ESTIMATION OF WHERE DEVIATIONS BECOME SYSTEMIC



Clarify Health uses a variety of techniques to fit the best model

Predictive Analytics

Similar to grouping data before modeling, we also have a variety of configurations for our models designed to gather different insights. For retrospective models, we use features and outcomes from the same year (also known as concurrent setup). If we are building models on our national datasets, we train and predict on the same set. For external client data prediction, we train on ground- truth dataset and predict on client dataset. For prospective models, we use features from prior year to predict outcome from a given year. We use one year as the training set and the next year as the inference set to ensure that the model identifies year-over-year patterns and does not overfit to a particular year for internal dataset prediction. For external client data prediction, we use previous year data from ground-truth dataset to train a model and predict on the client dataset for a given year. We use a variety of techniques to fit the best model:

- Initial plotting of outcome distributions
- Pre-modeling data verification reporting
- Automated correlation matrix: Feature selection for exploratory data analysis (EDA) runs
- Automated descriptive statistics: QQ-plots, residual plots, R2, bias-average, ROC plots, precision-recall plots, max-F1, AUC-ROC (automated validation reports)
- Physician reviews of models and insights (both with Clarify team members and customers)

Model Evaluation

Clarify's prediction models are evaluated on three factors: responsibility, usefulness, and validity.

Responsibility

As an analytics insight generator to a decision support organization, the Data Science Team at Clarify recognizes the duty we have in handling and drawing conclusions with health information from millions of individuals. We seek to make our domestic healthcare system one that is well-informed via analytics insights and capable of achieving better outcomes for everyone.

In pursuit of these goals, Clarify decided early on in its development that the best approach to this was through explainable modeling. At its most basic level, a model is simply a tool that is trained to make a prediction given a common set of features (independent variables). This training is performed by giving an algorithm as many examples as possible so the algorithm can encode the functional patterns for mapping feature values to a prediction.

As a team, we:

No black box

At Clarify, we only use algorithms that produce summaries about their encoded rules and highlight input features and recommendations in easy-to-understand ways.

SELECT FEATURES

Carefully select the features for our models to ensure we're not biasing the model. For example, using the race demographic feature in comparative utilization models may be biased because some demographic groups utilize fewer resources. Or, if we don't account for social and behavioral factors, then we incorrectly predict that a person experiencing homelessness or poverty would need the same resources as someone else.

REVIEW METHODOLOGY

Critically review the methodology of each model. For example, when comparing the performance of a physician during a year, we account for changes in case mix during the year. In this way, we account for case mix changes during the performance year. We do not include performance period features that are cost proxies, which may bias the model and reward over-utilization.

ASSESS BIASES

Review the encoded rules and feature importance to ensure our models are not encoding systematic or implicit biases before putting those models into production. And once those models are in production, those feature importance are available to any end-user of our model results.

Usefulness

British statistician George E. P. Box is famously attributed with the quotation, “all models are wrong, but some are useful.” At its core, what this quotation is saying is that no model is capable of accurately describing every data point that went into it. Noise, informational limits, and purpose are all reasons models have loss (and are all wrong).

At Clarify, the goal of our models and data insights is to help customers make better decisions. We target the utility of our models in a few ways:

1. Identifying actionable vs. non-actionable features
2. Grouping actionable features
3. Identifying cohorts to act on
4. Providing clear contributions of every feature to the outcome
5. Enabling the user to do what-if analyses by changing the values of one of more features to see what would be the effect on the outcome

> “All models are wrong, but some are useful.”
>
> – GEORGE E. P. BOX

Validity

We test for validity across two dimensions:

1. HOW PREDICTIVE IS THIS MODEL?
2. DOES THE MODEL CAPTURE THE CORRECT PATTERN?

How predictive is the model?

We use industry-standard accuracy metrics:

1. REGRESSION

When predicting amounts or counts (continuous outcomes), we typically use R2, Adjusted R2, RMSE (root mean square error), MAE (mean absolute error), or MAPE (mean absolute percentage error). In addition to single description metrics, we examine Q-Q and residual plots to evaluate prediction values over the full range of each observation. The choice of the accuracy metric is dependent on the data and the use case:

a. If the goal of the model is to penalize outliers more, we would typically use RMSE; otherwise, MAE.

b. If the goal is to compare models with different outcomes that have different value ranges, we would use MAPE.

c. If the data is very heavily skewed (e.g., extreme zero inflation), then we would choose to use an adjusted R2, which would, instead of comparing to the mean model, compare to a model that always predicts the most common value.

2. CLASSIFICATION (MODEL EVALUATION)

When predicting the probability of a binary outcome, we tend to use AUC (AUROC) (Area under the receiver operating curve) or AUPRC (Area under the Precision- Recall curve). Like with regressions, we also examine plots of prediction values relative to observations. We do this by examining ROC and precision-recall (PR) curves.

a. If the data is not heavily unbalanced with respect to the binary outcome, then we typically use the AUROC.

b. If the data is heavily unbalanced with respect to the binary outcome, then the ROC will tend to be misleading. In these cases, we would tend to use AUPRC.

3. CLASSIFICATION (THRESHOLD EVALUATION)

When evaluating the choice of a threshold, we tend to use F1 score.

a. If we want to weight precision and recall differently, then we tend to use F0.5 or F2 scores or similar.

Does the model capture the correct pattern?

Models can be highly accurate but pick the wrong pattern. For example, a highly predictive mortality model may not list age as one of the top predictors. We would consider this an accurate model but not a correct model.

We review the feature contributions of every feature in our models and use domain knowledge and expert review to ensure that the high predictors and low predictors are correct. This review includes:

1. Analyzing feature contributions to check if the features the model found to be most predictive:

a. Are found in published papers

b. Can be validated by our domain and clinical experts

2. Reviewing models with physicians inside Clarify

3. Reviewing models with customers

Validation

Model validation at Clarify is a critical gate for models/inference results to pass in order to make their way into our applications. The validation metrics and thresholds we use are not unique to Clarify. We use the same statistical methodologies that are used by any qualified statistician or data scientist to evaluate the performance of their models. We use these industry-accepted validation metrics and set minimum thresholds to determine how performant the models are and select the most performant algorithms and hyperparameter combinations for a given target-dependent variable. If models perform so poorly that they are not much better than random chance (or worse), then we flag those models to be withdrawn from release candidacy.

Data Interpretation

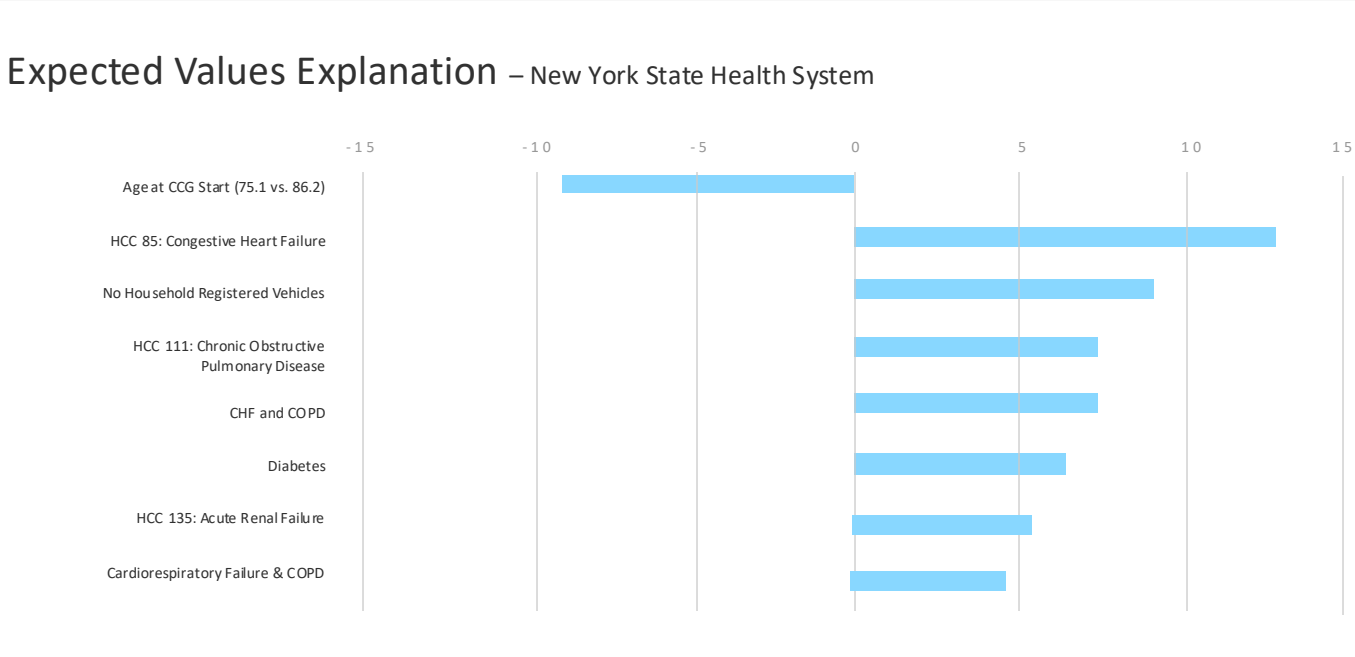
Prediction Explainer

The Clarify Prediction Explainer provides an additional level of detail to users, showing the contribution of each feature that drives the prediction. With this level of visibility, users can have more confidence in the prediction accuracy and evaluate how a potential course of action may impact an outcome.

More specifically, the Prediction Explainer enables users to:

- Evaluate the contribution of each feature to the prediction
- Compare the contribution of features between two sub-populations
- Conduct what-if analyses by changing the contribution value of each feature to see its effect on the outcome

FIGURE 4. PREDICTION EXPLAINER SHOWING THE TOP CONTRIBUTING FACTORS TO A PREDICTION: AN EXAMPLE PREDICTION OF ED VISITS FOR DIABETES



To further illustrate this point: if we use features X, Y, Z as inputs to the model M (X, Y, Z) to predict the value P, the prediction explainer provides the contributions of each of these features to the overall prediction: Xc + Yc + Zc = P. In the example shown above, users can assess the contribution of the feature 'Count of Comorbidities' on the prediction of 'ED Visits for Diabetes' in the population of 'physician Julia Kong's 301 patients.' Then, users can reduce 'Count of Comorbidities' by five to assess its impact on the 'ED Visit for Diabetes' in Doctor Kong's patient population. Additionally, this capability facilitates aggregation of contributions on temporal and population levels. Full transparency into the drivers of each prediction enables users to course correct.

Conclusion

Realizing the promise of big data in healthcare rests on the ability of data scientists, engineers, and domain experts to distill meaning from traditionally fractured, unconsumable data sources. We brought together experts across these domains to build the Clarify Atlas Platform in a way that transforms traditionally unactionable data into objective, clinically and financially meaningful business insights. The methodologies described in this paper showcase how we can uncover previously unrecognized insights about patients, clinicians, facilities, care networks, and therapies that are accurate and meaningful. It is with better market intelligence that payers, providers, and life science companies will truly move the needle in achieving their goals.

.

Clarify

About Clarify

Clarify Health Solutions® is a healthcare data and analytics company trusted by some of the most established organizations in healthcare, including providers, payers, tech and services, and life sciences. The Clarify Atlas Platform® is the foundation, leveraging the industry's largest and most robust dataset to map over 300M+ lives to deliver 20B+ AI-powered predictions to surface actionable insights with unparalleled speed and precision.

To learn more about how Clarify's data and analytics solutions can help your organization answer critical business questions about provider performance, visit clarifyhealth.com.

9/2024