MILLIMAN WHITE PAPER

Consumer data and its applications in health insurance underwriting

Michael Niemerg, FSA, MAAA
Joe Long, ASA, MAAA
Stephen Charlesworth
Meseret Woldeyes

Milliman

Executive summary

Advancements in machine learning and data mining have greatly increased the ease of generating, processing, and sharing insights from vast quantities of data. Companies in a variety of industries are using these sophisticated technologies in novel and entrepreneurial ways. Consumer data, for instance, has long been valuable to marketers but is now being analyzed for new applications within the insurance sector.

This white paper:

- Defines consumer data and provides examples of its application within and outside the insurance industry
- Presents the results of a study we conducted to understand the impact of incorporating consumer data in our Curv® product
- Provides background on the types and usage of data within health insurance underwriting
- Explains the methodology and results of the study

Ultimately, our study concludes that the addition of consumer data to other data commonly used in health insurance underwriting (e.g., age, gender, location, prescription history, medical claims data) yields limited incremental predictive value.

Consumer data defined

Consumer data, broadly defined, is data relating to an individual's purchase history and consumer preferences as well as any complementary attributes (e.g., demographic information) that can help to segment individuals into cohorts based upon likely preferences. The exact attributes of consumer data vary from vendor to vendor but typically include hundreds or even thousands of variables. The consumer data set analyzed in this paper includes variables such as: indicators of purchase history (house, car, etc.), consumer preferences (magazine subscriptions, categories of purchases, etc.), demographics (age, gender, marital status, dependents, ZIP Code, etc.), and basic credit attributes (income, mortgage size, payment history, etc.).

Consumer data includes a variety of attributes, so it may nominally overlap with other types of data such as social determinants of health, credit data, health data, or general demographics. For the purposes of this paper, these terms are defined as follows:

- **Social determinants of health:** "Conditions in the places where people live, learn, work, and play that affect a wide range of health risks and outcomes."[1][1]
- **Credit data:** Components of an individual's credit profile such as the number or age of accounts, credit limits, or average balances.
- **Health data:** Data collected in a clinical setting for either treatment or payment, such as diagnosis or procedure codes.

Consumer data remains distinct from these other sources of data because it focuses on consumer preferences and purchases while typically only encompassing elements of other types of data at a lower level of granularity. For instance, a credit profile may have hundreds of attributes related to lines of credit, account balances, etc., while consumer data may only have a few high-level credit-related variables such as outstanding mortgage balance.

Consumer data use cases

Consumer data has been used for years by marketers to help sell both consumer goods and insurance products to new and existing customers. More recently, with the advancements in machine learning algorithms and computing power, actuaries and data scientists in the insurance sector are actively exploring additional uses for this data.[2]

Though relatively new to group health insurance underwriting, consumer data already has accepted applications in other areas. For example, it is used in lead generation, customer segmentation, and targeted marketing in the retail and banking industries.

[1] CDC. Social Determinants of Health at CDC. Retrieved April 16, 2023, from https://www.cdc.gov/socialdeterminants/about.htmlhttps://www.cdc.gov/socialdeterminants/about.html.

[2] Gaweda, B., Krischanitz, C., Bellina, R. et al. (April 2022). Potential Data Sources for Life Insurance AI Modelling. Milliman Report. Retrieved April 16, 2023, from https://www.milliman.com/-/media/milliman/pdfs/2022-articles/4-22-22_lsc-data-science-report-ai-in-life-insurance.ashx.

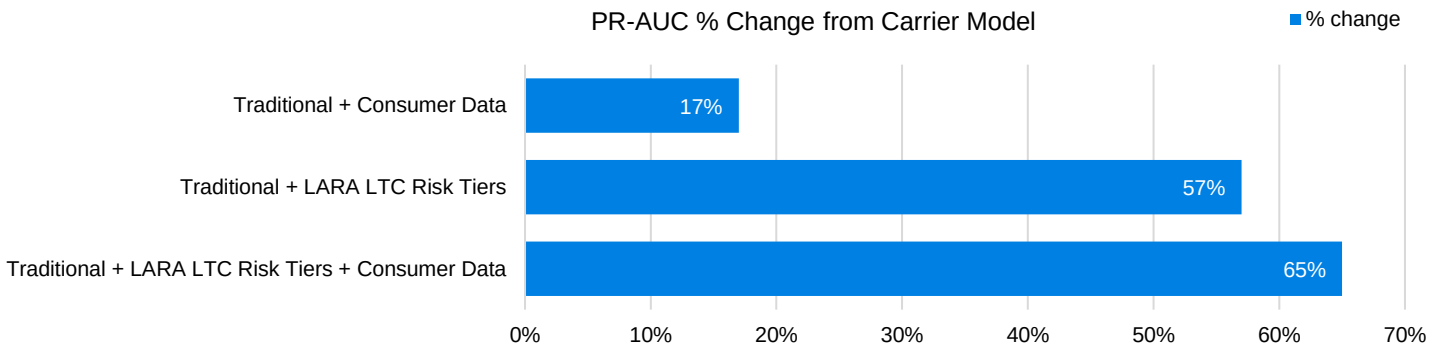
In the long-term care (LTC) insurance sector, consumer data can assist with prioritized outreach for wellness and intervention programs that aim to help LTC policyholders age in their homes longer. In this use case, consumer data has been shown to provide material predictive performance lift over traditional data collected and used by LTC insurance carriers for pricing and reserving. This is highlighted in the Milliman Long-term Care Advanced Risk Analytics (LARA™) predictive performance case study article.[3] For the purposes of this white paper, we have summarized the predictive performance of four separate models from the simulated pilot that predicts the probability an individual will have an LTC claim in the next year. The models that we developed in the analysis were tree-based gradient boosting machines (GBMs). GBMs were used to automatically perform feature selection, create feature interactions, and capture nonlinear relationships between the features and response variable. Our analysis of the performance of these models highlights the increased predictive power that is associated with the information extracted by the GBM from each data source.

Figure 1 summarizes the predictive performance change of these models relative to the baseline "traditional" model using the precision-recall area under curve (PR-AUC) metric. PR-AUC is a commonly used performance metric for imbalanced classification problems.[4] The traditional model contains traditional data collected and used by LTC insurance carriers for pricing and reserving. The "traditional + consumer data" model contains consumer data features in addition to the demographic variables in the traditional model. The "traditional + LARA LTC risk tiers" model includes LTC risk tiers that are calculated using a predictive model with deidentified medical and prescription histories as well as the attributes in the traditional model. Finally, the "traditional + consumer data + LARA LTC risk tiers" model combines all the elements—the carrier variables, the consumer data variables, and the LARA LTC risk tiers.

From this example we can clearly see that both the consumer data variables and the medical and prescription-based variables add meaningful lift when combined with the carrier variables in isolation. Additionally, when both variable sets are combined with the carrier data, their interactions provide the highest predictive performance.

Within our Life and Annuity Predictive Analytics (LAPA) team, two additional use cases involving consumer data have emerged. In one, annuity policyholders across carriers are clustered into segments via unsupervised clustering methods. These segments are the most distinct annuity policyholder personas (that is, individuals with similar behavior) and can be used to understand the specific needs of audiences more granularly than at the all-policyholders level. For annuity carriers, this means the ability to compare the performance, whether over-indexing or under-indexing, of current customer personas against an industry benchmark or index to better inform sales and marketing decisions. For instance, understanding that an organization underperforms on "affluent urban elite" policyholders may prompt the development of products to better serve that market or the launch of new marketing campaigns targeted toward this group.

FIGURE 1: PREDICTIVE PERFORMANCE IMPACT OF LARA RISK TIERS AND CONSUMER DATA OVER BASELINE MODEL



Note: % change = (AUC2/AUC1) – 1, where AUC1 = baseline of 9.8%. AUC2 = traditional + consumer data, traditional + LARA LTC risk tiers, or traditional + consumer data + LARA LTC risk tiers

[3] https://www.milliman.com/-/media/products/lara/superior-predictive-performance-of-milliman-lara-models.ashx

[4] Brownlee, J. (January 6, 2020). ROC Curves and Precision-Recall Curves for Imbalanced Classification. Machine Learning Mastery. Retrieved April 16, 2023, from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/.

LAPA's second use case is to satisfy an underserved need for annuity leads with a lead-generation service. As it stands, most annuity sales prospecting involving outreach is targeted with simple filters involving age and income. By building a supervised model to classify known annuity policyholders from non-policyholders using age, income, and the additional features in the consumer data set, we observed a substantial increase in PR-AUC over the status quo. While the final outcomes are still being studied in real-world experiments, this type of lead generation in other industries is well known to capture a larger and more interested audience than heuristic-based approaches.

Another example in which consumer data was successfully applied is seen in a study recently published in the *American Journal of Managed Care*, where publicly available and purchased social determinants of health data more accurately predicted inpatient and emergency department utilization without requiring clinical risk factors. The study result shows that air quality and income were more important to the decision-making capability of the model than age, ethnicity, or gender. Additionally, neighborhood in-migration, transportation, and purchasing channel preferences were more important than ethnicity or gender. Only three socioeconomic features of the 19 referenced in the study influenced the model's decision-making capability less than gender: retail access, employment sector, and percentage in group living quarters.[5]

Other potential insurance industry use cases involve marketing, mortality, lapsation prediction for life insurance, customer segmentation, and prioritization of outreach for care management. As more organizations explore the use of consumer data in the insurance industry, we expect to see additional use cases in which consumer data proves beneficial.

Uses of data in health insurance underwriting

Healthcare payers have a variety of tools and data sources available for use in rating, underwriting, and managing a health insurance book. For renewing populations, carriers typically use known claims history and demographic data such as age and geography to assess health status, risks, and care management opportunities associated with their memberships.

However, for new business blocks, claims history is often less detailed, if it is available to the payer at all. Instead, carriers may rely on questionnaires, industry benchmarks, and third-party de-identified data sources in models such as Curv to assess the health characteristics of a population or employer group. Deidentified data used in this manner is characterized by the fact that it can be obtained without a HIPAA authorization due to the anonymization of individuals within the group. This anonymization allows modelers using this data to know details about the medical history of individuals within the group—for instance whether there is an individual within the group that has cancer or diabetes—without knowing who the specific individual is, because personal identifiers do not exist within the data.

Emerging alternative data sources in health underwriting such as lab, behavioral health, and consumer data have attracted interest from vendors and healthcare payers as the industry gains access to more data and seeks to improve existing models or reduce costs associated with underwriting by using data sources that are cheaper to obtain. These new data sources seem promising, but their predictive accuracy is still being proven.

To test the potential improvement in predictive power when using consumer data in the context of new business underwriting for group health insurance, Milliman calibrated and analyzed the performance of models that predict healthcare claims cost using consumer data as model features in addition to de-identified prescription and medical data.

Study

OVERVIEW

To demonstrate the value of consumer data in predicting group healthcare claims, a rigorous statistical methodology is required. Healthcare claims are notoriously volatile, even for groups of individuals, and even small changes in predictive accuracy can have a sizeable financial impact for carriers. As described below, we carefully evaluated the data, developed predictive models, and employed a suite of validation measures to test the value proposition of consumer data.

DATA SET DESCRIPTION

To perform analyses on consumer data, we used a data set containing commercial group health insurance lives coming from a mix of carriers and geographies. For each one of the lives, consumer data as well as deidentified prescription and medical histories were collected. Overall, approximately 14 million individuals and 175,000 groups were represented in a mix of groups, with employee group sizes varying from five to 500.

[5] Chen, S. & Bergman, D. (January 15, 2020). Using Applied Machine Learning to Predict Healthcare Utilization Based on Socioeconomic Determinants of Care. American Journal of Managed Care. Retrieved April 16, 2023, from https://doi.org/10.37765/ajmc.2020.42142.

Allowed claims using experience from calendar year 2019 was used to generate the target variables. Consumer data features were generated using attributes generated from 2018 data; prescription and medical data features were generated based upon 2013-2018 data.

EXPLORATORY ANALYSIS

We performed data exploration after running basic validation and reasonability checks. Figure 2 shows a representative example of one of the data analyses. In this example, we graph an individual feature in the data set that measures financial risk tolerance against allowed claims per member per month (allowed PMPM), which measures the claim costs of individuals before accounting for member cost-sharing provisions like deductibles and coinsurance (that is, it is the total of patient liability and health plan liability). Financial risk tolerance is defined in the data dictionary as, "predicts the likelihood of an individual's risk tolerance for investments ranked into high (aggressive), moderate (moderately aggressive), and low (conservative)." In this encoding, 1 is low, 2 is moderate, and 3 is high. In some cases, data was unavailable, in which case it has the value "NA." In Figure 2, we can see that the median allowed PMPM (the dark line in the center of the white boxes) is roughly the same for each split. However, note how the third quartile (the upper bound of the white boxes) consistently shifts downward as the feature values increase, indicating different distributions of values and decreasing variance even though the medians are roughly the same.

This visualization only includes members who are female, aged 50 to 55, to isolate the analysis from any confounding effects from age and gender. Potential confounding from any other variables, such as geography, is still unaccounted for in this graphic. More generally, throughout this and all subsequent analyses, we paid careful attention to potential confounding with variables that would already be accounted for in a payer's estimates of costs, namely age, gender, and geography.

Figures 3, 4, and 5 illustrate the difference in distribution of this variable between males and females, and across age groups within the genders.

FIGURE 2: AGE 50-55 FEMALES – CLAIM DISTRIBUTION PARTITIONED BY THE VALUES OF THE CONSUMER DATA FEATURE INVESTMENT RISK TOLERANCE

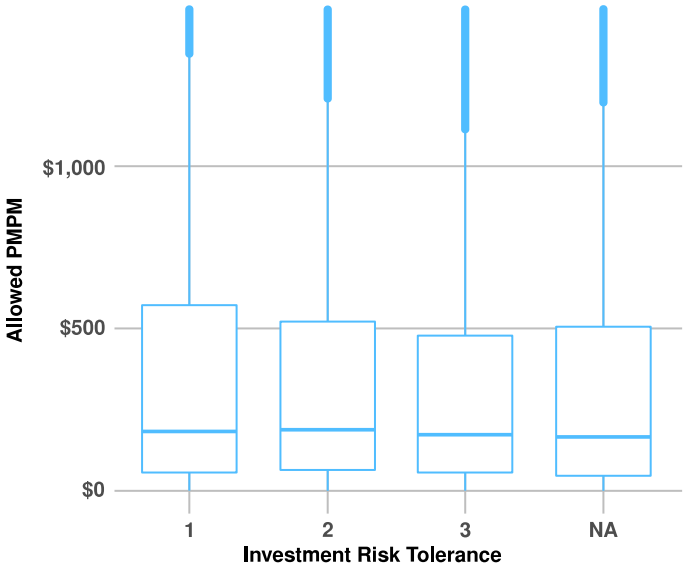


FIGURE 3: DISTRIBUTION OF THE INVESTMENT RISK TOLERANCE FEATURE BY GENDER

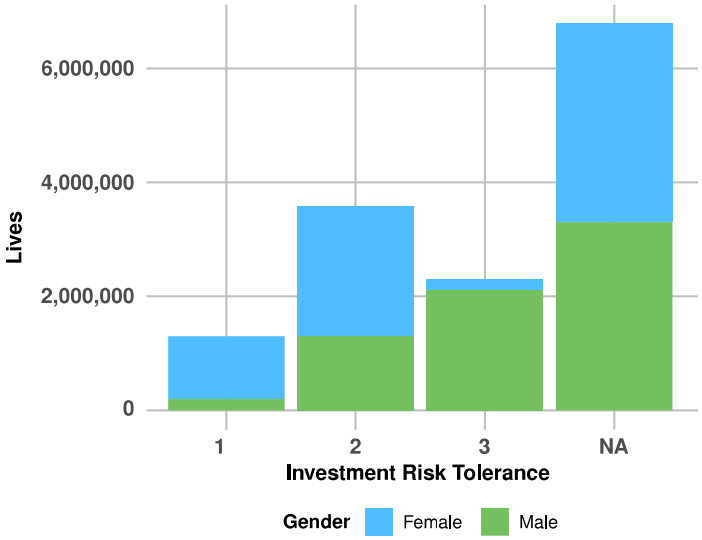


FIGURE 4: DISTRIBUTION OF THE INVESTMENT RISK TOLERANCE FEATURE FOR MALES, BY AGE GROUP

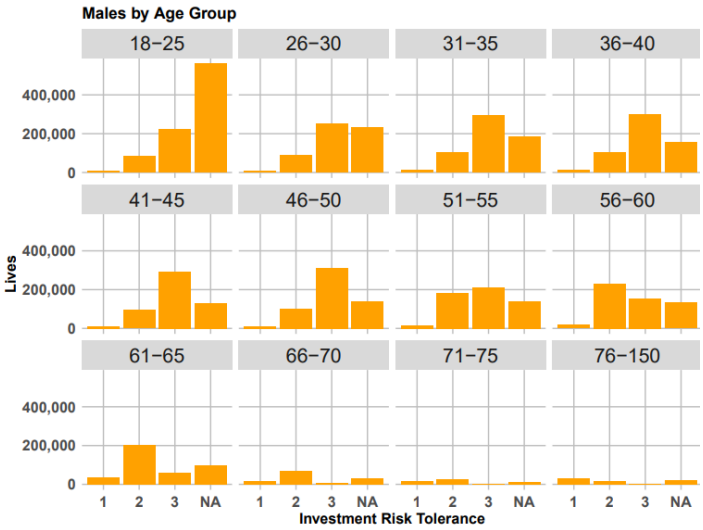
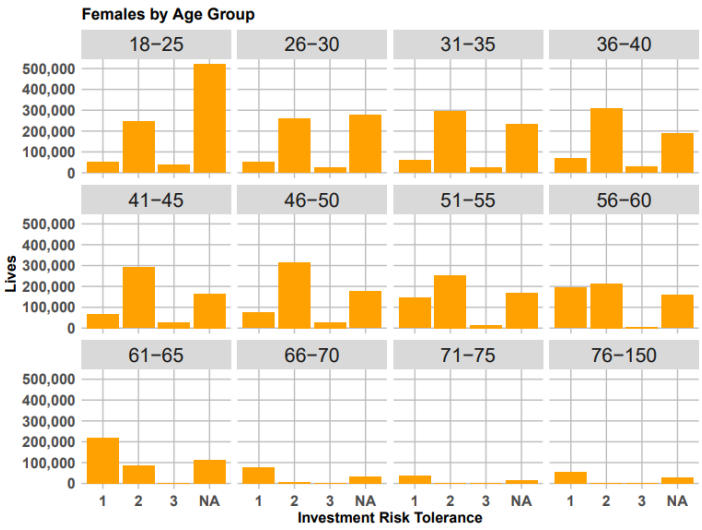


FIGURE 5: DISTRIBUTION OF THE INVESTMENT RISK TOLERANCE FEATURE FOR FEMALES, BY AGE GROUP



MODELING APPROACH

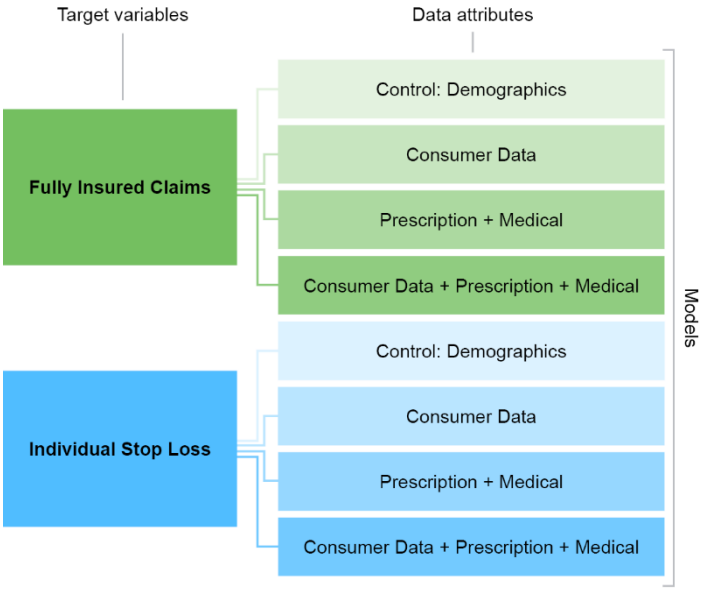
Models were developed using a prototypical machine learning framework. The validation approach employed in model-building was to split the data into a training set and a test set, perform cross-validation within the training data set during the model-building process, and use the test set as a final test of model performance. The 14 million lives in the analysis data set were split into approximately 11 million for training and 3 million for testing. The training and test partitions were developed using group membership so that all members of a particular group would be in either the training or the test set. Otherwise, the data sets were balanced across relevant metrics such as average group size and demographics. This method of data-splitting allowed us to measure both group-level and individual-level performance.

Four separate models were trained on each of two distinct target variables (for eight models total). The first target variable used was allowed PMPM, which was used as the target variable for fully insured business. The second target variable was a binary objective representing the occurrence of an individual member's annual allowed claims exceeding a threshold of $100,000 per year to represent the target variable on individual stop-loss. Throughout the course of the modeling effort, we developed 15 different permutations of feature sets for model training, using different combinations of features (for example, how census features and location were represented) to assess the effects on modeling results. The primary motivation throughout testing the different data representations was to ensure that we were extracting the maximum predictive power from the consumer data attributes.

The "control" model contains demographic features exclusively pertaining to age, gender, and member location. The "consumer data" model contains consumer data features in addition to the demographic variables in the control model. The "prescription + medical" model includes the prescription and medical data features as well as the attributes in the control model. Finally, the "consumer data + prescription + medical" model combines all the elements—the demographic variables, the consumer data variables, and the variables for the prescription and medical data features.

The control model establishes a modeling performance baseline from which the impact of supplemental data sources can be assessed. This model is meant to reflect the basic information that would be embedded within a group's manual rate structure. The consumer data model can be contrasted against the control model. It allows us to assess the impact of consumer data attributes relative to a manual rate. Comparing the "prescription + medical model" to the "consumer data + prescription + medical model" allows us to evaluate the impact of consumer data features in the context of the existing prescription and medical features available today in Curv.

FIGURE 6: MODEL INVENTORY



Target variables

- **Fully insured claims:** Numeric prediction of allowed PMPM claim costs for an individual.
- **Individual stop-loss:** Probabilistic binary classification of claims in excess of 100,000 for an individual.

Data attributes

- **Control model:** Demographics (age, gender, and member location); proxy for typical/traditional insurer manual rating approach in the absence of other data sources.
- **Consumer data:** Demographics plus consumer data attributes.
- **Prescription + medical:** Demographics plus deidentified prescription attributes plus deidentified medical attributes.
- **Consumer data + prescription + medical:** Demographics plus deidentified prescription attributes plus deidentified medical attributes plus consumer data attributes.

MODELING RESULTS

After the models were created, they were evaluated for predictive performance. No significant improvement was observed on key metrics, such as mean absolute error (MAE), root mean squared error (RMSE), or R^2 in either of the two contexts: control model versus consumer data or "prescription + medical" versus "consumer data + prescription + medical" as summarized in Figure 7.

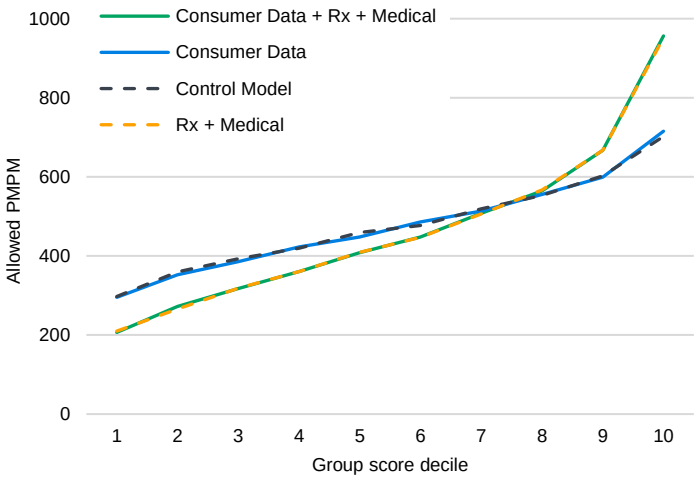
FIGURE 7: EVALUATION METRICS FOR THE MODELS AT THE INDIVIDUAL LEVEL

MODEL	MAE	RSME	R^2
CONTROL VS. CONSUMER DATA			
Control model	639.4	2,161.1	0.016
Consumer data	637.0	2,159.9	0.017
PRESCRIPTION + MEDICAL VS. PRESCRIPTION + MEDICAL + CONSUMER DATA			
Prescription + medical	528.1	1,993.8	0.163
Consumer data + prescription + medical	527.6	1,994.3	0.163

Figure 8 shows a lift chart that additionally affirms what the standard regression metrics demonstrated: that consumer data provides no practical differentiation in risk stratification. Thus, it is hard to distinguish the difference in the lift charts between the models that contain consumer data and those that do not.

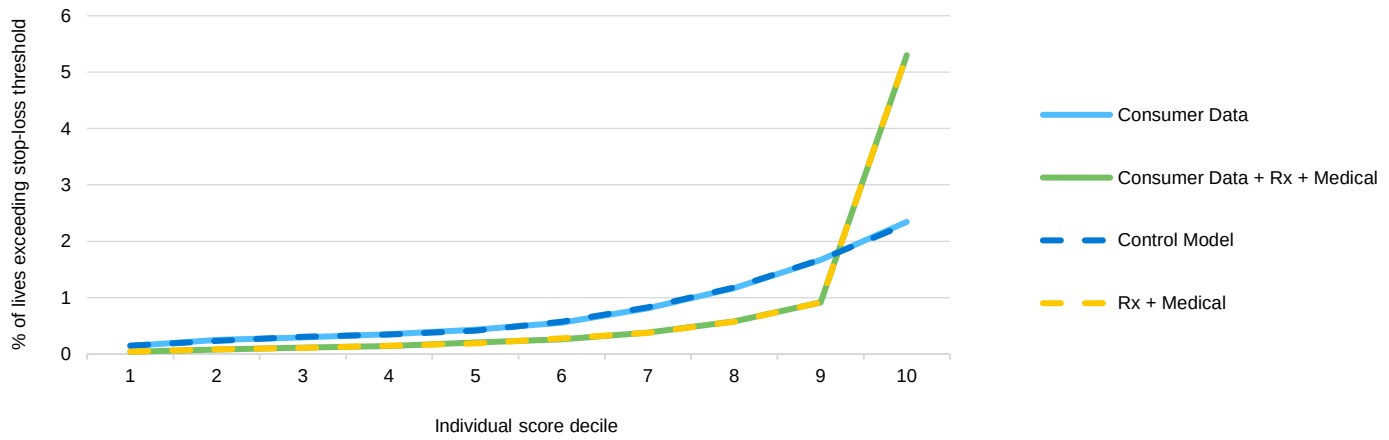
The consumer data model fails to significantly separate itself from the control model baseline. This suggests that it offers minimal value beyond the demographic and area factors currently in use. As is seen in Figure 8, the results are similar when a stop-loss objective is considered.

FIGURE 8: CLAIMS (ALLOWED PMPM) GROUP-LEVEL DECILE LIFT CHART



Predictions are binned into deciles based upon rank ordering (x-axis). For all groups falling within each decile the allowed PMPM value is calculated (y-axis). The better the separation from lowest to highest decile, the more the model is stratifying risk.

FIGURE 9: STOP-LOSS INDIVIDUAL-LEVEL DECILE LIFT CHART



As in Figure 8, this chart demonstrates that consumer data adds negligible predictive lift compared to a model that includes only demographics or compared to a model that includes demographics, prescription, and medical histories.

To gain additional insight into the inner workings of the models, the variables in the models were ranked by their feature importance to provide a relative measure of how much each variable contributed to the model predictions. For the "consumer data + prescription + medical models," 10 consumer data features that were not census elements broke into the top 100 features. Of these, one (Rx – number of drugs) was a lower-resolution version of the information used to build prescription and medical models. Aside from investment risk tolerance and "Rx – number of drugs," features breaking into the top 100 were related to facts about the individual's residence (ZIP Code home value, years since home was built, months since property was sold, land square footage, loan-to-value ratio), and a feature indicating the percentage of individuals without health insurance. As these features are related to an individual's financial situation and access to healthcare, careful consideration must be given to ensure using them does not reinforce existing disparities when the goal is to get an assessment of risk that is both accurate and fair.

Fairness considerations

As use cases for nontraditional data expand, one must consider the ways that using new types of data could impact individuals. While additional data may help improve the predictive performance of models, it may also further disadvantage or reinforce historical biases against particular groups, defined by race, ethnicity, sexual orientation, or other sensitive attributes. Therefore, before we put any of these models into use, we must test them for bias to ensure to the best of our ability that they are fair and equitable to all individuals. Any such analysis should be undertaken carefully and should involve definitions of fairness that directly correspond to the intended use of the model, as well as rigorous quantification of unfairness, deference to appropriate regulation, and professional judgment.

Depending on the use case, predictive performance may not be the most important criterion for determining which data sources should be used by the model. For example, the addition of new variables may not improve the overall predictive performance of a model; however, their inclusion could help address known bias issues for specific use cases. Regardless, sufficient testing must be performed to ensure the predictions from the model are fair and equitable to all individuals. Therefore, it is important to define the criteria for selecting data sources. There are also other important considerations one must evaluate when choosing data sources, which include but are not limited to legal, ethical, and operational cost implications.

Conclusions

Our results indicate that there is little additional lift to be gained by incorporating consumer data into group health underwriting when measured relative to existing data already commonly used for that purpose. Not every vendor of consumer data will offer the same attributes as the ones we tested, so results may vary by data set and vendor, but our conclusion was stark enough that we believe skepticism is warranted toward the use of consumer data in group health underwriting applications. We hypothesize that the addition of consumer data in our group health underwriting analysis was not beneficial in providing predictive performance lift because the important features that capture the variation in total healthcare costs were already accounted for in the medical, pharmacy, demographic, and area features.

While consumer data does not add unique value in this instance, that does not diminish its potential value in other applications within the insurance industry. Specifically, previously in this white paper we highlighted an LTC care management use case that showed consumer data in combination with medical and prescription-based variables adds meaningful predictive performance lift for models that are used to predict LTC claims. The addition of consumer data in the LTC care management use case allowed the model to gain more insight on the policyholders' living environments and financial stability, which were shown to be helpful at improving the performance of the models. Additional research could uncover other valuable uses such as the ability to more accurately predict specific healthcare events, medication nonadherence, or hospital readmission.

Milliman

Milliman is among the world's largest providers of actuarial, risk management, and technology solutions. Our consulting and advanced analytics capabilities encompass healthcare, property & casualty insurance, life insurance and financial services, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Michael Niemerg, FSA, MAAA
michael.niemerg@milliman.com

Joe Long, ASA, MAAA
joe.long@milliman.com

Stephen Charlesworth
stephen.charlesworth@milliman.com

Meseret Woldeyes
meseret.woldeyes@milliman.com