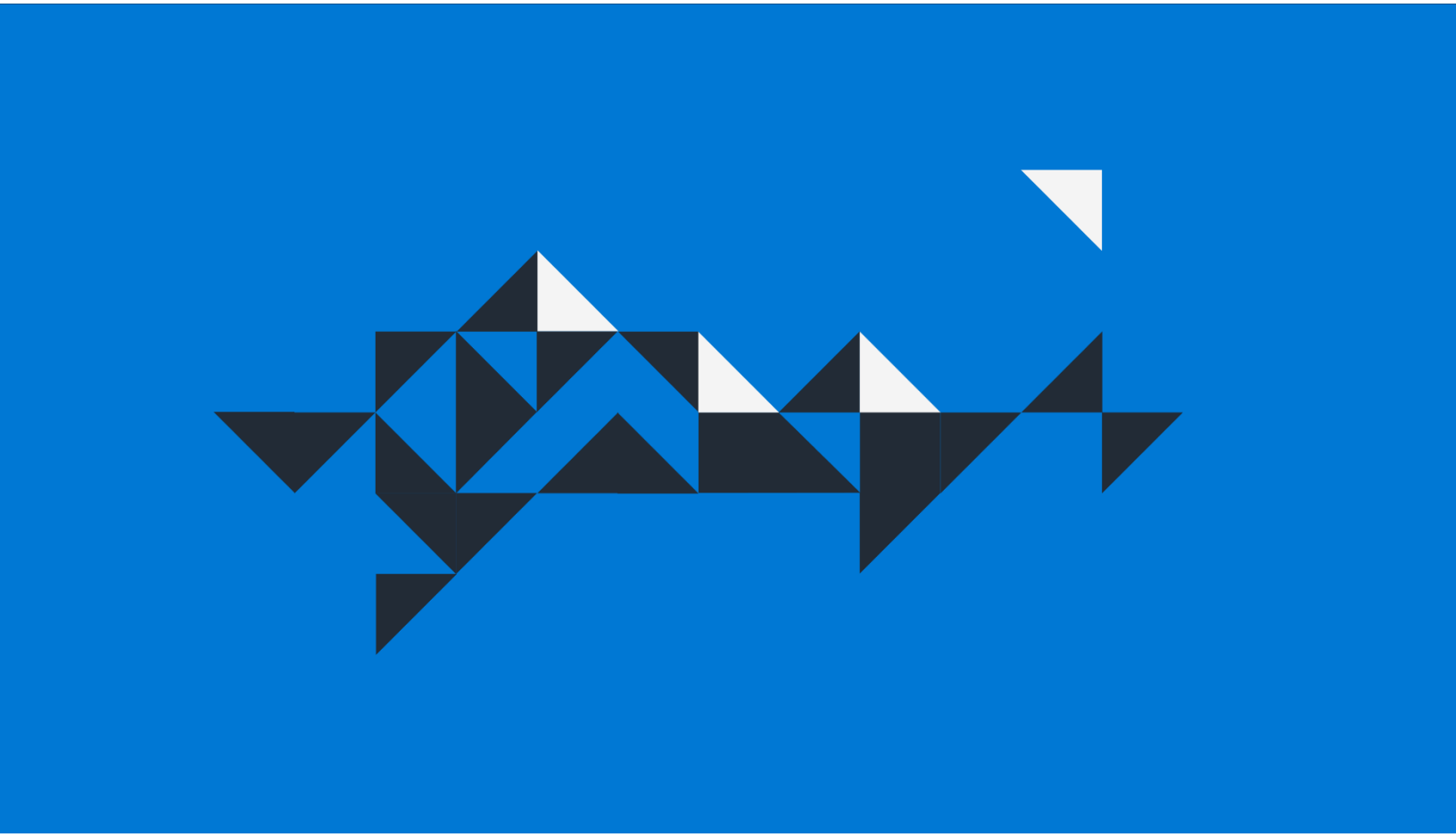
MILLIMAN REPORT

Analysis of Together Senior Health's RADAR algorithm to predict risk of undiagnosed dementia using claims data

Commissioned by Together Senior Health

October 2024

Spencer Marshall, FSA, MAAA
Brian Hartman, PhD, ASA
Christopher Kunkel, FSA, MAAA, PhD
June Chu, ASA MAAA



Milliman

Table of Contents

Executive Summary

This report is intended to provide a statistical analysis of Together Senior Health's (Together's) Risk of Alzheimer's and Dementia AlgoRithm (RADAR) family of models for predicting an individual's risk of having undiagnosed dementia[1]. RADAR is a proprietary cognitive-impairment identification and stratification model that builds on the previously published and validated electronic health record (EHR) Risk of Alzheimer's and Dementia Assessment Rule (eRADAR)[2][3]. Together engaged Milliman, Inc. (Milliman) to evaluate RADAR using administrative claims data and to test the impact of various modifications including recalculating the model weights and evaluating the marginal efficacy of including new prediction variables (predictors). Together is also developing and testing versions of RADAR that include claims, EHR, and other data sources that are not included in this report.

This report summarizes the findings from Milliman's engagement and is intended to provide feedback on the accuracy of predicting the risk of undiagnosed dementia using the RADAR models. We are only commenting on the specific results of our analysis. Actual dementia diagnoses may differ from predictions made by RADAR regardless of the source of the data used to calibrate the model. Results for any particular prediction will be unique to the characteristics of the input data, point in time, and other factors not considered in this analysis. This analysis utilizes administrative claims data and beneficiary data from the Centers for Medicare and Medicaid Services (CMS) Limited Data Set 5% Sample to identify the presence of variables used. Different input data may produce different results. This information does not constitute an endorsement of Together Senior Health or their RADAR model, nor does it quantify any financial value of dementia diagnosis predictions.

Any reader of this report must possess a certain level of expertise in areas relevant to this analysis to evaluate the significance and reasonability of the assumptions and the effect of these assumptions on the results. We recommend that all parties be aided by their own actuary, statistician, or other qualified professional when reviewing this report.

The analysis was performed as three distinct tasks:

1. **Create cohort datasets**
   - We created summaries of the CMS 5% Sample data by beneficiary with predictor flags. We also adapted EHR-based predictors for claims-based data and created additional predictor flags for consideration.
   - Each data set covers two consecutive years of historical data as model input and the subsequent third year as output for evaluation. For example, 2010 and 2011 historical data was used to develop predictors for outcomes assessed in 2012.
   - Predictor flags are based on beneficiary demographic information and reported International Classification of Diseases, 9th or 10th revision, Clinical Modification (ICD-9-CM or ICD-10-CM) diagnoses on non-laboratory claims, or by Healthcare Procedural Coding System (HCPCS) codes for specific services. The medical codes for each predictor were chosen to be consistent with the corresponding eRADAR predictor definitions. We did not perform a separate clinical evaluation of each definition.
2. **Run the models and calculate summary statistics**
   - The original RADAR model applies the model weights from eRADAR to the new claims-based predictors.
   - The Refit RADAR model is based on a reweighting of the predictors using the created cohort datasets.

[1] Incident dementia diagnoses in a 12-month outcome period are used as a proxy for undiagnosed dementia. The authors hypothesized that undiagnosed dementia was likely to be present at the beginning of the calendar year of the diagnosis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9904522/. We use the same language throughout this report.

[2] Barnes DE, Zhou J, Walker RL, Larson EB, Lee SJ, Boscardin WJ, Marcum ZA, Dublin S. Development and Validation of eRADAR: A Tool Using EHR Data to Detect Unrecognized Dementia. J Am Geriatr Soc. 2020 Jan;68(1):103-111. doi: 10.1111/jgs.16182. Epub 2019 Oct 14. PMID: 31612463; PMCID: PMC7094818. https://pubmed.ncbi.nlm.nih.gov/31612463/

[3] Coley RY, Smith JJ, Karliner L, Idu AE, Lee SJ, Fuller S, Lam R, Barnes DE, Dublin S. External Validation of the eRADAR Risk Score for Detecting Undiagnosed Dementia in Two Real-World Healthcare Systems. J Gen Intern Med. 2023 Feb;38(2):351-360. doi: 10.1007/s11606-022-07736-6. Epub 2022 Jul 29. PMID: 35906516; PMCID: PMC9904522.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9904522/

   - We compared C statistics[4] between the original and refit RADAR models.
3. **Evaluate new predictors**
   - We augmented the original RADAR model with a new set of beneficiary-level predictors identified by Together researchers and based on claims data.
   - We tested model performance when using these additional prediction variables.
   - We analyzed how often various predictors were included in the refit models and which were best at predicting the risk of undiagnosed dementia using C statistics.

KEY FINDINGS

- When using claim-based predictors, accuracy for predicting the risk of undiagnosed dementia is significantly improved by refitting the model to reweight the predictors.
- Using more potential predictors and allowing more flexibility in which predictors are selected results in statistically significant improvements. Whether the different models are practically significant (i.e., produce meaningfully different predictions of the risk of undiagnosed dementia) will depend on the application and dataset.
- When considering additional predictors, five of the new beneficiary-level variables were often included in the refit models.
- Model accuracy of predicting the risk of undiagnosed dementia varied across sex and self-reported race and ethnicity.

Data Preparation

The cohort datasets were created using the CMS 5% Sample for each calendar year from 2010-2021. Each cohort dataset uses data for three consecutive years, using two years to generate predictor flags and the final year as the outcome year. For instance, the 2012 outcome year is based on data from the 2010 and 2011 predictor years only. We restricted the cohorts to beneficiaries who were 65 and older and were enrolled in Medicare fee-for-service for the full 3-year period. We included over 8 million eligible beneficiaries (see Beneficiary eligibility criteria) and excluded approximately 1.6 million beneficiaries for not satisfying one or more criteria. For each predictor year (2010-2020), we included predictors using diagnoses and other medical codes for each beneficiary. For each outcome year (2012-2021), we classified each eligible beneficiary as receiving or not receiving an incident dementia diagnosis, which was used as a proxy for having undiagnosed dementia at the beginning of the outcome year.

BENEFICIARY ELIGIBILITY CRITERIA

Beneficiaries were included for a given period if they met the following criteria:

1. Had at least two claims from a primary care physician (PCP), specialist, urgent care, or emergency department visit in the predictor period.
2. Had no hospice care or dementia diagnosis in the predictor period.

HIGH-LEVEL DATA PROCESSING STEPS

1. We restricted eligibility records to exclude beneficiaries who did not meet the beneficiary eligibility criteria defined above.
2. For these beneficiaries, we analyzed their claims experience[5] to flag predictors and dementia diagnosis outcomes. Laboratory claims were excluded when evaluating predictors and outcomes.
3. We produced tables at the beneficiary level with corresponding predictor and outcome flags.

[4] Also known as the area under the receiver-operating characteristic (ROC) curve, the C statistic is a coarse measure of the ability of the model to differentiate between beneficiaries with a dementia diagnosis and those without.

[5] Prescription drug data is not available in the CMS 5% Sample data used. Predictors originally relying on prescription drugs were configured to rely on diagnoses and procedure codes using available sources.

Summary of Results

For Tasks 2 and 3, we fit six different models and compared how well they predicted the risk of undiagnosed dementia in the outcome year (i.e., the likelihood that a beneficiary was diagnosed with dementia in the outcome year), using the outcome flags in our data summaries. For example, to predict the probability of a member developing dementia in 2015, we:

1. Fit each model using predictors from 2012 and 2013 with a 2014 outcome year
2. Used those parameters with the predictors from 2013 and 2014 to predict the probability of developing dementia in 2015.

After calculating the probabilities of diagnosis, we compared those probabilities to the actual experience data to determine the models' accuracy. We compared model fit using C statistics, or area under the ROC curve[6]. We separately calculated the C statistics by self-reported race/ethnicity, and sex to check if model performance varied based on demographic characteristics.

To understand which variables improve the dementia diagnosis predictions, we fit least absolute shrinkage and selection operator (LASSO) models to each of the cohorts. For each of the LASSO models, we developed two versions: a "min" version and a "1se" version. The min version uses the lambda[7] value which provides the smallest cross-validated prediction error. The 1se version uses the largest lambda value whose error is within one standard deviation of the minimum error. The 1se models are more parsimonious potentially without losing too much predictive accuracy.

In this analysis we compared the following six models:

- RADAR – The published parameter estimates from the original eRADAR model[8] applied to the redefined predictors based on the CMS 5% Sample administrative claims data.

The following five models are variations of the original RADAR, refitting the results to the claims data source and modifying the list of included predictors:

- Refit RADAR – The RADAR model refit using a logistic regression to the redefined predictors.
- LASSO Min – A LASSO min model fit on the redefined predictors and the new beneficiary-level predictors.
- LASSO 1se – A LASSO 1se model fit on the redefined predictors and the new beneficiary-level predictors.
- Restricted LASSO Min – A LASSO min model that required all the redefined predictors to be in the final model with variable selection performed only on the new beneficiary-level predictors.
- Restricted LASSO 1se – A LASSO 1se model that required all the redefined predictors to be in the final model with variable selection performed only on the new beneficiary-level predictors.

For each of the LASSO models, we analyzed and summarized how often each predictor was included in the eventual model. Additionally, we analyzed and compared the C statistics for all the models.

C STATISTICS

We compared the predictive abilities of the models using the C statistic. There are six models for each cohort analyzed. The following general patterns emerged.

- Compared to using the weights from the original eRADAR model, accuracy for predicting the risk of undiagnosed dementia is statistically significantly improved by refitting the model to reweight the predictors.
- Using more potential predictors and allowing more flexibility in which predictors are included in the model results in statistically significant prediction improvements. Whether the results are practically significant (i.e. produce

[6] Hilden, Jørgen. "The area under the ROC curve and its competitors." Medical Decision Making 11.2 (1991): 95-101.

[7] The lambda parameter determines how large a penalty to apply to the coefficients. A larger lambda means that the model will be simpler, a smaller value allows the model to be more complex.

[8] https://pubmed.ncbi.nlm.nih.gov/31612463/

meaningfully different predictions of the risk of undiagnosed dementia) will depend on the application and dataset.

- Looking at self-reported race and ethnicity, all the models are less effective at predicting the risk of undiagnosed dementia for North American Native beneficiaries, and slightly less effective for Black beneficiaries, but show similar effectiveness for other races/ethnicities.
- Comparing reported sex, all models are more effective at predicting the risk of undiagnosed dementia for females than males.

Figure 1 shows the C statistics and 95% confidence intervals for each model separated by sex.

FIGURE 1: C STATISTICS AND 95% CONFIDENCE INTERVALS BY MODEL AND SEX (2021)

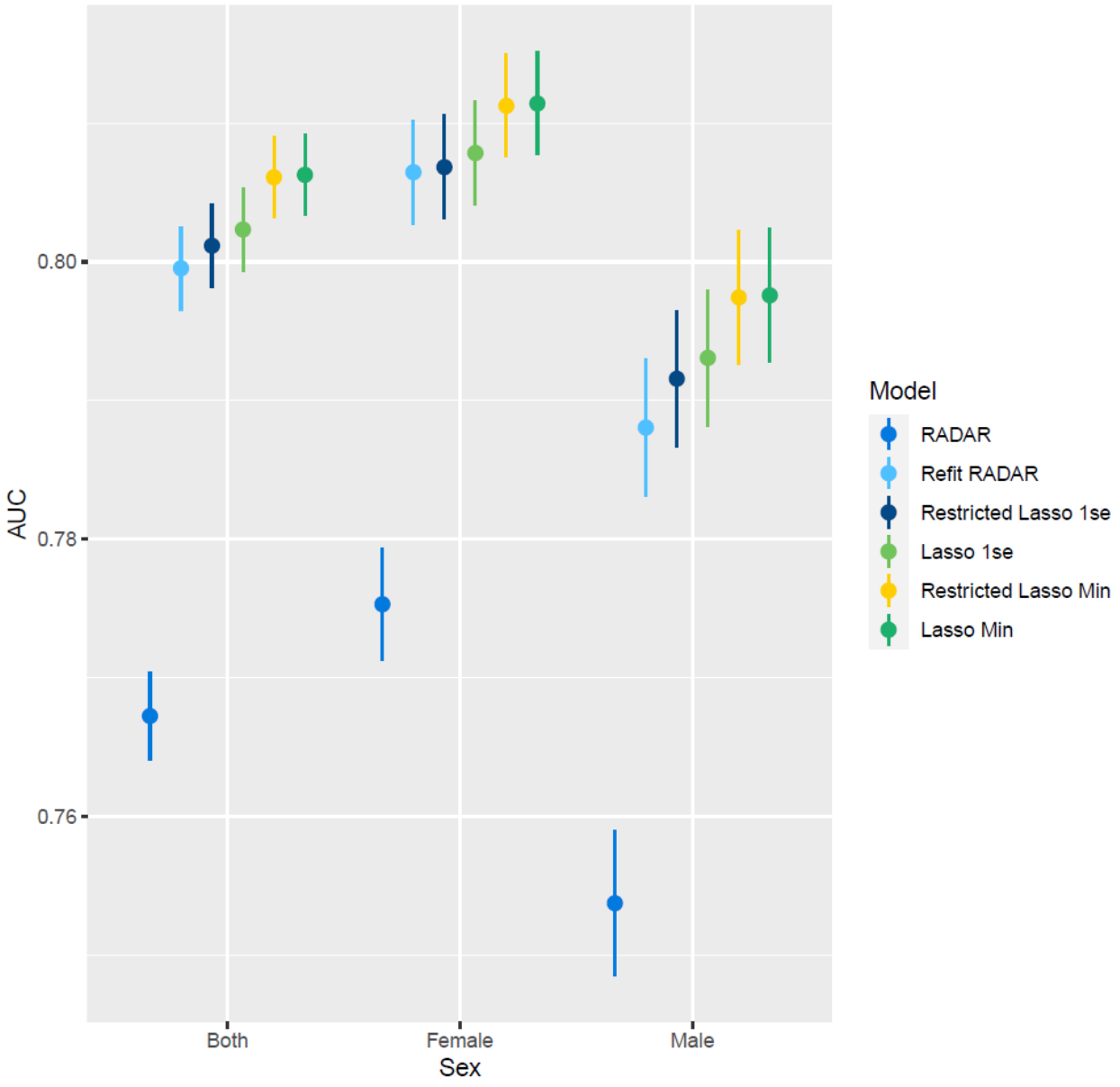
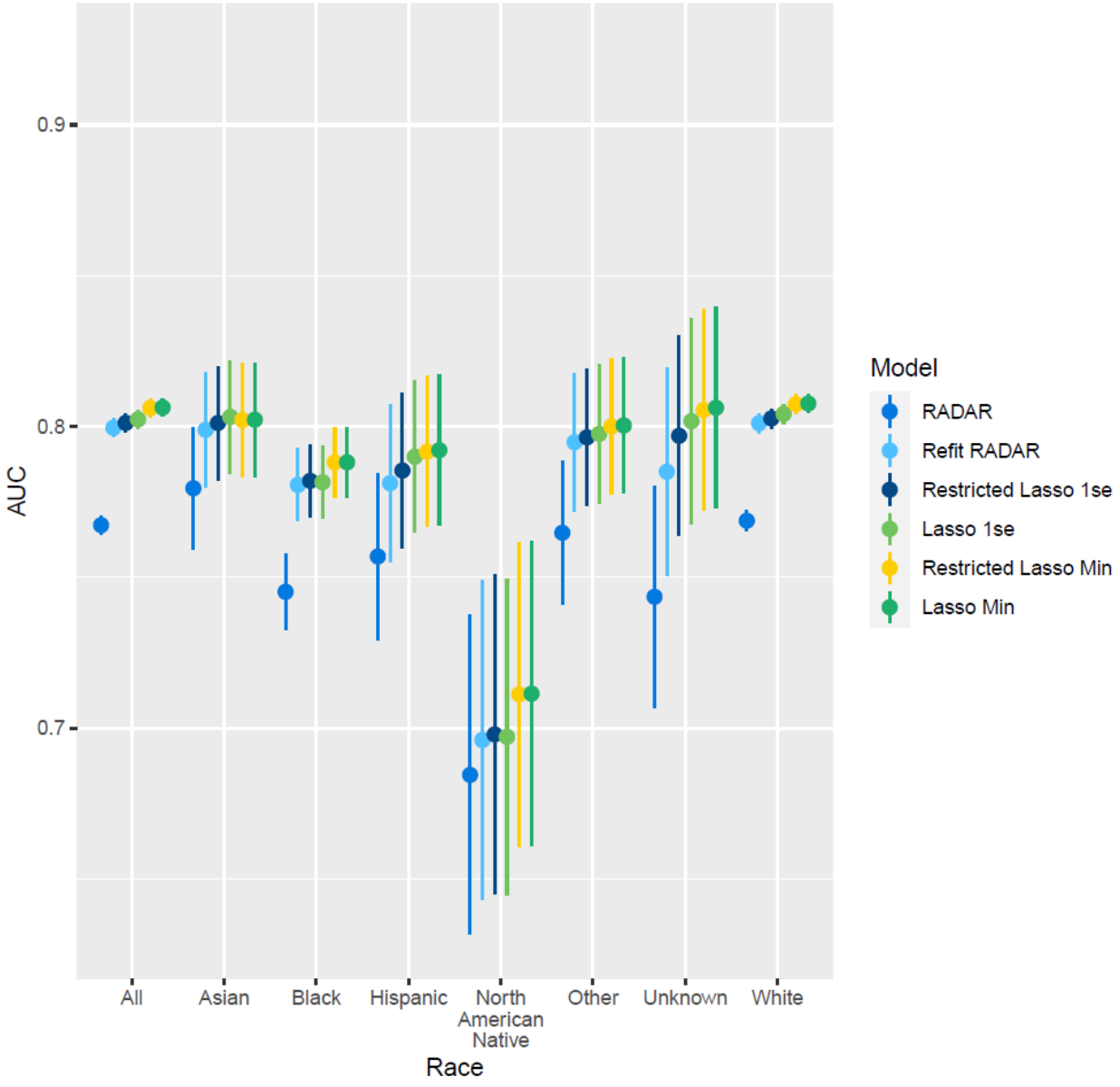


Figure 2 shows the C statistics and 95% confidence intervals for each model separated by race/ethnicity.

FIGURE 2: C STATISTICS AND 95% CONFIDENCE INTERVALS BY MODEL AND RACE/ETHNICITY (2021)

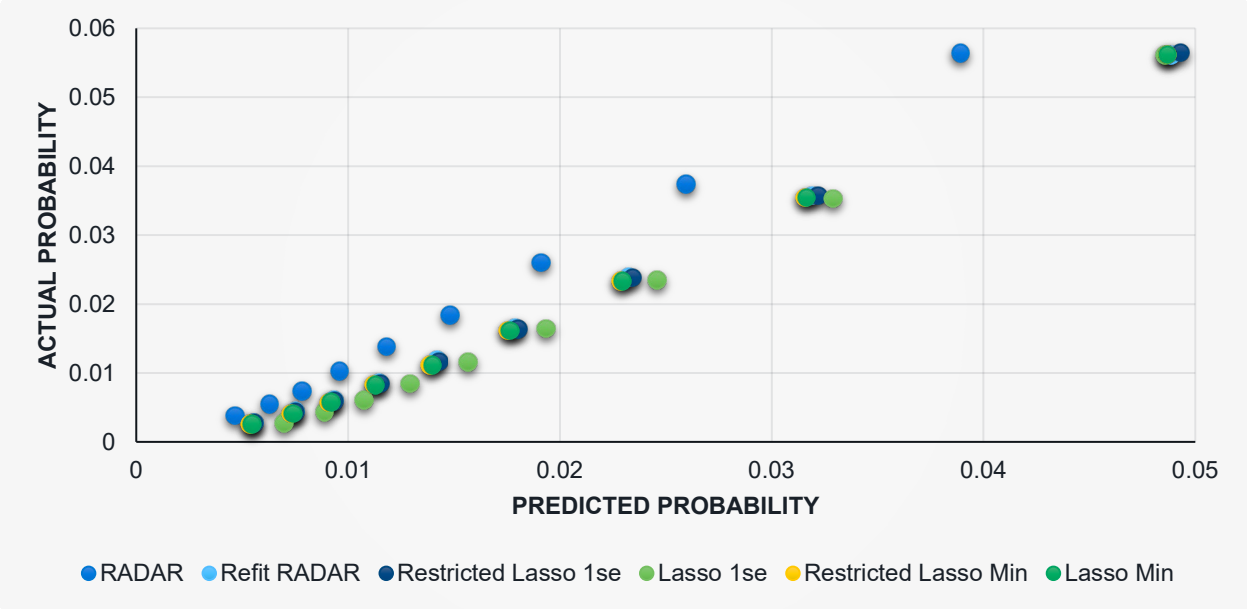


DECILE-BASED STATISTICS

We also examined the model fit by sorting the beneficiaries by predicted value and then divided them up into deciles to examine the average predicted probability in each decile and compare it to the actual proportion of beneficiaries in each decile who are diagnosed in the future year. All the models perform reasonably well by this metric.

Figure 3 compares the actual and predicted probability of a dementia diagnosis by model.

FIGURE 3: AVERAGE PREDICTED PROBABILITY BY DECILE



THRESHOLD-BASED STATISTICS

For each of the models, we compared how they would perform in a real setting by choosing a threshold above which we would say the beneficiary is likely to be diagnosed with dementia next year. Given that threshold, we measured how accurately we would have correctly predicted those who would be diagnosed with dementia next year.

	DIAGNOSED NEXT YEAR	NOT DIAGNOSED NEXT YEAR	
High predicted score	A (true positive)	B (false positive)	PPV = A / (A + B)
Low predicted score	C (false negative)	D (true negative)	NPV = D / (C + D)
	Sensitivity = A / (A + C)	Specificity = D / (B + D)	

Sensitivity

Sensitivity is a measure of how well the model identified beneficiaries who received a dementia diagnosis in the outcome year. Lower thresholds will generally have higher sensitivities because more beneficiaries will have a sufficiently high predicted score. Comparing the models, the original RADAR has the lowest sensitivity, with the other models having similar sensitivities to each other. For example, using a threshold of 90% (i.e., individuals with scores in the top 10% are classified as having a high likelihood of undiagnosed dementia), sensitivity was 0.37 for RADAR and ranged from 0.42 to 0.43 for the other models (Table 1). This can be interpreted as meaning that, among those who will be diagnosed with dementia in the next year, 37% to 43% would have been identified as high risk.

Specificity

Specificity measures how many false negatives are avoided. It is the proportion of all negatives which are true negatives. Higher thresholds will generally have higher specificity. In our example, each of the models performed essentially identically with specificity of 91% using the 90% threshold (Table 1). This can be interpreted as meaning that, among those who will not be diagnosed with dementia in the next year, 91% will be identified as low risk.

Positive Predictive Value (PPV)

The PPV measures the likelihood that a beneficiary who was predicted to have a dementia diagnosis in the outcome year actually did. This measure will generally increase as the threshold increases. Using this statistic, most models perform similarly with a slightly worse performance for RADAR (0.10) compared to the other models (0.12) using the 90% threshold (Table 1).

Negative Predictive Value (NPV)

The NPV measures the likelihood that a beneficiary who was predicted to have no dementia diagnosis in the outcome year actually did not. Lower thresholds generally improve this metric. Each model performed very similarly in this metric with very high NPV (Table 1).

Table 1 summarizes these threshold-based statistics for the 90% quantile threshold by model.

TABLE 1: THRESHOLD STATISTICS FOR THE 90% QUANTILE THRESHOLD BY MODEL

	90% QUANTILE THRESHOLD			
	SENSITIVITY	SPECIFICITY	PPV	NPV
RADAR	0.37	0.91	0.10	0.98
REFIT RADAR	0.42	0.91	0.12	0.98
LASSO MIN	0.43	0.91	0.12	0.98
LASSO 1SE	0.42	0.91	0.12	0.98
RESTRICTED LASSO MIN	0.43	0.91	0.12	0.98
RESTRICTED LASSO 1SE	0.42	0.91	0.12	0.98

Data Sources

The CMS Medicare 5% Sample contains claims and beneficiary data for 5% of the Medicare fee-for-service population. Claim types include inpatient facility, outpatient facility, home health agency, hospice, skilled nursing facility, durable medical equipment, and Part B claims. For cleaning and summarizing the CMS Medicare 5% Sample, the Milliman Health Cost Guidelines (HCG) Grouper[9] was used to assign service categories to all claim lines.

Predictor definitions for RADAR were provided by Together. In some cases, such as prescription drugs, the CMS 5% Sample data could not be used and were adapted to rely on available information using definitions from other sources including CMS, the National Institute of Health, clinical review, and others.

Caveats, Limitations, and Considerations

Spencer Marshall, FSA, MAAA, Christopher Kunkel, FSA, MAAA, PhD, and June Chu, ASA, MAAA are members of the American Academy of Actuaries and meet the qualification standards for performing the analyses in this report. To the best of our knowledge and belief, this report is complete and accurate and has been prepared in accordance with generally recognized and accepted actuarial principles and practices.

This report is intended to provide feedback on the accuracy of the RADAR family of models when using administrative claims data and the impact on accuracy when recalculating model weights and including new predictors. The results may not be appropriate and are not intended to be used for other purposes. In particular, actual dementia diagnoses for specific individuals may differ from predictions made by any of the models described in this report, regardless of the source of input data, choice of model or model calibration.

If distributed to third parties, the report must be shared in its entirety. We do not intend for this information to benefit or create a legal liability to any third party, even if we permit the distribution of our work product to such third party. Those reviewing this report should take full responsibility for interpreting the results, which should be reviewed by someone knowledgeable in areas of healthcare data and statistical modeling.

In performing our analysis, we relied on information provided to us by Together and other publicly available data sources, which we reviewed for reasonableness, but accepted without audit. We have relied on Together Senior Health for the definitions of the claims-based predictors in the models described in this report; we have not performed a separate clinical review of the appropriateness of the predictors. If the underlying data or information is inaccurate or incomplete, the results of our analysis may likewise be inaccurate or incomplete.

Models used in the preparation of our analysis were applied consistently with their intended use. We have reviewed the models, including their inputs, calculations, and outputs for consistency, reasonableness, and appropriateness to the intended purpose and in compliance with generally accepted actuarial practice and relevant actuarial standards of practice (ASOP). The models, including all input, calculations, and output may not be appropriate for any other purpose.

This report is based on historical data and may not reflect current and future relationships. Emerging experience should be monitored, and updates made as appropriate. The figures present averages and percentiles, and significant variation may exist within the data underlying these values. Some of the figures reflect different time periods and patient populations that may not be directly comparable.

[9] The Milliman HCG Grouper processes medical claims, pharmacy claims and member eligibility data using the Milliman HCG categorization and utilization counting methodology. The application also produces validation and cost summary reports

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

Spencer Marshall
Spencer.Marshall@milliman.com

Brian Hartman
Brian,Hartman@milliman.com

Christopher Kunkel
Christopher.Kunkel@milliman.com

June Chu
June.Chu@milliman.com