Original Manuscript

SGS SOUTHERN GERONTOLOGICAL SOCIETY

Journal of Applied Gerontology
2023, Vol. 0(0) 1–10

DOI: 10.1177/07334648231223449
journals.sagepub.com/home/jag
Sage

Comparative Performance of Three Claims-Based Frailty Measures Among Medicare Beneficiaries

Sara E. Heins[1], Denis Agniel[2], Jacob Mann[3], and Melony E. Sorbero[1]

Abstract

Frailty is an important predictor of mortality, health care costs and utilization, and health outcomes. Validated measures of frailty are not consistently collected during clinical encounters, making comparisons across populations challenging. However, several claims-based algorithms have been developed to predict frailty and related concepts. This study compares performance of three such algorithms among Medicare beneficiaries. Claims data from 12-month continuous enrollment periods were selected during 2014–2016. Frailty scores, calculated using previously developed algorithms from Faurot, Kim, and RAND, were added to baseline regression models to predict claims-based outcomes measured in the following year. Root mean square error and area under the receiver operating characteristic curve were calculated for each model and outcome combination and tested in subpopulations of interest. Overall, Kim models performed best across most outcomes, metrics, and subpopulations. Kim frailty scores may be used by health systems and researchers for risk adjustment or targeting interventions.

Keywords

frailty, Medicare, claims analysis, risk-adjustment

What this paper adds

- Compared performance of three previously developed claims-based algorithms to predict hospitalizations, nursing home stays, and days home in the following year.
- Data used to test algorithm performance is many times larger than used in previous validations.
- First study to study potential bias of frailty algorithms by comparing performance of frailty algorithms by race/ethnicity and neighborhood socioeconomic status.

Application of study findings

- Models including Kim frailty scores best predicted three outcomes and require only routinely collected claims data.
- Kim frailty scores may be used by health systems and researchers for risk adjustment or to target interventions towards frail individuals.

Introduction

Frailty is a significant concern among older adults and identifying frailty is important to health care providers, health systems, and researchers. While there is no single definition of frailty, it is generally characterized by a collection of symptoms including unintentional weight loss, decreased strength, and increased susceptibility to falls and fractures (Ensrud et al., 2009; Fried et al., 2001; Xue, 2011). Higher levels of measured frailty are associated with increased mortality, health care utilization, and medical spending (Graham et al., 2009; Kan et al., 2018; McIsaac et al., 2016).

Despite frailty's importance, it is challenging to quantify (Rockwood, 2005; Sternberg et al., 2011). A common approach is to assess dependency in activities of daily living (ADLs), such as bathing, dressing, and preparing and eating food, and to use these as proxies for

Manuscript received: April 24, 2023; **final revision received:** August 8, 2023; **accepted:** September 11, 2023.

[1]RAND Corporation Pittsburgh, Pittsburgh, PA, USA
[2]RAND Corporation, Santa Monica, CA, USA
[3]Solutions By Design, Vienna, VA, USA

Corresponding Author:
Sara E. Heins, RAND Corporation Pittsburgh, 4570 Fifth Ave #600, Pittsburgh, PA 15213, USA.
Email: sheins@rand.org

frailty (Costenoble et al., 2021; Faurot et al., 2015). ADL dependency is a strong predictor of early mortality (Keeler et al., 2010; Rosenberg et al., 2019; Walter et al., 2001) and increased health care utilization and spending (Rosenberg et al., 2019; Williams et al., 1994). Functional impairment, encompassing ADLs, mobility, and memory limitations, has also been used as a proxy for frailty (Heins et al., 2023). Yet another approach to measuring frailty is to count an "accumulation of deficits" including the previously discussed ADLs, symptoms such as hearing or vision loss, and abnormal laboratory values such as low calcium or abnormal kidney function (Kim et al., 2018; Mitnitski et al., 2001). However, many common measures of frailty, such as ADL dependency, are not routinely collected for all patients. Individuals are usually assessed for ADL dependency when admitted to a post-acute care (PAC) setting such as a skilled nursing facility (SNF) or when initiating Home Health Agency (HHA) care, but these assessments are not typically made as part of routine clinical care in broader populations of older adults.

To address this limitation, several claims-based frailty indices (CFIs) have been developed. A recent systematic review of claims-based frailty indices identified 14 CFI development studies (Shashikumar et al., 2020). These CFIs were calculated using diagnosis and procedure codes from routinely collected administrative claims to predict risk of frailty. CFIs have been developed using different populations and predictors and validated using different outcomes (Shashikumar et al., 2020), and CFIs may be optimized for different populations or applications. For example, predictors of frailty among younger populations may differ from predictors that are significant in older populations. Two previously developed and validated algorithms include those by Faurot et al. (Faurot et al., 2015) and Kim et al. (Kim et al., 2018) Both of these algorithms were developed using data from the Medicare Current Beneficiary Survey (MCBS), which links fee-for-service (FFS) Medicare claims with panel survey data for a representative sample of beneficiaries. Both algorithms started with a pool of claims-based predictors derived from diagnosis and procedure codes and used selection methods to empirically identify predictors and develop coefficients for models used to predict outcomes measuring proxies for frailty. A third algorithm developed by RAND (Heins et al., 2023) used a larger sample of Medicare FFS beneficiaries linked with PAC assessment data to develop two algorithms. These algorithms started with predictors from Faurot et al. and Kim et al. and chronic and disabling conditions from the Chronic Conditions Data Warehouse (CCW) (Centers for Medicare and Medicaid Services, 2022) and used similar selection methods to identify predictors and develop coefficients for a model predicting activity and mobility limitations and a second predicting memory limitations. RAND used inverse probability weighting to make the population of patients receiving PAC services on which the algorithms were developed more representative of the overall Medicare FFS population. The objective of this paper is to compare the relative performance of these three frailty algorithms (Faurot, Kim, and RAND) at predicting claims-based outcomes among Medicare FFS beneficiaries.

Methods

Data and Study Population

Data for this study consisted of all Medicare FFS beneficiaries with at least 12 months of continuous enrollment during 2014–2016. For each eligible beneficiary, we randomly selected an eligible 12-month reference period which was used to ascertain predictors from the Faurot, Kim, and RAND algorithms. The 12 months following the reference period was used to calculate claims-based outcomes which were then used to independently evaluate model performance. This study was approved by the institutional review board of the lead author.

Score Calculation

For each beneficiary's reference period, we ascertained Faurot, Kim, and RAND predictors, based on methods described in their respective manuscripts. Summaries of each algorithm are presented in Table 1 and detailed model specifications are in Supplementary Tables 1-5. All three algorithms used International Classification of Disease, Version 9 or 10 (ICD-9/10) diagnosis codes and Healthcare Common Procedure Coding System (HCPCS) codes to identify the presence of a specific condition or health care service associated with frailty. The Faurot and RAND algorithms also included age and sex and the Faurot model also included race. The Faurot algorithms used claims from the last eight months of the reference period while the Kim and RAND algorithms used claims from the entire 12-month reference periods. Using these specifications, we calculated the following four "predicted scores" for each beneficiary on a random 80 percent of the study population to serve as our "training" set: 1) Predicted probability of having a memory limitation (RAND), 2) Predicted number of activity/mobility limitations out of a possible six (RAND), 3) Predicted Survey Frailty Index (SFI)—representing the proportion of abnormalities present out of a total of 56 possible self-reported symptoms, diagnoses, and functional limitations in the MCBS (Kim), and 4) Predicted probability of having at least one self-reported dependency of six ADLs (Faurot).

Claims-Based Outcome definitions

Next, we constructed three claims-based outcomes for each beneficiary in the sample based on the 12 months following the randomly selected 12-month reference period. These three outcomes were: 1) Number of hospitalizations, 2) Nursing facility stay (including long-term stays, dichotomized as yes or no), and 3) Days at home (Number of days

Table 1. Comparison of RAND, Kim, and Faurot Frailty Models.

	RAND Memory Limitations	RAND Activity and Mobility Limitations	Kim Frailty	Faurot Frailty
Total number of predictors	134	114	93	29
Types of predictors	Age, sex, "proximal" and "multiple prior" versions of indicators from Kim, Faurot, and CCW		Indicators of diagnoses and services	Age, sex, race, indicators of diagnoses and services
Timeframe, Frequency, and setting of diagnoses	"Proximal" indicators: At least one inpatient claim in last two weeks of reference period "Multiple Prior" indicators: At least two claims in the rest of the 12-month reference period		At least one claim in 12 months	At least one claim in 8 months
Population for development	Medicare FFS beneficiaries with a PAC stay following ≥12 months of continuous enrollment		MCBS Participants ≥65	MCBS Participants >65
Outcomes	Dichotomous memory limitations Item	Count of activity and mobility limitations 0-6	Survey frailty index (deficit accumulation approach)	ADL dependency

Abbreviation: ADL, Activities of Daily Living; CCW, Chronic Conditions Warehouse; FFS, Fee-For-Service; MCBS, Medicare Current Beneficiary Survey.

beneficiary was alive, not in the hospital, and not in a nursing facility). Hospitalizations and nursing facility stays were determined based on Medicare claim type and place of service codes. These outcomes were chosen because they were not used in the development and validation of the original Faurot, Kim, and RAND studies, they were available for all beneficiaries in the study data set, and they are indicative of poor health outcomes (hospitalization and nursing home stay) and decreased independence (nursing home stay), while days at home reflects both health status and ability to maintain independence.

Model Specifications

Because their relationship with outcomes may be nonlinear, we grouped each of these predicted scores by decile (hereafter, "decile" predictor) and based on categories suggested by Kim

(Kim et al., 2020) representing the ≤10th, 11th–25th, 26th–75th, 76th–90th, and >90th percentiles (hereafter, "categorical" predictor). Each decile/percentile was calculated from the full study population of Medicare FFS beneficiaries. Decile and categorical predictors were represented by nine and four dummy variables, respectively.

We next performed regressions in our training set using the three claims-based outcomes described above. For each outcome, we used age (both as a continuous variable and in five-year age categories to allow for non-linear relationships) and sex to define our baseline model. We then ran five models using the baseline model predictors plus one or more of the previously defined predicted scores: 1) Baseline model plus continuous predicted probability of having at least one ADL dependency (Faurot), 2) Baseline predictors plus continuous predicted SFI (Kim), 3) Baseline predictors plus continuous predicted probability of having a memory limitation (RAND), 4) Baseline predictors plus continuous predicted number of activity/mobility limitations (RAND), and 5) Baseline predictors plus continuous predicted probability of having a memory limitation AND continuous predicted number of activity/mobility limitations (RAND).

For each of the above five specifications we also developed versions of each model where the continuous predictor was replaced by the decile-based predictor (e.g., continuous predicted Kim score replaced by decile-based predicted Kim score) and where the continuous predictor was replaced by the categorical predictor.

For each model, we calculated the Root Mean Squared Error (RMSE) and AUC (Area Under the Receiver Operating Characteristic Curve) in the 20 percent "validation" set. RMSE is a summary measure of how far the actual outcome values are from the predicted outcome values, with smaller RMSE indicating better model fit. To calculate AUC for the number of hospitalizations outcome, we dichotomized the outcome as 0 and 1+ hospitalizations. AUC is a measure of the ability of a model to correctly identify a dichotomous outcome. AUC values range from 0 to 1 with higher numbers indicating better diagnostic ability and .5 indicating a model no better than chance.

Subpopulation Analyses

We hypothesized that some algorithms may perform better than others in specific subpopulations. For example, the Faurot and Kim algorithms did not include individuals <65 in their developments whereas the RAND algorithm did, so we hypothesized that the RAND algorithm may perform best in this group. For other stratifications, we were interested in comparing the differences between the baseline model that just included demographic information with other models, by group. For example, we hypothesized that underreporting of diagnoses may be worse for certain subpopulations and would appear as variation in the magnitude of the difference

between the baseline algorithm and the other algorithms. Given the large number of outcomes and subpopulations tested, we performed the subpopulation analyses only on the two best performing models in the overall population. The definitions and rationale for examining these subpopulations are as follows:

ICD Version. In October 2015, the diagnosis coding system used in claims transitioned from ICD-9 to ICD-10. ICD-10 codes tend to contain significantly more granularity. Crosswalking between diagnoses is straightforward using CMS general equivalence mapping (National Bureau of Economic Research, 2012) and an ICD-10 version of Kim's predictors has previously been validated (Gautam et al., 2021). However, in practice, certain diagnoses may be used more or less commonly following the transition irrespective of true condition prevalence (Slavova et al., 2018). For example, one study compared diagnosed condition prevalence for 34 chronic conditions immediately before and after the ICD-10 transition (Yoon & Chow, 2017). While this study found that prevalence estimates were similar for most conditions, there were significant differences for some conditions that are important predictors in the three algorithms of interest (e.g., Alzheimer's disease and spinal cord injury had over twice the odds of diagnosis in the ICD-10 period while arthritis had half the odds of being diagnosed in the ICD-10 period). If important predictors are coded more or less frequently between the ICD-9 and ICD-10 periods, it is possible that the algorithms using these diagnosis codes may perform differently in these periods as well. We therefore examined performance separately among beneficiaries with reference periods entirely prior to October 1, 2015 (ICD-9 only), entirely on or after October 1, 2015 (ICD-10 only), and spanning the ICD-9 to ICD-10 transition (both).

Race/Ethnic Group. A documented shortcoming of claims-based indicators as proxies of health status is the tendency to underreport conditions among Black and Hispanic individuals who may have more limited health care utilization relative to health status (Obermeyer et al., 2019). We thus wanted to determine whether there were differences in algorithm performance between racial/ethnic groups.

Area Deprivation Index. The ADI provides rankings of zip codes by socioeconomic disadvantage based on factors such as education and income (Kind et al., 2014). Beneficiaries residing in more disadvantaged deciles may have lower health care utilization relative to health status and, as a result, model performance may be poorer among these individuals.

Age Group. The relationship between conditions and functional limitations may vary by age. Of particular interest is the difference between the over- and under-65 populations. Medicare beneficiaries under the age of 65 typically qualify based on disability whereas beneficiaries 65 and older can qualify solely on age, so we may expect to see differences in these populations. Furthermore, the RAND algorithms included beneficiaries under 65 in their development whereas the Kim and Faurot algorithms did not.

Receiving PAC Services. The Faurot and Kim algorithms were developed using the MCBS, a representative population of Medicare FFS beneficiaries. In contrast, RAND algorithms were developed on a population of Medicare FFS beneficiaries who received PAC services. We hypothesized that the RAND models would perform better among patients who received PAC services at some point during the reference period, and worse among patients who did not receive such services.

Results

Study Population

Study population characteristics are presented in Table 2. The mean age of Medicare beneficiaries was 71.50 years (standard deviation = 12.57 years) and 83.35% of the population was older than 65. Approximately one in five beneficiaries (19.01%) were dually enrolled in Medicaid and a similar percentage (21.65%) received a low-income Part D subsidy. A plurality (38.4%) of beneficiaries were from the southern United States and the majority of beneficiaries (81.63%) were white. Approximately one in ten beneficiaries (10.66%) received SNF or HHA services during the reference period. In the 12 months following the reference period, 5.11% of beneficiaries had a nursing facility stay and had an average of .26 hospitalizations and 344.78 days at home.

Comparison of Models

RMSE and AUC for each of the models tested on the overall population are presented in Table 3. The Kim decile and continuous models performed the best overall. The Kim decile model had the lowest RMSE for the number of hospitalizations outcome (.718 compared to .747 for the baseline model, a 3.88% improvement) and the Kim continuous model had the highest AUC for the nursing facility outcome (.882 compared to .754 for the baseline model, a 16.98% improvement). The RAND memory recall and activity/mobility limitations decile model had the lowest RMSE for the nursing facility outcome (.200) followed closely by the Kim decile model (.201). The Kim continuous model and the Kim decile model had the lowest RMSE for the days home outcome and the highest AUC for the ≥1 hospitalization outcome (.174 and .734, respectively, for both models). The Kim continuous model had a high RMSE (.810) for the number of hospitalizations; this was in fact 8.43% higher than the baseline model RMSE (.747) for the same outcome. This was likely because there were a small number of Kim scores that were outliers with respect to the median. These outlier scores

Table 2. Population Characteristics[a] (*N* = 38,248,756).

Characteristic	Statistic
Age,[b] mean (SD)	71.50 (12.57)
>65,[b] *N* (%)	31,878,792 (83.35%)
Male, *N* (%)	17,813,493 (46.57%)
Medicaid enrolled, full or partial,[c] *N* (%)	7,271,177 (19.01%)
Low income subsidy (part D),[c] *N* (%)	8,280,054 (21.65%)
ADI,[d] mean (SD)	50.81 (18.68)
Region[e]	
Midwest, *N* (%)	8,366,812 (21.87%)
Northeast, *N* (%)	7,069,938 (18.48%)
South, *N* (%)	14,697,326 (38.43%)
West, *N* (%)	7,251,569 (18.96%)
Missing or U.S. Territory, *N* (%)	863,111 (2.26%)
Race	
Non-Hispanic White, *N* (%)	31,225,888 (81.64%)
Non-Hispanic Black, *N* (%)	3,661,825 (9.57%)
Hispanic, *N* (%)	840,531 (2.20%)
Asian, *N* (%)	785,479 (2.05%)
Other/Unknown, *N* (%)	1,735,033 (4.54%)
Received SNF and/or HHA services during reference period, *N* (%)	4,076,790 (10.66%)
NF Stay in 12 months following reference period, *N* (%)	1,955,619 (5.11%)
Number of hospitalizations in 12 months following reference period, mean (SD)	.26 (.75)
Days home in 12 months following reference period, *N* (SD)	344.78 (69.66)

Abbreviations: ADI, Area Deprivation Index; DME, Durable Medical Equipment; HHA, Home Health Agency; PAC, Post-Acute Care; NF, Nursing Facility; SNF, Skilled Nursing Facility.
[a]All beneficiaries included in population must have had continuous enrollment data during reference period.
[b]Age as of end of twelve-month reference period.
[c]Enrolled at any point during twelve-month reference period.
[d]Based on county of residence at end of the twelve-month reference period. ADI not calculated for 1.37% of beneficiaries with missing county.
[e]Based on state of residence at the end of the six-month reference period.

created more variability and a higher RMSE for the continuous model but exerted less influence over the results in the decile and categorical models. Thus, between the Kim decile model and Kim continuous model specifications, the decile model was preferred for additional analyses. Based on these results, we next tested both the Kim decile model and the RAND memory and activity/mobility limitations decile model to compare performance in subpopulations of interest.

Subpopulation Analyses

Results showing model performance in subpopulations of interest for the number of hospitalizations, nursing facility stay, and days at home outcomes as measured by RMSE and AUC are shown in Tables 4 and 5, respectively. Alternate versions of these tables showing percentage improvement from the baseline model for RMSE and AUC are in Supplementary Tables 6 and 7, respectively. Overall, results showed differences in RMSE and AUC between subpopulations of interest. In general, the results for subpopulations were consistent with our main findings: Algorithms that performed better at predicting an outcome overall tended to also predict that outcome better across subpopulations.

The RAND and Kim models both performed worse among Black beneficiaries than among other racial groups as measured by the highest RMSE in both the RAND and Kim models for all but the "days at home" outcome and the lowest AUC for the Kim and RAND models across all three outcomes. However, similar trends were seen in the baseline model, which did not include any indicators of diagnoses or service which are dependent on health care utilization. The percentage improvement from the baseline model to the RAND and Kim models were fairly consistent across subgroup and actually showed the greatest improvement, as indicated by percent reduction in RMSE, among Black beneficiaries for all three outcomes. Similar trends were seen using AUC. Likewise, with ADI deciles, more advantaged deciles tended to have better model performance, but the improvement from the baseline to models including RAND and Kim predictors was relatively consistent across ADI.

ICD-9 and ICD-10 versions of the models performed similarly, with ICD-10 versions having a slightly lower RMSE across all three outcomes. There were no differences

Table 3. Model Comparison on Claims-Based Outcomes in Full Medicare Populations.

	RMSE (% Difference from Baseline)			AUC (% Difference from Baseline)		
Model	# Hospitalizations	≥1 NF Stay	Days Home	≥1 Hospitalization	≥1 NF Stay	Days Home
Baseline (age and sex)	.747	.213	.184	.616	.754	.719
RAND-M, continuous	.747 (<.01%)	.203 (−4.69%)	.18 (−2.17%)	.615 (−.16%)	.821 (8.89%)	.744 (3.48%)
RAND-M, categorical	.744 (−.40%)	.207 (−2.82%)	.182 (−1.09%)	.647 (5.03%)	.815 (8.09%)	.745 (3.62%)
RAND-M, decile	.743 (−.54%)	.207 (−2.82%)	.182 (−1.09%)	.649 (5.36%)	.818 (8.49%)	.748 (4.03%)
RAND-AM, continuous	.743 (−.54%)	.204 (−4.23%)	.178 (−3.26%)	.645 (4.71%)	.852 (13%)	.78 (8.48%)
RAND-AM, categorical	.737 (−1.34%)	.202 (−5.16%)	.178 (−3.26%)	.656 (6.49%)	.854 (13.26%)	.785 (9.18%)
RAND-AM, decile	.736 (−1.47%)	.202 (−5.16%)	.177 (−3.8%)	.658 (6.82%)	.857 (13.66%)	.788 (9.6%)
RAND-M-AM, continuous	.741 (−.80%)	.202 (−5.16%)	.178 (−3.26%)	.65 (5.52%)	.857 (13.66%)	.78 (8.48%)
RAND-M-AM, categorical	.732 (−2.01%)	.201 (−5.63%)	.177 (−3.80%)	.68 (10.39%)	.86 (14.06%)	.787 (9.46%)
RAND-M-AM, decile	.731 (−2.14%)	**.200**[a] **(−6.10%)**	.177 (−3.80%)	.685 (11.2%)	.864 (14.59%)	.79 (9.87%)
Kim, continuous	.810 (8.43%)	.203 (−4.69%)	**.174**[a] **(−5.43%)**	**.734**[a] **(19.16%)**	**.882**[a] **(16.98%)**	**.825**[a] **(14.74%)**
Kim, categorical	.720 (−3.61%)	.202 (−5.16%)	.175 (−4.89%)	.723 (17.37%)	.875 (16.05%)	.819 (13.91%)
Kim, decile	**.718**[a] **(-3.88%)**	.201 (−5.63%)	**.174**[a] **(−5.43%)**	**.734**[a] **(19.16%)**	.881 (16.84%)	.823 (14.46%)
Faurot, continuous	.757 (1.34%)	.205 (−3.76%)	.176 (−4.35%)	.675 (9.58%)	.869 (15.25%)	.804 (11.82%)
Faurot, categorical	.730 (−2.28%)	.201 (−5.63%)	.176 (−4.35%)	.672 (9.09%)	.864 (14.59%)	.801 (11.4%)
Faurot, decile	.728 (−2.54%)	.201 (−5.63%)	.175 (−4.89%)	.681 (10.55%)	.872 (15.65%)	.807 (12.24%)

Abbreviations: AM, Activity/Mobility Limitations, AUC, Area Under the Curve; ADI, Area Deprivation Index; ICD-9/10, International Classification of Disease, Versions 9/10; M-AM, Memory and Activity/Mobility Limitations; M, Memory Limitations; PAC, Post-Acute Care; RMSE, Root Mean Squared Error.
[a]The best performing model as indicated by the lowest RMSE or the highest AUC.

Table 4. Root Mean Squared Error of Models by Subpopulation.

		Hospitalizations			Nursing Facility Stay			Days Home		
Category	Subpopulation	Baseline	M-AM Decile	Kim Decile	Baseline	M-AM Decile	Kim Decile	Baseline	M-AM Decile	Kim Decile
ICD Version	ICD-9	.791	.776	**.761***	.238	**.224***	.225	.245	.233	**.228***
	ICD-10	.694	.678	**.667***	.180	**.169***	.171	.119	.118	**.116***
	Both	.729	.713	**.700***	.204	**.191***	.192	.144	.141	**.140***
Race/Ethnic group	Asian	.591	.583	**.566***	.178	.175	**.168***	.164	.162	**.158***
	Black	.960	.933	**.914***	.229	**.211***	.214	.187	.179	**.176***
	Hispanic	.789	.771	**.752***	.185	.179	**.176***	.171	.168	**.165***
	White	.727	.713	**.700***	.216	**.203***	.204	.187	.180	**.177***
	Other	.617	.603	**.589***	.139	.133	**.131***	.138	.134	**.132***
ADI decile (1 = most advantaged decile, 10 = least advantaged decile)	1	.701	.687	**.674***	.213	.203	**.202***	.174	.168	**.164***
	2	.747	.732	**.718***	.217	.206	**.206***	.179	.173	**.170***
	3	.717	.703	**.691***	.208	.197	**.196***	.176	.170	**.167***
	4	.736	.721	**.708***	.211	**.199***	.200	.181	.174	**.171***
	5	.748	.734	**.720***	.205	**.194***	.195	.182	.175	**.173***
	6	.746	.731	**.718***	.208	**.195***	.196	.185	.178	**.175***
	7	.765	.749	**.736***	.214	**.200***	.202	.188	.180	**.177***
	8	.771	.755	**.741***	.230	**.213***	.215	.190	.183	**.179***
	9	.770	.754	**.741***	.222	**.206***	.209	.192	.184	**.181***
	10	.801	.783	**.769***	.216	**.201***	.204	.194	.187	**.184***
	Missing	.369	.318	**.252***	.100	.073	**.052***	.175	**.171***	.173
Age group	<65	.948	.929	**.903***	.170	**.162***	.164	.148	.146	**.143***
	≥65	.700	.685	**.675***	.221	**.207***	.208	.190	.183	**.180***
PAC	Yes	1.363	1.306	**1.280***	.424	**.395***	.399	.361	.339	**.330***
	No	.635	.629	**.617***	.172	**.162***	**.162***	.150	.147	**.145***

Abbreviations: AM, Activity/Mobility Limitations; ADI, Area Deprivation Index; ICD-9/10, International Classification of Disease, Versions 9/10; M-AM, RAND Memory and Activity/Mobility Limitations; PAC, Post-Acute Care; RMSE, Root Mean Squared Error.
*The best performing model as indicated by the lowest RMSE.

Table 5. Area Under the Curve by Subpopulation.

		Hospitalizations			Nursing Facility Stay			Days Home		
Category	Subpopulation	Baseline	M-AM Decile	Kim Decile	Baseline	M-AM Decile	Kim Decile	Baseline	M-AM Decile	Kim Decile
ICD Version	ICD-9	.613	.676	**.731***	.754	.864	**.881***	.712	.786	**.818***
	ICD-10	.619	.699	**.734***	.747	.852	**.872***	.715	.795	**.814***
	Both	.610	.685	**.731***	.763	.882	**.887***	.704	.781	**.810***
Race/Ethnic group	Asian	.640	.706	**.761***	.749	.865	**.880***	.738	.793	**.817***
	Black	.586	.687	**.744***	.764	.862	**.894***	.670	.772	**.808***
	Hispanic	.604	.692	**.767***	.684	.856	**.869***	.721	.780	**.811***
	White	.617	.682	**.728***	.744	.856	**.895***	.724	.792	**.824***
	Other	.604	.687	**.747***	.761	.865	**.880***	.699	.771	**.807***
ADI decile (1 = most advantaged decile, 10 = least advantaged decile)	1	.645	.701	**.746***	.727	.857	**.892***	.745	.806	**.837***
	2	.632	.695	**.740***	.755	.850	**.881***	.732	.798	**.828***
	3	.624	.687	**.732***	.751	.855	**.879***	.732	.799	**.832***
	4	.621	.687	**.731***	.763	.864	**.884***	.728	.795	**.829***
	5	.621	.684	**.730***	.760	.862	**.880***	.727	.795	**.827***
	6	.614	.683	**.729***	.766	.866	**.882***	.718	.790	**.825***
	7	.612	.681	**.728***	.761	.874	**.886***	.716	.789	**.823***
	8	.606	.676	**.724***	.760	.870	**.880***	.711	.784	**.819***
	9	.600	.672	**.720***	.751	.868	**.878***	.706	.782	**.816***
	10	.593	.673	**.722***	.749	.867	**.875***	.692	.773	**.808***
	Missing	.541	.680	**.858***	.734	.864	**.872***	.715	.727	**.731***
Age group	<65	.552	.642	**.735***	.646	.772	**.908***	.614	.692	**.767***
	≥65	.623	.692	**.734***	.611	.809	**.833***	.725	.799	**.827***
PAC	Yes	.510	.554	**.586***	.761	.868	**.884***	.603	.642	**.675***
	No	.602	.653	**.707***	.599	**.699***	.687	.709	.750	**.783***

Abbreviations: AM, Activity/Mobility Limitations; ADI, Area Deprivation Index; ICD-9/10, International Classification of Disease, Versions 9/10; M-AM, RAND Memory and Activity/Mobility Limitations; PAC, Post-Acute Care.
The best performing model as indicated by the highest Area Under the Curve.

in which model performed better when stratifying by age group.

While the Kim model performed better overall, the RAND model performed better on a few metrics, primarily with the nursing facility outcome. The RAND model had slightly lower RMSE for all ICD categories, among Black and White Medicare beneficiaries, among the seven most disadvantaged deciles, among both age groups, and among both PAC and non-PAC patients for the nursing facility outcome. However, the Kim model had a higher AUC for all subpopulations using the nursing facility outcome, except among those who received SNF or HHA services during the reference period. The only other metric where the RAND model was superior was RMSE for "missing" ADI decile with the days at home outcome.

Discussion

This study examined the relative performance of three frailty algorithms using multiple metrics and outcomes and in different subpopulations. Overall, we found that using deciles of scores predicted by the Kim algorithm in combination with sex and age best predicted the claims-based outcome measures of hospitalizations and days at home across most measures of model fit and subgroups. The Kim model and the RAND model incorporating age, sex, and decile scores predicted by the activity and mobility limitations and memory limitations algorithms were comparable at predicting nursing facility stay in the following year, with the RAND model performing better among PAC patients.

The RAND algorithms had some potential advantages and disadvantages in their development. The data set used to develop the RAND algorithms, which integrated Medicare claims data with PAC assessment data, contained substantially more beneficiaries than the MCBS data used to develop the Kim and Faurot algorithms. Furthermore, the RAND population included Medicare beneficiaries under the age of 65, eligible for Medicare due to a disability or end stage renal disease, allowing for examination of an important population with potentially high rates of functional impairment. The structure of the data used to develop the RAND models allowed for inclusion of all claims up until the assessment date, whereas the MCBS used panel data that may result in up to a four-month delay between the end of the claims reference period and outcome measurement. The primary disadvantage of the RAND algorithms was that outcome data were only available for beneficiaries with PAC assessment data, who are not representative of the overall Medicare population. A final difference between the outcomes in the PAC assessment data

used to develop the RAND algorithms and the survey outcomes in the MCBS used to calculate the Faurot and Kim algorithms was that the former were assessed by a health care professional while the latter were self-reported. There is not a clear consensus in the literature as to whether self-reported or observed ADL outcomes are more accurate, and relative performance of the measures may depend on the assessors and study population. For example, self-reported outcomes may be more accurate for assessing low levels of disability (Kuhn et al., 2006), while clinician-reported outcomes may be preferable if the assessed population has cognitive impairment (Sager et al., 1992). While the Faurot and Kim algorithms were developed using MCBS data, the Kim algorithms likely benefited from using an outcome that captured more dimensions of frailty than the ADL dependency outcome used by the Faurot algorithms. The Kim study also started with a larger pool of candidate predictors to select from and retained more variables in its final algorithm in comparison to the Faurot study.

While we focused on the two highest performing models for our outcomes of interest (those using decile scores from the Kim and RAND algorithms), there are tradeoffs involved with using different types of models and the appropriate selection depends on several different factors. For example, although the models using Kim scores performed better in our study across most outcomes and subpopulations, the magnitude of the differences in performance were small. Similarly, in a prior study, the Kim and Faurot algorithms along with two other algorithms (Davidoff et al., 2013; Segal et al., 2017) were compared against additional outcomes (a frailty phenotype based on an accumulation of deficit approach and ADL dependencies). In this study, the Kim algorithms best predicted both outcomes and both the Kim and Faurot algorithms outperformed the Davidoff and Segal algorithms, though again the magnitude of the differences was small and the Davidoff and Segal algorithms used fewer predictors than both the Kim and Faurot algorithms. If computation time or algorithm complexity is a concern, using an algorithm with fewer predictors may be preferable. However, if scores are used for large scale risk adjustment in research studies, the additional predictors may be worth the improvement in performance. Adding such scores to existing Medicare databases would eliminate researcher concerns about increased computation time or complexity of the algorithms.

The present study is the first to examine the performance of these three algorithms by subpopulations of interest and these results are important for understanding how these algorithms should be applied and interpreted. Based on prior research indicating bias in claims-based risk adjustment algorithms, we were interested in understanding how these algorithms would perform among different races, ethnicities, and socioeconomic groups. The results of this study suggested that model performance was worse for some metrics for non-Hispanic Black beneficiaries as compared to other racial/ethnic groups. Similarly, for beneficiaries living in more disadvantaged areas the algorithms tended to have worse performance as compared to beneficiaries in more advantaged areas. However, the gap between the performance of the best Kim and RAND models and a baseline model that solely included age and sex variables and no indicators of health care diagnoses or utilization was fairly consistent across racial/ethnic and ADI subgroups. These results suggested that differences in model performance by outcome were related more to outcome variability by subgroup than by bias in the model related to differential service utilization. Differences in health care utilization relative to health status by race, ethnicity, and socioeconomic status could be less pronounced within the Medicare FFS population where levels of health care coverage are similar. It is possible that the algorithms may exhibit greater bias if applied to populations with less consistent insurance coverage (Obermeyer et al., 2019). It is also possible that the claims-based outcomes, such as hospitalization or nursing facility stay, may be subject to some degree of the same biases as the predictors if race is correlated with health care utilization, more broadly (Obermeyer et al., 2019). While the results of these analyses do not support bias inherent in the models, the worse performance noted in specific subgroups point to a need for caution in using claims-based algorithms for risk-adjustment, more broadly. Additional research on potential bias in these and other claims-based algorithms is needed, particularly in populations with greater variation in health care access. All three outcome measures tested were claims-based, measured similar constructs, and were subject to similar measurement biases. In the future, it would be useful to repeat these analyses using other non-claims-based outcome measures.

We also found that ICD-9 and ICD-10 versions of the models performed similarly. Our results indicated no concerns about using RAND or Kim algorithms differently before or after the ICD-9 to ICD-10 transition. Across all outcomes and metrics, there were no differences in algorithm performance by age group. Overall, the Kim model performed better than the RAND model on <65 population for most metrics despite not including these beneficiaries in its development.

Overall, we found that algorithms developed by Kim performed the best at predicting claims-based outcomes of interest. The advantage of using the more representative MCBS for development likely outweighed the advantage of the larger data set used to develop the RAND models. Algorithms which can predict frailty using routinely collected administrative data are valuable to many different stakeholders. For example, claims-based algorithms for identifying frailty may be used for risk-adjustment for value-based payments or for targeting interventions to frail individuals. The Kim CFI could be a useful resource for these stakeholders and making these scores more widely available, for example, in CMS Medicare data sets, may encourage their use.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was conducted under a contract with the Office of the Assistant Secretary for Planning and Evaluation (contract #HHSP233201500038I) with funding from the Office of the Secretary Patient Centered Outcomes Research Trust Fund. The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the U.S. Department of Health and Human Services

Research Ethics

IRB approval number: RAND Human Subject Protection Committee 2019-0952-AM01.

ORCID iD

Sara E. Heins https://orcid.org/0000-0003-0196-9000

Supplemental Material

Supplemental material for this article is available online.

References

Centers for Medicare and Medicaid Services. (2022). Other chronic health, Mental health, and potentially disabling conditions. Cetners for Medicare and Medicaid Services. Retrieved from. https://www2.ccwdata.org/web/guest/condition-categories-other

Costenoble, A., Knoop, V., Vermeiren, S., Vella, R. A., Debain, A., Rossi, G., Bautmans, I., Verté, D., Gorus, E., De Vriendt, P., & De Vriendt, P. (2021). A comprehensive overview of activities of daily living in existing frailty instruments: A systematic literature search. *The Gerontologist*, *61*(3), Article e12–e22. https://doi.org/10.1093/geront/gnz147

Davidoff, A. J., Zuckerman, I. H., Pandya, N., Hendrick, F., Ke, X., Hurria, A., Lichtman, S. M., Hussain, A., Weiner, J. P., Edelman, M. J., & Edelman, M. J. (2013). A novel approach to improve health status measurement in observational claims-based studies of cancer treatment and outcomes. *Journal of geriatric oncology*, *4*(2), 157–165. https://doi.org/10.1016/j.jgo.2012.12.005

Ensrud, K. E., Ewing, S. K., Cawthon, P. M., Fink, H. A., Taylor, B. C., Cauley, J. A., Dam, T. T., Marshall, L. M., Orwoll, E. S., & Cummings, S. R. (2009). A comparison of frailty indexes for the prediction of falls, disability, fractures, and mortality in older men. *Journal of the American Geriatrics Society*, *57*(3), 492–498. https://doi.org/10.1111/j.1532-5415.2009.02137.x

Faurot, K. R., Jonsson Funk, M., Pate, V., Brookhart, M. A., Patrick, A., Hanson, L. C., Castillo, W. C., Stürmer, T., & Stürmer, T. (2015). Using claims data to predict dependency in activities of daily living as a proxy for frailty. *Pharmacoepidemiology and Drug Safety*, *24*(1), 59–66. https://doi.org/10.1002/pds.3719

Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W. J., Burke, G., McBurnie, M. A., & Burke, G. (2001). Frailty in older adults: Evidence for a phenotype. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *56*(3), M146–M156. https://doi.org/10.1093/gerona/56.3.m146

Gautam, N., Bessette, L., Pawar, A., Levin, R., & Kim, D. H. (2021). Updating international classification of diseases 9th revision to 10th revision of a claims-based frailty index. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *76*(7), 1316–1317. https://doi.org/10.1093/gerona/glaa150

Graham, J. E., Snih, S. A., Berges, I. M., Ray, L. A., Markides, K. S., & Ottenbacher, K. J. (2009). Frailty and 10-year mortality in community-living Mexican American older adults. *Gerontology*, *55*(6), 644–651. https://doi.org/10.1159/000235653

Heins, S. E., Agniel, D., Mann, J., & Sorbero, M. E. (2023). Development and validation of algorithms to predict activity, mobility, and memory limitations using medicare claims and post-acute care assessments. *Journal of Applied Gerontology*, *42*(7), 1651–1661. https://doi.org/10.1177/07334648231162613

Kan, H. J., Kharrazi, H., Leff, B., Boyd, C., Davison, A., Chang, H.-Y., Kimura, J., Wu, S., Anzaldi, L., Richards, T., Lasser, E. C., Weiner, J. P., & Richards, T. (2018). Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. *Medical care*, *56*(3), 233–239. https://doi.org/10.1097/MLR.0000000000000865

Keeler, E., Guralnik, J. M., Tian, H., Wallace, R. B., & Reuben, D. B. (2010). The impact of functional status on life expectancy in older persons. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *65*(7), 727–733. https://doi.org/10.1093/gerona/glq029

Kim, D. H., Patorno, E., Pawar, A., Lee, H., Schneeweiss, S., & Glynn, R. J. (2020). Measuring frailty in administrative claims data: Comparative performance of four claims-based frailty measures in the US medicare data. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *75*(6), 1120–1125. https://doi.org/10.1093/gerona/glz224

Kim, D. H., Schneeweiss, S., Glynn, R. J., Lipsitz, L. A., Rockwood, K., & Avorn, J. (2018). Measuring frailty in medicare data: Development and validation of a claims-based frailty index. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *73*(7), 980–987. https://doi.org/10.1093/gerona/glx229

Kind, A. J., Jencks, S., Brock, J., Yu, M., Bartels, C., Ehlenbach, W., Greenberg, C., Smith, M., & Smith, M. (2014). Neighborhood socioeconomic disadvantage and 30-day rehospitalization: A retrospective cohort study. *Annals of internal medicine*, *161*(11), 765–774. https://doi.org/10.7326/M13-2946

Kuhn, R., Rahman, O., & Menken, J. (2006). Survey measures of health: How well do self-reported and observed indicators measure health and predict mortality. In B. Cohen & J. Menken (Eds.), *Aging in sub-Saharan Africa: Recommendations for furthering research* (pp. 314–342). National Academies Press.

McIsaac, D., Beaulé, P. E., Bryson, G., & Van Walraven, C. (2016). The impact of frailty on outcomes and healthcare resource

usage after total joint arthroplasty: A population-based cohort study. *The bone & joint journal*, *98-B*(6), 799–805. https://doi.org/10.1302/0301-620X.98B6.37124

Mitnitski, A. B., Mogilner, A. J., & Rockwood, K. (2001). Accumulation of deficits as a proxy measure of aging. *The Scientific World Journal*, *1*, 323–336. https://doi.org/10.1100/tsw.2001.58.

National Bureau of Economic Research. (2012, May 11, 2016). CMS' ICD-9-CM to and from ICD-10-CM and ICD-10-PCS Crosswalk or general equivalence mappings. Retrieved from, https://data.nber.org/data/icd9-icd-10-cm-and-pcs-crosswalk-general-equivalence-mapping.html

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Rockwood, K. (2005). Frailty and its definition: A worthy challenge. *Journal of the American Geriatrics Society*, *53*(6), 1069–1070. https://doi.org/10.1111/j.1532-5415.2005.53312.x

Rosenberg, T., Montgomery, P., Hay, V., & Lattimer, R. (2019). Using frailty and quality of life measures in clinical care of the elderly in Canada to predict death, nursing home transfer and hospitalisation-the frailty and ageing cohort study. *BMJ Open*, *9*(11), Article e032712. https://doi.org/10.1136/bmjopen-2019-032712

Sager, M. A., Dunham, N. C., Schwantes, A., Mecum, L., Halverson, K., & Harlowe, D. (1992). Measurement of activities of daily living in hospitalized elderly: A comparison of self-report and performance-based methods. *Journal of the American Geriatrics Society*, *40*(5), 457–462. https://doi.org/10.1111/j.1532-5415.1992.tb02011.x

Segal, J. B., Chang, H.-Y., Du, Y., Walston, J., Carlson, M., & Varadhan, R. (2017). Development of a claims-based frailty indicator anchored to a well-established frailty phenotype. *Medical care*, *55*(7), 716–722. https://doi.org/10.1097/MLR.0000000000000729

Shashikumar, S. A., Huang, K., Konetzka, R. T., & Joynt Maddox, K. E. (2020). Claims-based frailty indices: A systematic review. *Medical care*, *58*(9), 815–825. https://doi.org/10.1097/MLR.0000000000001359

Slavova, S., Costich, J. F., Luu, H., Fields, J., Gabella, B. A., Tarima, S., & Bunn, T. L. (2018). Interrupted time series design to evaluate the effect of the ICD-9-CM to ICD-10-CM coding transition on injury hospitalization trends. *Injury epidemiology*, *5*(1), 1–12. https://doi.org/10.1186/s40621-018-0165-8

Sternberg, S. A., Wershof Schwartz, A., Karunananthan, S., Bergman, H., & Mark Clarfield, A. (2011). The identification of frailty: A systematic literature review. *Journal of the American Geriatrics Society*, *59*(11), 2129–2138. https://doi.org/10.1111/j.1532-5415.2011.03597.x

Walter, L. C., Brand, R. J., Counsell, S. R., Palmer, R. M., Landefeld, C. S., Fortinsky, R. H., & Covinsky, K. E. (2001). Development and validation of a prognostic index for 1-year mortality in older adults after hospitalization. *JAMA*, *285*(23), 2987–2994. https://doi.org/10.1001/jama.285.23.2987

Williams, B. C., Fries, B. E., Foley, W. J., Schneider, D., & Gavazzi, M. (1994). Activities of daily living and costs in nursing homes. *Health Care Financing Review*, *15*(4), 117–135.

Xue, Q.-L. (2011). The frailty syndrome: Definition and natural history. *Clinics in Geriatric Medicine*, *27*(1), 1–15. https://doi.org/10.1016/j.cger.2010.08.009

Yoon, J., & Chow, A. (2017). Comparing chronic condition rates using ICD-9 and ICD-10 in VA patients FY2014–2016. *BMC Health Services Research*, *17*, 572–576. https://doi.org/10.1186/s12913-017-2504-9.