Development of frailty measurement tools using administrative health data: A systematic review

Author

Nghiem, Son, Sajeewani, Disna, Henderson, Katrina, Afoakwah, Clifford, Byrnes, Joshua, Moyle, Wendy, Scuffham, Paul

Published

2020

Journal Title

Archives of Gerontology and Geriatrics

Version

Accepted Manuscript (AM)

DOI

10.1016/j.archger.2020.104102

Development of frailty measurement tools using administrative health data: A systematic review

Running title: Frailty measurement using administrative data: a review

Son Nghiem[1*], Disna Sajeewani[1], Katrina Henderson[2], Clifford Afoakwah[1], Joshua Byrnes[1], Wendy Moyle[3,4], Paul Scuffham[1,3]

[1]Centre for Applied Health Economics

[2]Health Library,

[3]Menzies Health Institute Queensland

[4]School of Nursing and Midwifery

Griffith University

Queensland, Australia

[*]Corresponding Author: 177 Kessels Road, Nathan, QLD 4111 Australia. Email: s.nghiem@griffith.edu.au. Phone: +61-7-3735-9103

1. Introduction

Frailty is a complex condition associated with declining physical functioning in older adults. Frail older adults are at increased risk of multiple adverse outcomes. Some of the more common adverse outcomes are falls, hospitalisation, physical or mental disability and death (Evans, Sayers, Mitnitski, & Rockwood, 2014). The assessment of the risk for hospital admission has attracted increasing attention in both research and clinical practice internationally as it can play a crucial role in effective planning of health services. Whilst the ability for current frailty measurement methods to result in better outcomes is considered unlikely, electronic frailty indexes based on digitized and routinely collected data are a potentially useful approach to unlocking improved outcomes and save healthcare costs (Turner & Clegg, 2014).

There are a plethora of frailty measurement tools, but the two most common approaches to measurement are frailty phenotypes (Fried et al., 2001) and frailty indices (Rockwood, 1999). A frailty phenotype sums up the number of frailty indicators while a frailty index estimates the fraction of total health deficits. Despite the differences in measurement approaches, frailty measurement tools often require respondents to complete questionnaires. For example, the most-cited "Fried frailty phenotype" (Fried et al., 2001) requires respondents to report their weight loss status by answering the question "In the last year, have you lost more than 10 pounds unintentionally?" Furthermore, they must undertake clinical assessments, physical assessments for walking speed and grip strength under the judgements of health professionals.

Existing measures of frailty use a combination of clinical assessment and survey questionnaires and are time- and resource-intensive, especially to assess frailty at the population level (Gilbert et al., 2018). Whilst it may take only a few minutes to assess the frailty status of one individual, assessing every older patient in a population would require an overwhelming amount of time and resources, and hence most elderly people in hospitals do

not have their frailty status assessed. However, trends in population ageing require frailty screening and monitoring at the population level for more effective planning of health services (Morley et al., 2013). Thus, at the population level, existing tools are not suitable to provide timely frailty assessments. Therefore, a reliable, automated and cost-effective frailty measurement tool is required to meet the healthcare needs of the ageing-population, occurring across the globe. With the increasing digitization of routinely-collected data and analytical power, including artificial intelligence algorithms, frailty can be detected and monitored more effectively at the population level.

In this study, we systematically reviewed articles that developed and validated automated frailty measurement tools using routinely-collected (administrative) health data. Since frailty measurement using routinely-collected health data is a relatively new development, this systematic review implicitly tests the hypothesis that these tools are available and clinically effective. We also discuss opportunities and challenges to apply these tools to healthcare practices.

2. Methods

2.1 Search strategy

A systematic literature search was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009) for articles published in six databases: PubMed, EMBASE, MEDLINE, CINAHL, Scopus and Web of Science. The literature search was conducted from April to December 2019. We also conducted a hand search from the references of the selected articles and contacted authors of the hospital frailty risk score (Gilbert et al., 2018) to increase the number of relevant studies. There is no reporting guideline for prediction models using administrative data. Therefore, we report the findings using the guidelines for meta-

analysis of observational studies in epidemiology (MOOSE) (Straup et al., 2000) because this review includes studies using routinely-collected health data (i.e., observational studies). However, as our review did not focus on clinical interventions, and we did not perform a meta-analysis, many items of the MOOSE guideline were not applicable. Search records were managed using EndNote version 17. The study protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO registration number: CRD42019130357).

The title and abstract searches were conducted using the following search phrases.

[Frail]*

AND [tool OR score OR instrument OR measurement OR assessment OR outcome OR indicators OR index]

AND ["administrative data" OR "health record" OR "routine data" OR "hospital data" OR "big data"]

2.2 Study selection

After removing duplicate records, two researchers initially screened the titles and abstracts to exclude irrelevant articles. Any disagreement between the two researchers was solved by a discussion with the team, and a conclusion was made after detailed examination of those articles.

The following selection criteria were used to select relevant studies from titles and abstracts:

a) a new frailty measurement tool was developed using routinely-collected or linked health data.
b) the statistical validity of the new tool was conducted internally and/or externally (internal validation refers to the use of a subset of the original data to validate findings

while external validation refers to using different data set than the training data to validate findings) (Steyerberg & Harrell, 2016).

c) the study population included older adults.

d) the study was published in a peer-review academic journal.

e) a full text of the study was available in English.

The exclusion criteria were:

a) Clinical studies.

b) Studies published in non-peer review outlets.

c) Studies focused on specific diseases.

d) Studies that did not include older adults (65 years and above).

e) Studies did not conduct any validation test.

2.3 Risk of bias

We used both the CHARMS (Moons et al., 2014) and PROBAS (Wolff et al., 2019) guidelines to assess the risk of bias. Since both guidelines were not designed to assess predictive models using administrative data, some of their criteria such as treatment details or follow-ups were not relevant to studies in this review (see Supplement data for details). In addition, we tailored the assessment by including two additional criteria that we believe are important for this review:

a) Benchmarking against an established frailty measurement tools such as the Fried phenotype (Fried et al., 2001) or Rockwood scale (Rockwood et al., 2005).

b) Being validated independently by other studies.

The popularity of established tools such as the Fried phenotype (Fried et al., 2001) or Rockwood scale (Rockwood et al., 2005) indicates that they are widely accepted, especially

in clinical settings. Thus, studies that benchmark the performance of their proposed tools against the established tools could have high quality (criterion). Tools were verified independently by other authors would be more reliable than tools that were validated by their developers only (criterion b). The same argument has been used to put less weight on self-citation to assess the impact of academic publications (Epstein, 2007).

2.4 Participants and result synthesis

Since there is no universally accepted definition of frailty, we include studies with diverse frailty definitions in their study population. We selected the minimum age cut-off of selected studies to define older adults, which in this study is defined as 65 years and above. The objective of this systematic review was to evaluate the development and application of frailty measurement tools rather than the results of studies. As such, no meta-analysis or synthesis of results was conducted. Instead, a narrative synthesis of the findings from the included studies is presented. Key summary measures include discriminative power and the prediction of adverse health outcomes and the goodness-of-fit of models.

3. Results

The systematic search of the literature using the above-mentioned databases and phrases yielded 464 studies. After removing duplicates, 249 studies were screened on titles and abstracts for further assessment. Applying the exclusion criteria, 235 studies were removed, and 14 articles were retained as potentially relevant full-text articles (Figure 1). Two independent reviewers evaluated the full-text of the 14 articles and reached a substantial agreement with a rate of 85.5% (i.e., Cohen's kappa (Cohen, 1960) is 0.74). After discussions among all team members about the disagreement, five studies were selected for detailed

reviews. Nine excluded studies consisted of three studies that focused on specific diseases, and six studies did not use administrative health data.

All of the five studies included were recently published in the 2016-2019 period (Table 1). Logistic regression was the main analytical method used in all selected studies. Only Bertini et al. (2018) employed other methods (recursive partitioning, generalized boosted, random forest, support vector machines) for robustness checks. All new frailty measures produced fair discriminative power with the value of C-statistics, or its equivalence (Hanley & McNeil, 1982) – area under the curve (AUC), of at least 0.6 on mortality, the most common outcome. Gilbert et al. (2018) was the only study that benchmarked results against the Fried phenotype and the Rockwood scale (Rockwood et al., 2005) but they were also the sole study that did not present any measure of model fit. Two of the studies, both from the UK (Clegg et al., 2016; Gilbert et al., 2018), were validated by additional independent studies.

Table 1. Summary of characteristics of studies included in the narrative synthesis

Criteria	Clegg et al. (2016)	Gilbert et al. (2018)	Bertini et al. (2018)	McIsaac et al. (2019)	Soong et al. (2019)
Data sources	Electronic health records (EHRs)	Linked hospital admission data	12 socio-clinical databases	Linked health data	Linked multinational hospital admission data
Sample size	931,541	22,139	95,368	415,704	294,998
Countries	UK	UK	Italy	Canada	9 countries*
Age of subjects	65-95	75+	65+	>65	>75
Outcomes	Mortality, hospital admission, nursing home admission	30-day mortality, long hospital stay, 30-day ED readmitted	Emergency admission or death	1-year morality, institution discharge	Long hospital stay, 30-day non-elective readmission, mortality
Methods	Cox regression	Cluster analysis, Logistic regression	Logistic regression+ 5 other methods	Logistic regression, generalized estimating equation	Logistic regression

Number of predictors	36	109	27	30	11
C-statistics/AUC					
-Mortality	0.76	0.60	0.70-0.77	0.81	0.62-0.70
-Readmission	0.71	0.56	-	-	0.63-0.64
-Long stay	-	0.68	-	-	0.61-0.69
Goodness-of-fit	Pseudo-R^2 (0.02-0.04)	NA	Hosmer-Lemeshow (test>0.05)	OR for motality (>1)	Homser-Lemeshow (p<0.001)
Benchmark	-	Fried, Rockwood	-	-	Elixhauser comorbidity
Validation					
Internal	Y	Y	Y	Y	Y
External	Y	Y	N	N	Y
Independent	Y(Abbasi et al., 2019; Ambagtsheer et al., 2019; Boyd et al., 2019; Brundle et al., 2019; Lansbury et al., 2017; Stow, Matthews, & Hanratty, 2018)	Y(Eckart et al., 2019; McAlister & van Walraven, 2019)	N	N	N

*Australia, Belgium, Denmark, Finland, Italy, Netherlands, Norway, UK and USA

The summary of predictors in Table 2 shows that all selected studies use the presence of multiple health conditions (e.g., multimorbidity) in frailty assessment. There was also considerable overlap between predictors of the five selected studies and two of the most common frailty measurement tools: the Fried phenotype (Fried et al., 2001) and the Rockwood scale (Rockwood et al., 2005). In this regard, two UK studies (Clegg et al., 2016; Gilbert et al., 2018) were also ranked the highest with at least four predictors overlapping with the Fried phenotype (Fried et al., 2001) and the Rockwood scale (Rockwood et al., 2005).

Table 2. Summary of predictors used in the five selected studies

Criteria	Clegg et al. (2016)	Gilbert et al. (2018)	Bertini et al. (2018)	McIsaac et al. (2019)	Soong et al. (2019)
Multimorbidity[#]	+	+	+	+	+
Low grip strength[*]	?	?	?	?	?
Mobility problem[*#]	+	+	?	?	+
Physical inactivity[*#]	+	+	?	?	?
Exhaustion[*#]	?	?	0	0	0
Weight loss[*]	+	0	0	+	0
Cognitive problem[#]	+	+	+	?	?
Mental problem	?	+	?	?	+
Social vulnerability	+	0	+	+	0
Falls	0	+	0	0	+
Multi-medication	+	0	+	0	0
Demographics (age, gender, nationality, etc.)	?	?	+	+	+

+: clearly included;?: seems to be included, 0: not included; [*]: included in Fried phenotype; [#]included in the Rockwood scale

The selected studies contain a low risk of bias and high applicability using the PROBAS guidelines (Table 3). However, using our additional criterion of being independently validated by other authors, only the two UK studies satisfy. One possible reason for the last three studies not being independently validated could be due to their data structure. Building a data warehouse from a dozen sources by Bertini et al. (2018) or from nine countries across three continents by Soong et al. (2019) were particularly challenging. In addition, the use of comorbidity measures such as Charlson scores by McIsaac, Wong, Huang, Moloo, and van Walraven (2018) and Elixhauser by Soong et al. (2019) may prevent their studies being replicated by other authors because comorbidity may not fully reflect frailty. Overall, the availability of independent validations was the main reason we gave the two UK studies (Clegg et al., 2016; Gilbert et al., 2018) low risk of bias and high applicability.

Table 3. Assessments for risk of bias and applicability of included studies

Authors (year)	Participants	Predictors	Outcome	Analysis	Overall
Risk of bias					
Clegg et al. (2016)	+	+	+	+	+
Gilbert et al. (2018)	+	+	+	+	+
Betini et al. (2018)	+	+	+	-	-
McIsaac et al. (2019)	+	+	+	-	-
Soong et al. (2019)	+	+	+	-	-
Applicability					
Clegg et al. (2016)	+	+	+	+	+
Gilbert et al. (2018)	+	+	+	+	+
Betini et al. (2018)	+	+	+	-	-
McIsaac et al. (2019)	+	+	+	-	-
Soong et al. (2019)	+	+	+	-	-

Note: * + indicates a low risk of bias/low concern regarding applicability; − indicates a high risk of bias/high concern regarding applicability; and? indicates unclear risk of bias/unclear concern regarding applicability.

3.1 Electronic Frailty Index (eFI)

The electronic Frailty Index (eFI) (Clegg et al., 2016) was the first frailty measurement tool that was developed and validated using routinely collected health record data from the UK. The eFI was developed for use as part of routine UK primary care to facilitate early frailty identification and management using, for example, tiered population health management approaches. This frailty measurement tool was developed using anonymised data of 931,541 patients aged 65-95 years old (207,814 in the development cohort; 207,720 in the internal validation cohort; 516,007 in the external validation cohort).

The development and internal validation cohorts were established from electronic health records (EHRs) through the 'ResearchOne' database while the external validation

cohort was established from 'the Health Improvement Network' (THIN) database. The eFI was developed using 36 deficits, constructed from 2,171 Clinical Terms Version 3 (CTV3) Read codes. Their 36-item deficits were then defined by selecting individual codes that increase with age and has a population prevalence greater than 0.5% but less than 100% by age 65. The authors then conducted a hand search to ensure that frailty was measured by organ and physio-social systems. The eFI was estimated by the fraction of total 36 deficits presence in each patient. Quartiles of the eFI were used to categorise patients into four groups: fit, mild, moderate and severe frailty.

The eFI was used to predict all-cause mortality, unplanned hospitalisation, and nursing home admission using Cox regressions. The eFI was shown to have good discrimination with the C-statistics of 0.76, 0.75 and 0.75 for mortality at 1, 3 and 5 years, respectively, with the external validation data. The eFI was also found to have achieved consistent discriminative power in predicting mortality by independent validations. One of the independent validation studies (Brundle et al., 2019) revealed that the eFI was strongly correlated with the Marshal frailty index(Marshall, Nazroo, Tampubolon, & Vanhoutte, 2015) (Spearman correlation coefficient (ρ)=0.68, [95% confidence interval-CI: 0.62-0.74]) and Edmonton Frailty Scale (Rolfson, Majumdar, Tsuyuki, Tahir, & Rockwood, 2006) (ρ=0.63 [0.57-0.69]) while moderately correlated with the Rockwood scale (Rockwood et al., 2005) (ρ=0.59 [0.49-0.65]) and the Fried phenotype (Fried et al., 2001) (ρ=0.51 [0.42-0.59]). The Australian validation of the eFI (Ambagtsheer et al., 2019) also showed moderate agreement with the Fried frailty phenotype (Fried et al., 2001) with the inter-rater reliability (Kappa) of 0.52 [0.29-0.75]) but the eFI produced a very high level of accuracy with a C-statistics of 0.9. The Canadian validation (Abbasi et al., 2019) also revealed a strong correlation between the eFI and a comprehensive geriatric assessment with a correlation coefficient of 0.72 [CI: 0.60-0.81].

3.2 Hospital Frailty Risk Score (HFRS)

The second study from the UK (Gilbert et al., 2018) used a three-step approach to develop and validate the Hospital Frailty Risk Score (HFRS). First, cluster analysis was conducted to categorise patients by frailty characteristics using International Classification of Disease (ICD) codes and resources used (bed days and hospital costs). Second, the HFRS was estimated using parameters from a logistic regression that uses ICD codes as predictors of frailty group. For the ease of interpretation, the HFRS was categorized into three groups: low-, intermediate- and high-risk. Third, the validation of the HFRS was conducted. The authors also found that their HFRS performed well against the Fried phenotype (Fried et al., 2001) and the Rockwood scales (Rockwood et al., 2005) with the respective kappa scores of 0.22 [0.15-0.30] and 0.30 [0.22-0.38].

The HFRS has been validated independently in Switzerland (Eckart et al., 2019) and Canada (McAlister & van Walraven, 2019). The Swiss study was the first independent validation of the HFRS conducted by Eckart et al. (2019) using a cohort of 4,957 emergency admission patients aged 75 years and above. Similar to the original study, the authors found a moderate discriminative performance of the HFRS with the respective AUC for 30-day mortality and a long hospital stay of 0.66 and 0.72. However, this study did not benchmark the HFRS against other frailty measures. The Canadian validation by McAlister and van Walraven (2019) utilised electronic health records of 452,785 patients aged 75 years and older. Logistic regressions were used to determine the association between HFRS groups and adverse health outcomes: mortality and readmission. They found that the HFRS was most strongly associated with prolonged hospital stay and mortality with a C-statistics of 0.71 for both outcomes.

3.3 Frailty Risk Score (FRS)

The third study was conducted in Italy by Bertini et al. (2018) with a comprehensive strategy to develop a data warehouse from 12 socio-clinical sources. The authors developed a frailty risk model based on the prediction of two adverse outcomes: the probability of emergency hospitalization or death within a year. An important contribution of this study is that it also developed an early prediction model to estimate the probability for non-frail patients to become frail within one year. They selected 27 clinical and socio-economic variables from the rich data warehouse using a stepwise approach. The prediction model was validated both internally by randomly splitting the data into a training set (2/3) and a validation set (1/3). External validation was also implemented using the Access Management of Integrated and Automated Social-Health Network (GARCIA). The frailty risk scores were calculated based on parameters from the logistic regressions with higher scores indicating more severe frailty. The authors also conducted some machine learning methods (recursive partitioning, generalised boosted, support vector machines and random forest) for the robustness test. They found that logistic regression provided better accuracy and specificity. Although the researchers suggest the applicability of the model in national and international settings, we did not find any evidence of the validation of the FRS. One possible reason is that the authors used an older version of the international classification of disease (i.e., ICD-9) codes and their frailty risk scores have not been benchmarked against previously established measures. The authors acknowledged that future development needs to include other clinical information such as gait speed, grip strength, presence of tremor and mood disorders, which were used in other frailty measures such as the Fried frailty phenotype (Fried et al., 2001).

This study also developed the "worsening risk predictive model", which predicts the probability of becoming frail within one year among non-frail individuals. The worsening risk model includes 26 clinical and socio-economic variables, all of which were overlapped

with predictors in the FRS model. This model was only validated internally. The model revealed good discriminative ability with the respective C-statistics for the training set and the test set of 0.875 and 0.879.

3.4 Preoperative Frailty Index (pFI)

The next study was developed in Canada by McIsaac et al. (2018), who developed and validated a generalized preoperative frailty index (pFI). The index was developed using 30 variables, constructed from various linked health data. One noticeable difference of this index is that nine variables included a choice of 0.5 points (apart from the possible 0 or 1 point for each variable) to capture an intermediate level of condition. The index is presented as the ratio of total score divided by 30 variables. The index was found to be a robust and significant predictor of one-year mortality, with an odds ratio of 2.85 for a 0.1 point increase in the frailty index. The results are consistent in those obtained from the validation cohort and sub-groups (e.g., joint replacement, cancer). However, the frailty index developed in this study has not been benchmarked against other measures or independently validated using external data. One possible reason, as the authors acknowledged, is that data for some of the 30 variables selected to develop this index (e.g., home-support care data) may not be available in other contexts.

3.5 Dr Forster Global Frailty Score (FGFS)

The last study selected was an international collaboration from 34 hospitals in nine countries by Soong et al. (2019), who developed and validated the Dr Forster Global Frailty Score (FGFS). The FGFS was estimated using a two-step process: 1) a weighted frailty score was estimated using parameters from logistic regressions for each frailty syndrome group; 2) the weighted frailty score was included in a logistic regression to predict three outcomes of

interest: long hospital stay (upper quartile), 30-day non-elective readmission, and in-hospital mortality. The external validation was conducted using the Hospital Episodes Statistics data set from England. The FGFS was a significant predictor of all three outcomes, after controlling for age, gender and countries. The AUC for non-elective a long hospital stay, 30-day readmission and in-hospital mortality was 0.68, 0.63 and 0.66, respectively. The main limitation of this study was the considerable variations in the frequency of frailty syndrome coding across countries, which may affect its reliability and generalisability.

4. Discussion

All frailty measurements using administrative health data show significant association with adverse health outcomes such as long hospital stay, readmission and mortality. However, the prediction of these outcomes is of most use for tertiary prevention (i.e., preventing complications caused by frailty progression (Dekker & Sibai, 2001)) to manage frailty or other morbidities (Sternberg, Schwartz, Karunananthan, Bergman, & Mark Clarfield, 2011). A significant element for the clinical usefulness of any frailty measure may lie in their ability for the early detection of frailty for primary prevention (e.g., to control risk factors) and secondary prevention (e.g., exercise, nutrition). Among the studies reviewed, only Bertini et al. (2018) developed such a prediction model to estimate the risk of becoming frail within one year. We argue that the use of routine primary healthcare data to predict outcome such as hospitalisation (Clegg et al., 2016) would also contribute to earlier detection of frailty and hance prediction of adverse outcomes than otherwise would have occurred. A study from Japan, the country with the most severe ageing population issue (Muramatsu & Akiyama, 2011), showed the effectiveness of preventing frailty at the primary care level (Shinkai et al., 2016).

Apart from administrative health data, some studies have developed frailty measurement tools using other data sources such as insurance claims (Kim & Schneeweiss, 2014; Segal et al., 2017) and inputs from motion-detection devices (Martini et al., 2018). Since frailty is a complex condition that can be measured from various approaches (e.g., clinical, biological, psychological, social), future studies may seek to combine administrative health data with other data sources (e.g., government linkage data, social media outlets) to generate a holistic measure of frailty.

While contemporary literature demonstrated that administrative health data could be used to measure frailty, we need to assess the clinical usefulness of these automated frailty measurement tools. For example, future studies should evaluate the effects of automated frailty measurements on timely and effective healthcare planning. Potential side effects of automated frailty measurements (e.g., increase demand for clinical assessments for those with high automated frailty risk scores), should also be investigated. These assessments will provide crucial inputs for policymakers to decide on the application of automated frailty measurement tools into practices.

We hypothesise that the automated measurement and prediction of frailty will be cost-effective compared with clinical assessment approaches, however, this hypothesis needs to be formally tested. We anticipate that cost-effectiveness studies of automating frailty measurement and prediction relative to traditional approaches will be implemented in the future. To what extent will predictive frailty models affect change in clinical practice and subsequently achieve improved patient health outcomes? A direction for future studies may well be to investigate the extent to which automated frailty measurement tools affect medical decision-making. Particularly, future studies should use frailty measurement tools to identify a refined target population for clinical trials of interventions, with a clear translational pathway to implementation. In addition, future studies should conduct economic modelling of

the impact of implementing interventions for people with different degrees of frailty based on the use of these tools.

The main limitation of this review is the small number of selected studies. We believe that the trend of using administrative data to measure frailty is increasing and an updated review in the future will find more eligible studies.

5. Conclusions and Implications

To our best knowledge, this study is the first systematic review of the selected topic. To date, there are only five studies that have developed and validated frailty measurement tools using administrative health data. All the five tools have a strong predictive power of common adverse health outcomes, including mortality, readmission and long hospital stay. There is considerable overlap in the predictors used in the five studies and those of the Fried phenotype (Fried et al., 2001) and the Rockwood scale (Rockwood et al., 2005). Two of the five selected studies have been independently validated in Australia, Switzerland and Canada.

Based on the finding of this systematic review, we expect that more validations of the UK studies will be conducted. Also, for automated frailty measurement tools to be implemented in practice, their clinical usefulness and cost-effectiveness will need to be examined thoroughly in the future.

Funding

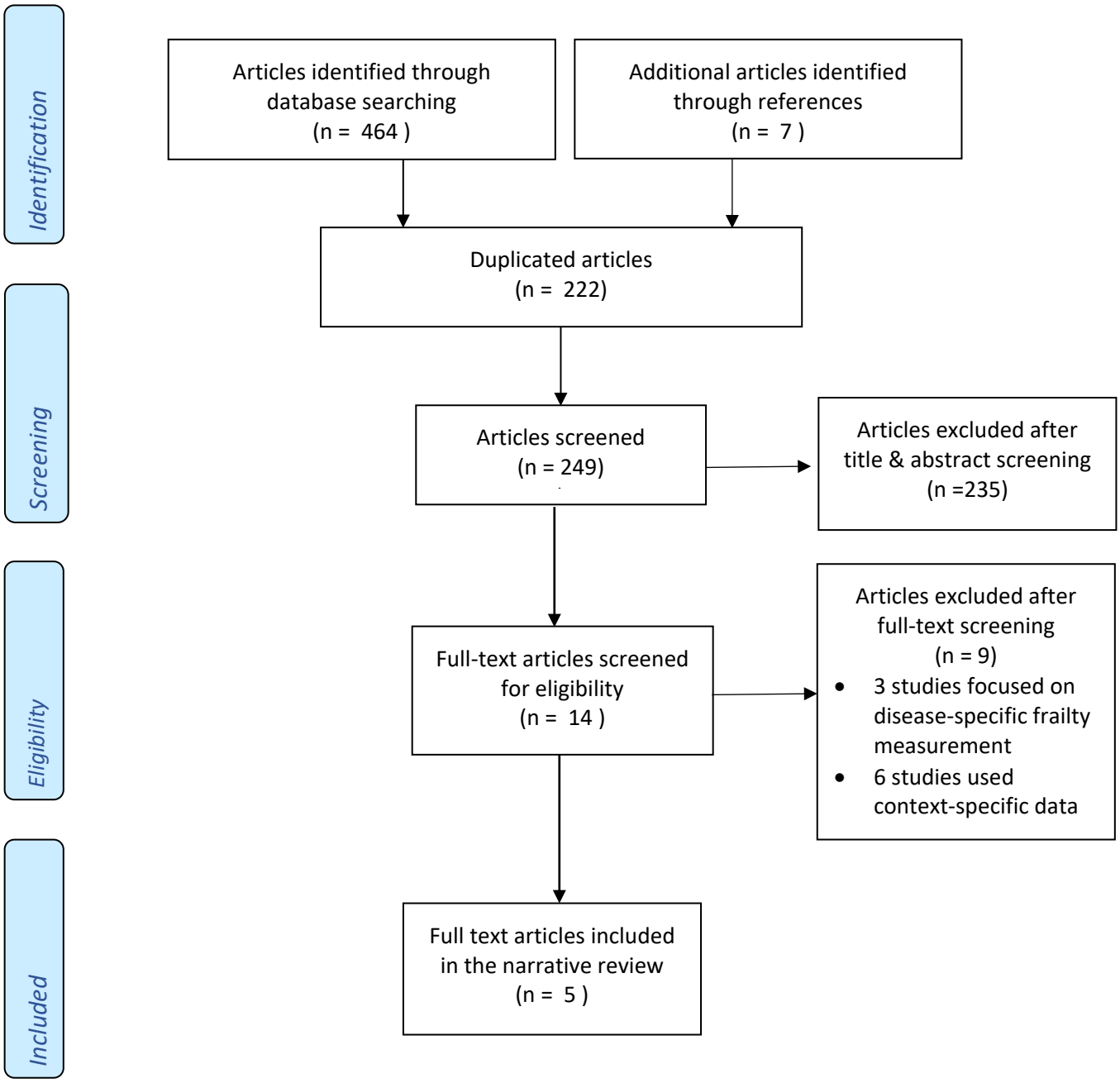
SN received a New Researcher Grant (No. 51594) from Griffith University to conduct a research project on the validation of the hospital frailty risk score in Australia. This systematic review is a part of that project.

Conflicts of Interest

None declared

Supplementary data:

Figure 1. PRISMA flow diagram



PROBAS Risk of bias assessment of included studies

Authors (year)	Participants	Predictors	Outcome	Analysis	Overall
Risk of bias					
Clegg et al. (2016)	+	+	+	+	+
Gilbert et al. (2018)	+	+	+	+	+
Betini et al. (2018)	+	+	+	-	-
McIsaac et al. (2019)	+	+	+	-	-
Soong et al. (2019)	+	+	+	+	+
Applicability					
Clegg et al. (2016)	+	+	+	+	+
Gilbert et al. (2018)	+	+	+	+	+
Betini et al. (2018)	+	+	+	-	-
McIsaac et al. (2019)	+	+	+	-	-
Soong et al. (2019)	+	+	+	-	-

Note: * + indicates low risk of bias/low concern regarding applicability; − indicates high risk of bias/high concern regarding applicability; and? indicates unclear risk of bias/unclear concern regarding applicability.

Judgement criteria for Risk of bias and Applicability

Risk of bias	
Low	• If all domains were rated low risk of bias. • If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to high risk of bias. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High	• If at least one domain is judged to be at high risk of bias.
Unclear	• If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.
Applicability	

Low	• If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have low concerns regarding applicability.
High	• If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have high concerns regarding applicability.
Unclear	• If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have unclear concerns regarding applicability overall.

CHARM checklist

Domain	Key items	Reported on page #				
		Clegg et al. (2016)	Gilbert et al. (2018)	Bertini et al. (2018)	McIsaac et al. (2019)	Soong et al. (2019)
Sources of Data	Source of data (e.g., cohort, case-control, randomized trial participants, or registry data)	354	1776-1777	724	103	2
Participants	Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria)	NA	NA	NA	NA	NA
	Participant description	354	1776	NA	103	2
	Details of treatments received, if relevant	NA	NA	NA	NA	NA
	Study dates	354	1776	729	103	2
Outcome(s) to be Predicted	Definition and method for measurement of outcome	354	1777	724	103	2
	Was the same outcome definition (and method for measurement) used in all patients?	Y	Y	Y	Y	Y
	Type of outcome (e.g., single or combined endpoints)	354	1777	724	103	2
	Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?	Y	Y	Y	Y	Y
	Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?	Y	Y	Y	Y	Y
	Time of outcome occurrence or summary of duration of follow-up	NA	NA	NA	NA	NA
Candidate predictors	Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)	354	1777	727	104	3
	Definition and method for measurement of candidate predictors	354	1777	727	104	3
	Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)	NA	NA	NA	NA	NA
	Were predictors assessed blinded for outcome, and for each other (if relevant)?	NA	NA	NA	NA	NA
	Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or categorised)	NA	NA	NA	NA	NA
Sample size	Number of participants and number of outcomes/events	354	1777	729	103	4
	Number of outcomes/events in relation to the number of	NA	NA	NA	NA	NA

	candidate predictors (Events Per Variable)					
Missing data	Number of participants with any missing value (include predictors and outcomes)	NA	NA	NA	104	NA
	Number of participants with missing data for each predictor	NA	NA	NA	NA	NA
Model development	Handling of missing data (e.g., complete-case analysis, imputation, or other methods)	NA	NA	NA	104	NA
	Modelling method (e.g., logistic, survival, neural network, or machine learning techniques)	354	1776-1777	732	103	3
	Modelling assumptions satisfied	NA	NA	NA	NA	NA
	Method for selection of predictors **for inclusion** in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)	354-355	1776-1777	728-729	104	3
	Method for selection of predictors **during multivariable modelling** (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)	NA	NA	NA	NA	NA
	Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)	1777	NA	NA	NA	NA
Model performance	Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals	356-357	1778	730	105	4
	Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used	356-357	1778	730	NA	NA
Model evaluation	Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators)	354-355	1780	730	103	NA
	In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)	NA	NA	NA	NA	NA

Results	Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)	356-357	1778-1780	730-731	105	5-6
	Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance	NA	NA	NA	NA	NA
	Comparison of the distribution of predictors (including missing data) for development and validation datasets	NA	NA	NA	NA	NA
Interpretation and Discussions	Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)	357	1780-1781	732-733	105-106	6-8
	Comparison with other studies, discussion of generalizability, strengths and limitations.	358	1781	733	107	8

MOOSE Checklist for Meta-analyses of Observational Studies

Item No	Recommendation	Reported on Page No
Reporting of background		
1	Problem definition	1
2	Hypothesis statement	2
3	Description of study outcome(s)	1
4	Type of exposure or intervention used	NA
5	Type of study designs used	2
6	Study population	4
Reporting of search strategy		
7	Qualifications of searchers (eg, librarians and investigators)	Title page
8	Search strategy, including time period included in the synthesis and key words	2
9	Effort to include all available studies, including contact with authors	3
10	Databases and registries searched	2
11	Search software used, name and version, including special features used (eg, explosion)	2
12	Use of hand searching (eg, reference lists of obtained articles)	2
13	List of citations located and those excluded, including justification	15
14	Method of addressing articles published in languages other than English	3
15	Method of handling abstracts and unpublished studies	3
16	Description of any contact with authors	2
Reporting of methods		
17	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested	3
18	Rationale for the selection and coding of data (eg, sound clinical principles or convenience)	3
19	Documentation of how data were classified and coded (eg, multiple raters, blinding and interrater reliability)	3
20	Assessment of confounding (eg, comparability of cases and controls in studies where appropriate)	NA
21	Assessment of study quality, including blinding of quality assessors, stratification or regression on possible predictors of study results	NA
22	Assessment of heterogeneity	NA
23	Description of statistical methods (eg, complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated	NA
24	Provision of appropriate tables and graphics	5-6
Reporting of results		
25	Graphic summarizing individual study estimates and overall estimate	NA
26	Table giving descriptive information for each study included	7
27	Results of sensitivity testing (eg, subgroup analysis)	NA

28	Indication of statistical uncertainty of findings	NA

Item No	Recommendation	Reported on Page No
Reporting of discussion		
29	Quantitative assessment of bias (eg, publication bias)	8-9
30	Justification for exclusion (eg, exclusion of non-English language citations)	8-9
31	Assessment of quality of included studies	8-9
Reporting of conclusions		
32	Consideration of alternative explanations for observed results	14
33	Generalization of the conclusions (ie, appropriate for the data presented and within the domain of the literature review)	14
34	Guidelines for future research	15
35	Disclosure of funding source	16

From: Stroup DF, Berlin JA, Morton SC, et al, for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. Meta-analysis of Observational Studies in Epidemiology. A Proposal for Reporting. *JAMA*. 2000;283(15):2008-2012. doi: 10.1001/jama.283.15.2008.

PRISMA checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	2
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	3
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	3
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	3-4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	3-4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	4
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	4
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	5-6

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	4
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	4
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	5-6
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	5-6
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	5-6
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	5-15
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	8-9
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14-15
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	16

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

References

Abbasi, M., Khera, S., Dabravolskaj, J., Vandermeer, B., Theou, O., Rolfson, D., et al. (2019). A cross-sectional study examining convergent validity of a frailty index based on electronic medical records in a Canadian primary care program. *BMC Geriatrics, 19*(1), 109.

Ambagtsheer, R. C., Beilby, J., Dabravolskaj, J., Abbasi, M., Archibald, M. M., & Dent, E. (2019). Application of an electronic frailty index in Australian primary care: Data quality and feasibility assessment. *Aging Clinical Experimental Research, 31*(5), 653-660.

Bertini, F., Bergami, G., Montesi, D., Veronese, G., Marchesini, G., & Pandolfi, P. (2018). Predicting frailty condition in elderly using multidimensional socioclinical databases. *Proceedings of the IEEE, 106*(4), 723-737. doi:10.1109/jproc.2018.2791463

Boyd, P. J., Nevard, M., Ford, J. A., Khondoker, M., Cross, J. L., & Fox, C. (2019). The electronic frailty index as an indicator of community healthcare service utilisation in the older population. *Age and Ageing, 48*(2), 273-277. doi:10.1093/ageing/afy181

Brundle, C., Heaven, A., Brown, L., Teale, E., Young, J., West, R., et al. (2019). Convergent validity of the electronic frailty index. *Age and Ageing, 48*(1), 152-156. doi:10.1093/ageing/afy162

Clegg, A., Bates, C., Young, J., Ryan, R., Nichols, L., Ann Teale, E., et al. (2016). development and validation of an electronic frailty index using routine primary care electronic health record data. *Age and Ageing, 45*(3), 353-360.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Dekker, G., & Sibai, B. (2001). Primary, secondary, and tertiary prevention of Pre-Eclampsia. *The Lancet, 357*(9251), 209-215.

Eckart, A., Hauser, S. I., Haubitz, S., Struja, T., Kutz, A., Koch, D., et al. (2019). Validation of the hospital frailty risk score in a tertiary care hospital in Switzerland: Results of a prospective, observational study. *BMJ Open, 9*(1), e026923.

Epstein, D. (2007). Impact factor manipulation. *The Write Stuff, 16*(3), 133-134.

Evans, S. J., Sayers, M., Mitnitski, A., & Rockwood, K. (2014). The risk of adverse outcomes in hospitalized older patients in relation to a frailty index based on a comprehensive geriatric assessment. *Age and Ageing, 43*(1), 127-132.

Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., et al. (2001). Frailty in older adults: Evidence for a phenotype. *56*(3), M146-M157.

Gilbert, T., Neuburger, J., Kraindler, J., Keeble, E., Smith, P., Ariti, C., et al. (2018). Development and validation of a hospital frailty risk score focusing on older people in acute care settings using electronic hospital records: An observational study. *The Lancet, 391*(10132), 1775-1782. doi:10.1016/S0140-6736(18)30668-8

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Kim, D. H., & Schneeweiss, S. (2014). Measuring frailty using claims data for pharmacoepidemiologic studies of mortality in older adults: evidence and recommendations. *Pharmacoepidemiology and Drug Safety, 23*(9), 891-901.

Lansbury, L. N., Roberts, H. C., Clift, E., Herklots, A., Robinson, N., & Sayer, A. A. (2017). Use of the electronic frailty index to identify vulnerable patients: A pilot study in primary care. *British Journal General Practice, 67*(664), e751-e756.

Marshall, A., Nazroo, J., Tampubolon, G., & Vanhoutte, B. (2015). Cohort differences in the levels and trajectories of frailty among older people in England. *Journal of Epidemiology and Community Health, 69*(4), 316-321.

Martini, C., Barla, A., Odone, F., Verri, A., Cella, A., Rollandi, G. A., et al. (2018). Data-driven continuous assessment of frailty in older people. *Frontiers in Digital Humanities, 5*, 6.

McAlister, F., & van Walraven, C. (2019). External validation of the hospital frailty risk score and comparison with the hospital-patient one-year mortality risk score to predict outcomes in elderly hospitalised patients: A retrospective cohort study. *BMJ Quality and Safety, 28*(4), 284-288. doi:10.1136/bmjqs-2018-008661

McIsaac, D. I., Wong, C. A., Huang, A., Moloo, H., & van Walraven, C. (2018). Derivation and validation of a generalizable preoperative frailty index using population-based health administrative data. *Annals of Surgery*. doi:10.1097/sla.0000000000002769

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., et al. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine, 11*(10), e1001744.

Morley, J. E., Vellas, B., Van Kan, G. A., Anker, S. D., Bauer, J. M., Bernabei, R., et al. (2013). Frailty consensus: A call to action. *Journal of the American Medical Directors Association, 14*(6), 392-397.

Muramatsu, N., & Akiyama, H. (2011). Japan: super-aging society preparing for the future. *The Gerontologist, 51*(4), 425-432.

Rockwood, K. (1999). A brief clinical instrument to classify frailty in elderly people. *Lancet, 353*(9148), 205-206.

Rockwood, K., Song, X., MacKnight, C., Bergman, H., Hogan, D. B., McDowell, I., et al. (2005). A global clinical measure of fitness and frailty in elderly people. *Canadian Medical Association Journal, 173*(5), 489-495.

Rolfson, D. B., Majumdar, S. R., Tsuyuki, R. T., Tahir, A., & Rockwood, K. (2006). Validity and reliability of the edmonton frail scale. *Age and Ageing, 35*(5), 526-529.

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.

Shinkai, S., Yoshida, H., Taniguchi, Y., Murayama, H., Nishi, M., Amano, H., et al. (2016). Public health approach to preventing frailty in the community and its effect on healthy aging in Japan. *Geriatrics & Gerontology International, 16*, 87-97.

Soong, J. T. Y., Kaubryte, J., Liew, D., Peden, C. J., Bottle, A., Bell, D., et al. (2019). Dr Foster global frailty score: An international retrospective observational study developing and validating a risk prediction model for hospitalised older persons from administrative data sets. *BMJ Open, 9*(6), e026759. doi:10.1136/bmjopen-2018-026759

Sternberg, S. A., Schwartz, A. W., Karunananthan, S., Bergman, H., & Mark Clarfield, A. J. J. o. t. A. G. S. (2011). The Identification of frailty: A systematic literature review. *Journal of the American Geriatrics Society, 59*(11), 2129-2138.

Steyerberg, E. W., & Harrell, F. E. (2016). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology, 69*, 245-247.

Stow, D., Matthews, F. E., & Hanratty, B. (2018). Frailty trajectories to identify end of life: A longitudinal population-based study. *BMC Medicine, 16*(1). doi:10.1186/s12916-018-1148-x

Straup, D., Berlin, J., Morton, S., Olkin, I., Williamson, G., Rennie, D., et al. (2000). For the meta-analysis of observational studies in epidemiology (MOOSE) group: Meta-analysis of observational studies in epidemiology. A proposal for reporting. *Journal of American Medical Association, 283*, 2008-2012.

Turner, G., & Clegg, A. (2014). Best practice guidelines for the management of frailty: A British geriatrics society, age UK and royal college of general practitioners report. *Age and Ageing, 43*(6), 744-747.

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., et al. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51-58.