PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Dr Foster Global Frailty Score: An international retrospective observational study developing and validating a risk prediction model for hospitalised older persons from administrative datasets.
AUTHORS	Soong, John; Kaubryte, Jurgita; Liew, Danny; Peden, Carol; Bottle, Alex; Bell, Derek; Cooper, Carolyn; Hopper, Adrian

VERSION 1 - REVIEW

REVIEWER	Andrew Clegg University of Leeds, UK
REVIEW RETURNED	19-Oct-2018

GENERAL COMMENTS	The authors describe a retrospective cohort study to develop and validate a risk prediction model based on frailty syndromes using routine hospital data from 1.4M patients across nine countries. The study is novel, particularly the international element. I have a number of comments for consideration Page 4 line 15 Why only those aged >75? Frailty is prevalent in hospitalised patients who are younger than this, so this has limited the validity of findings to this age group, and findings cannot automatically be extrapolated to younger age groups (e.g. >65). What will the advice of the team be when clinicians naturally wish to use the score with younger populations? Page 16 line 42 The team state that the GFS has significant predictive capacity beyond that of comorbidity. I think this needs clarifying. They have just demonstrated that it has lower predictive capacity that comorbidity (Elixhauser), when analysed separately, with extremely minimal (negligible) increase in discrimination when combined. I accept that table 4 does support to some extent, but I think that the authors need to be clearer about the clinical interpretation of any increase when both scores are in the model. Given the size of the dataset, any findings are likely to be statistically significant, so there needs to be a clearer interpretation of the clinical relevance of the findings. Page 19 line 5 This appears again in the conclusion, but I do need convincing that this is the case here. Although I accept that table 4 might

	support this statement to some extent, the evidence in table 5 does not support this conclusion as any increase in discrimination is negligible and probably clinically meaningless.

REVIEWER	Simon Conroy University of Leicester, UK Author of paper in same field, active researcher in the field
REVIEW RETURNED	10-Feb-2019

GENERAL COMMENTS	The title contains the name of a commercial entity – is this permitted? It is certainly distasteful Methods How were the Global Comparator hospitals selected? What is the role of selection bias? Why restrict inclusion to 24 hours+? With the increasing impact of Same Day Emergency Care in the UK, significant proportion of older people will have been managed in ambulatory setting, and so their outcomes (which are plausibly systematically and clinically significantly different to admitted patients) are not taken account of, thus limiting generalisability. Frailty codes are likely to be under-reported in some situations and over-reported in others – where is the underpinning examination of the face/content/construct validity of these codes that allows confidence to be placed in their meaning? 'Within the Global Comparators dataset, 30 models were created' – why, what, how? There appears to be a fundamental difference in how the investigators have created a frailty score. They have summed the number of frailty syndromes, in contrast to the more conventional frailty index/score methodology from Searle et al1 which has been using in a large number of studies globally. Moreover, most clinicians would use frailty to identify the risk of developing the hyperacute geriatric syndromes such as delirium, with high risk prompting interventions to reduce the risk of developing frailty syndromes. Why did they adopt this methodology? The coding variation between and within countries is an issue – it is perhaps impossible to tell if this variation is clinically driven, arises from a result of different system configurations, or coding error/differences – this a fundamental concern that is not robustly addressed in this paper Can the authors clarify of the derivation and validation cohorts were separate temporally and spatially? Results The odds ratios shown in table 7 are statistically significant given the large dataset, but do not appear to be clinically meaningful. Discussion 'Our study found that frailty syndromes are feasibly coded' – what does this mean? 'The Dr Foster Global Frailty Score has significant predictive capacity beyond that of other known predictors of poor outcome in older persons, such as co-morbidity and chronological age.' This over-inflates the importance of the findings and does not reflect the literature. The odds ratios for mortality, length of stay and readmission were no greater than 1.1. Here are a few of many

	papers showing stronger clinically significant predictive power using frailty2-7. ‘The ORs and predictive capacity in the validation cohort were generally lower than the derivation cohort, but are in keeping with other risk prediction models for older persons within the English secondary care administrative data’ – this is not accurate reflection of previous studies which describe a gradient of increasing risk of adverse outcomes with increasing frailty risk scores – which the present paper does not appear to have done. This is important as the whole point of the frailty concept is to unpick the heterogeneity that age and comorbidities alone do not address fully. References 1. Searle SD, Mitnitski A, Gahbauer EA, et al. A standard procedure for creating a frailty index. BMC geriatrics 2008;8:24. 2. Gilbert T, Neuburger J, Kraindler J, et al. 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 2018;391(10132):1775-82. 3. Hubbard RE, Peel NM, Samanta M, et al. Frailty status at admission to hospital predicts multiple adverse outcomes. Age and Ageing 2017:1-6. 4. Hewitt J, Moug SJ, Middleton M, et al. Prevalence of frailty and its association with mortality in general surgery. American Journal of Surgery 2015;209(2):254-9. 5. Basic D, Shanley C. Frailty in an older inpatient population: using the clinical frailty scale to predict patient outcomes. Journal of Aging and Health 2015;27:670-85. 6. Evans SJ, Sayers M, Mitnitski A, et al. The risk of adverse outcomes in hospitalized older patients in relation to a frailty index based on a comprehensive geriatric assessment. Age and Ageing 2014;43:127-32. 7. Wou F, Gladman JRF, Bradshaw L, et al. The predictive properties of frailty-rating scales in the acute medical unit. Age and Ageing 2013;42:776-81.

VERSION 1 – AUTHOR RESPONSE

Reviewers' Comments to Author:

Reviewer: 1

Reviewer Name: Andrew Clegg

Institution and Country: University of Leeds, UK

Please state any competing interests or state ‘None declared’: None declared

The authors describe a retrospective cohort study to develop and validate a risk prediction model based on frailty syndromes using routine hospital data from 1.4M patients across nine countries. The study is novel, particularly the international element.

Many thanks for your kind and accurate reflection of this study.

I have a number of comments for consideration

Page 4 line 15

Why only those aged >75? Frailty is prevalent in hospitalised patients who are younger than this, so this has limited the validity of findings to this age group, and findings cannot automatically be extrapolated to younger age groups (e.g. >65). What will the advice of the team be when clinicians naturally wish to use the score with younger populations?

Many thanks for your comments. Previous study within English national datasets (Hospital Episode Statistics) suggest a much larger frequency of coding for frailty syndromes from age groups 75 and above onwards (Soong J, Poots A, Scott S, et al Quantifying the prevalence of frailty in English hospitals BMJ Open 2015;5:e008456. doi: 10.1136/bmjopen-2015-008456; Appendix 3 Supp Data), though a risk prediction model was built on a population that included a younger age group (>65 years) as well (Soong J, Poots AJ, Scott S, et al Developing and validating a risk prediction model for acute care based on frailty syndromes BMJ Open 2015;5:e008457. doi: 10.1136/bmjopen-2015-008457). Given these findings, for this study, we decided to focus on a more elderly population as it seemed to have the greatest burden. We note a similar approach was taken in a recent study (Gilbert T, Neuburger J, Kraindler J, Keeble E, Smith P, Ariti C, et al. 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. 2018;391(10132):1775-82.). Thank you for highlighting this. The text ‘The score was developed on hospitalised populations of age ≥ 75 years as the majority of frail older persons fall within this age-group, particularly in Western Europe. This score is therefore not validated in those who fall below 75 years of age.’ has been added to the limitations section.

Page 16 line 42

The team state that the GFS has significant predictive capacity beyond that of comorbidity. I think this needs clarifying. They have just demonstrated that it has lower predictive capacity that comorbidity (Elixhauser), when analysed separately, with extremely minimal (negligible) increase in discrimination when combined. I accept that table 4 does support to some extent, but I think that the authors need to be clearer about the clinical interpretation of any increase when both scores are in the model. Given the size of the dataset, any findings are likely to be statistically significant, so there needs to be a clearer interpretation of the clinical relevance of the findings.

Many thanks for your comments. We have removed the text ‘The Dr Foster Global Frailty Score has significant predictive capacity beyond that of other known predictors of poor outcome in older persons, such as co-morbidity and chronological age.’ to more accurately reflect findings.

Page 19 line 5

This appears again in the conclusion, but I do need convincing that this is the case here. Although I accept that table 4 might support this statement to some extent, the evidence in table 5 does not support this conclusion as any increase in discrimination is negligible and probably clinically meaningless.

Many thanks for your comments. We have removed the text ‘It has predictive power beyond that of the Elixhauser co-morbidity score within these datasets.’ to more accurately reflect findings

Reviewer: 2

Reviewer Name: Simon Conroy

Institution and Country: University of Leicester, UK

Please state any competing interests or state ‘None declared’: Author of paper in same field, active researcher in the field

The title contains the name of a commercial entity – is this permitted? It is certainly distasteful

Many thanks for your comment. The Global Comparators programme at Dr Foster® was a not-for-profit international hospital collaborative which ran from 2011-2017. Amongst it’s many activities, it focused on pooling and benchmarking data, knowledge-sharing networks and health services research to better understand variations in outcomes and disseminate international best practice. The development of this score was a not-for-profit analysis. Dr Foster® also has a separate commercial entity.

Methods

How were the Global Comparator hospitals selected? What is the role of selection bias?

Many thanks for the comment. We have included the text ‘The hospitals that contributed data to the Global Comparators dataset were mainly large academic centres with reputations of clinical excellence. As such, the quality of coding and patient outcomes represented may not be representative of other institutions.’ within the limitations section

Why restrict inclusion to 24 hours+? With the increasing impact of Same Day Emergency Care in the UK, significant proportion of older people will have been managed in ambulatory setting, and so their outcomes (which are plausibly systematically and clinically significantly different to admitted patients) are not taken account of, thus limiting generalisability.

Many thanks for comment. Hospital administrative data codes for indicating admission to hospital varies from country to country. Day case admissions for observation, investigations or procedures are common in many countries and would be included in patients admitted for less than 24 hours, and less reflective of the population of interest. As this study focuses on hospitalised patients, the inclusion criteria were adjusted to reflect this. The text ‘This was to exclude records with inadequate quality data, and patients admitted into observations units or day-case attendances. Overall, 0.17% of data were missing within the derivation dataset.’ within the Study Population part of the Methods section had previously been added to reflect this. We have added the text ‘ Additionally, the study focused on hospitalised patients of ≥24 hours to exclude patients admitted to observational units, for investigations or procedures. There is increasing acceptance for the acute medical management of older persons in an ambulatory setting. This methodology will exclude same-day discharges, limiting generalisability.’ to the limitations section. Alternatively, frailty contributes to patients having the longest lengths of stay, highest readmission rates, and highest rate of use of long-term care after discharge. (Sager MA, Franke T, Inouye SK et al. Functional outcomes of acute medical illness and hospitalization in older persons. Arch Intern Med 1996;156(6):645–52.), so this ambulatory population may be small in the international context.

Frailty codes are likely to be under-reported in some situations and over-reported in others – where is the underpinning examination of the face/content/construct validity of these codes that allows confidence to be placed in their meaning?

Many thanks for your comment. The variables chosen have undergone content and face validity through a national (UK) Delphi consensus process (Soong JTY, Poots AJ, Bell D Finding consensus on frailty assessment in acute care through Delphi method BMJ Open 2016;6:e012904. doi: 10.1136/bmjopen-2016-012904). This methodology of coding frailty has had previous study exploring coding frequency and predictive power in English national hospital administrative data (Soong J, Poots A, Scott S, et al Quantifying the prevalence of frailty in English hospitals BMJ Open 2015;5:e008456. doi: 10.1136/bmjopen-2015-008456 and Soong J, Poots AJ, Scott S, et al Developing and validating a risk prediction model for acute care based on frailty syndromes BMJ Open 2015;5:e008457. doi: 10.1136/bmjopen-2015-008457).

'Within the Global Comparators dataset, 30 models were created' – why, what, how?

Many thanks for your comment. We have amended the text to read 'Within the Global Comparators dataset, 30 separate regression models were undertaken, to account for admission status, frailty, Elixhauser co-morbidity and combination of frailty and Elixhauser for the three outcomes above.' We have also included a summary diagram, figure 1.

There appears to be a fundamental difference in how the investigators have created a frailty score. They have summed the number of frailty syndromes, in contrast to the more conventional frailty index/score methodology from Searle et al1 which has been using in a large number of studies globally.

Many thanks for the comment. The score was developed from summation of weights from the natural log value of the Odds Ratios developed from multivariable regression, and not the summation of number of frailty syndromes. The text 'A two-step process for each outcome was utilised to model the frailty and comorbidity scores. First, binary logistic regression was utilised to ascertain odds ratios (ORs) for each frailty syndrome group and each outcome, within the population subgroups separately (elective and non-elective). The natural log of OR (ln OR) was used to create weights for each frailty syndrome group, using the smallest ln OR as reference (weighted 1.0). Secondly, the summation of the weights for each frailty syndrome group was utilised to create a frailty score.' within the Risk Models part of the Methods section had previously been added to reflect this. This methodology is well described and has been used to generate other risk scores, such as the Elixhauser Co-morbidity Score (Bottle A, Aylin P. Comorbidity scores for administrative data benefited from adaptation to local coding and diagnostic practices. J Clin Epidemiol. 2011;64(12):1426-33).

The procedure described by Searle et al. describes a specific methodology to create a Frailty Index, with very distinct data requirements, according to the operationalised model of accumulated deficits. For instance, each variable should be associated with health status, it's prevalence must increase with age but not saturate too early, the variables should cover a range of systems and, if the frailty index is to be used serially on the same individuals, the variables need to remain the same. It remains to be seen if secondary use of Diagnostic Codes in structured administrative data (e.g. ICD-10 or ICD-9) from hospitalised older patients may fulfil the criteria above, and thus successfully operationalised into a frailty index. We note the recent publication of the Hospital Frailty Risk Score (Gilbert T, Neuburger J, Kraindler J, Keeble E, Smith P, Ariti C, et al. 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. 2018;391(10132):1775-82.), which also does not utilise the Frailty Index methodology.

Moreover, most clinicians would use frailty to identify the risk of developing the hyperacute geriatric syndromes such as delirium, with high risk prompting interventions to reduce the risk of developing frailty syndromes. Why did they adopt this methodology?

Many thanks for your comment. The results of a national (UK) Delphi consensus (Soong JTY, Poots AJ, Bell D Finding consensus on frailty assessment in acute care through Delphi method BMJ Open 2016;6:e012904. doi: 10.1136/bmjopen-2016-012904) exploring indicators of frailty in the acute care setting revealed that frailty syndromes and indicators of high utilisation (e.g. non-elective hospital readmission)gained consensus as useful and appropriate measures of frailty after two rounds. Additionally, UK National guidelines for care of the older persons in the acute care setting (Acute Care Toolkit 3. Acute medical care for frail older people. London: Royal College of Physicians, 2012; Quality care for older people with urgent and emergency care needs(the Silver Book) and Recognising Frailty, Good Practice Guidelines, British Geriatric Society) have recommended frailty syndromes as a potential method of recognizing frailty in the acute care or hospital setting. The text 'The approach of targeting frailty syndromes for hospitalised patients has support in existing literature, and in keeping with national standards bodies recommendations in the UK' has been added to the discussion section to better reflect this, and references added.

The coding variation between and within countries is an issue – it is perhaps impossible to tell if this variation is clinically driven, arises from a result of different system configurations, or coding error/differences – this a fundamental concern that is not robustly addressed in this paper

Many thanks for comment. The text ' The variability in frequency of coding of frailty syndromes across countries may limit reliability and generalisability, although the country of origin was accounted for in the multivariable regression. Further subgroup analysis in countries with similar frequency of coding, or hierarchical regression to account for clusters, may be the next step.' had previously been included in the limitations section to reflect this. This variation in coding frailty utilising this methodology is further explored and expanded in a planned publication.

Can the authors clarify of the derivation and validation cohorts were separate temporally and spatially?

Many thanks for comment. The text 'A small proportion of the validation cohort may have been duplicated from the derivation cohort (eight hospitals in calendar year 2013). However, using national data from several calendar years minimises the effect of this overlap.' had been previously included in the limitations section to reflect this.

Results

The odds ratios shown in table 7 are statistically significant given the large dataset, but do not appear to be clinically meaningful.

Many thanks for your comment. We note that none of our models have predictive powers suitable for clinical risk prediction at the patient's bedside (AUC >0.80). We note similar findings from the Hospital Frailty Risk Score (Gilbert T, Neuburger J, Kraindler J, Keeble E, Smith P, Ariti C, et al. 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. 2018;391(10132):1775-82.): 'The Hospital Frailty Risk Score discriminated weakly between individuals with different outcomes within hospitals; the c statistics were 0·60 for 30-day mortality, 0·68 for a long hospital stay, and 0·56 for 30-day readmission. The inclusion of patients' other characteristics (age, sex, deprivation, admission history, and comorbidity) improved discrimination to 0·69 for 30-day mortality, 0·73 for long hospital stay, and 0·61 for readmission.' There may be a ceiling effect with methodologies utilising standardised Diagnostic Codes from structured fields within administrative data to code for frailty. We have added the text 'However, we note that our model's predictive powers are not suitable for clinical risk prediction at the patient's bedside (AUC >0.80).' in the Discussion section to better reflect this.

Discussion

'Our study found that frailty syndromes are feasibly coded' – what does this mean?

Many thanks for your comment. We have amended to text to read 'Our study found that frailty syndromes are coded with variable frequency within a large (N≈1.3m) international dataset of hospitalised older persons (aged over 75 years) utilising readily available administrative data, with Falls & Fractures and Dementia & Delirium being the most frequently coded syndromes.'

'The Dr Foster Global Frailty Score has significant predictive capacity beyond that of other known predictors of poor outcome in older persons, such as co-morbidity and chronological age.' This over-inflates the importance of the findings and does not reflect the literature.

Many thanks for your comments. We have removed the text 'The Dr Foster Global Frailty Score has significant predictive capacity beyond that of other known predictors of poor outcome in older persons, such as co-morbidity and chronological age.' to more accurately reflect findings.

The odds ratios for mortality, length of stay and readmission were no greater than 1.1. Here are a few of many papers showing stronger clinically significant predictive power using frailty2-7.

Many thanks for your comment. We note that the supplied references do not reflect coding of frailty in large administrative datasets, but represent mainly observational trials both prospective and retrospective, or secondary analysis of prospective controlled trial data or clinical datasets. These clinical or trial datasets are 'richer' and have variables associated with frailty not readily coded within the structured diagnostic codes of administrative datasets, but are smaller in recruitment number, and require frailty to be measured manually (e.g. Clinical Frailty Scale) and as such is resource intensive, and subject to inter-operator error. Thus, we feel this is not comparable to our present study. Our hypothesis is that there is utility for widespread coding of frailty in readily available administrative data. As such, we have conducted a literature review of prognostic studies for older persons utilising administrative data, summarised in Appendix 4. The text ‘ Appendix 4 summarises the characteristics, setting, data sources, predictor and outcome variables and performance of recent case-mix studies for older persons utilising administrative data.’ had been previously included in the discussion section to reflect this.

‘The ORs and predictive capacity in the validation cohort were generally lower than the derivation cohort, but are in keeping with other risk prediction models for older persons within the English secondary care administrative data’ – this is not accurate reflection of previous studies which describe a gradient of increasing risk of adverse outcomes with increasing frailty risk scores – which the present paper does not appear to have done. This is important as the whole point of the frailty concept is to unpick the heterogeneity that age and comorbidities alone do not address fully.

Many thanks for the comment. Again we are comparing to risk prediction models for older persons within administrative datasets. We have conducted a literature review of prognostic studies for older persons utilising administrative data, summarised in Appendix 4. The text ‘ Appendix 4 summarises the characteristics, setting, data sources, predictor and outcome variables and performance of recent case-mix studies for older persons utilising administrative data.’ had been previously included in the discussion section to reflect this. We have used Area under the Receiver Operator Characteristic Curve(AUROC) to express predictive power, in keeping with existing literature (Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Contin Educ Anaesth Crit Care Pain 2008;8:2213). We note that we have not stratified the score at specific cut-points to demonstrate separation for the three outcomes measured, particularly as we feel these cut-points should reflect the sensitivity and specificity of it’s intended use (e.g. health resource planning). The text ‘Further investigation of appropriate cut-points based on desired model sensitivity and specificity for the above outcomes depending on how the model is used (e.g. health resource planning) represents future work.‘ has been added to the Discussion section to reflect this, and this has been highlighted under the limitations section.

References

1. Searle SD, Mitnitski A, Gahbauer EA, et al. A standard procedure for creating a frailty index. BMC geriatrics 2008;8:24.

2. Gilbert T, Neuburger J, Kraindler J, et al. 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 2018;391(10132):1775-82.

3. Hubbard RE, Peel NM, Samanta M, et al. Frailty status at admission to hospital predicts multiple adverse outcomes. Age and Ageing 2017:1-6.

4. Hewitt J, Moug SJ, Middleton M, et al. Prevalence of frailty and its association with mortality in general surgery. American Journal of Surgery 2015;209(2):254-9.

5. Basic D, Shanley C. Frailty in an older inpatient population: using the clinical frailty scale to predict patient outcomes. Journal of Aging and Health 2015;27:670-85.

6. Evans SJ, Sayers M, Mitnitski A, et al. The risk of adverse outcomes in hospitalized older patients in relation to a frailty index based on a comprehensive geriatric assessment. Age and Ageing 2014;43:127-32.

7. Wou F, Gladman JRF, Bradshaw L, et al. The predictive properties of frailty-rating scales in the acute medical unit. Age and Ageing 2013;42:776-81.

VERSION 2 – REVIEW

REVIEWER	Andrew Clegg University of Leeds UK I led the development, validation and national implementation of the primary care electronic frailty index (eFI) in th UK.
REVIEW RETURNED	22-Mar-2019

GENERAL COMMENTS	The authors have made appropriate modifications to the manuscript based on reviewer comments.