RESEARCH ARTICLE

Clinical frailty scale: Inter-rater reliability of retrospective scoring in emergency abdominal surgery

Katarina M. Fornæss¹ | Pia L. Nome¹ | Elin Kismul Aakre¹,² | Tor-Arne Hegvik³,⁴ | Ib Jammer²

¹Department of Clinical Medicine, University of Bergen, Bergen, Norway
²Department of Anaesthesia and Intensive Care, Haukeland University Hospital, Bergen, Norway
³Department of Biomedicine, University of Bergen, Bergen, Norway
⁴Department of Breast and Endocrine Surgery, Haukeland University Hospital, Bergen, Norway

Abstract

Background: Frailty is a complex syndrome shown to be an independent predictor of morbidity and mortality after surgery in older patients. Frailty scoring may, therefore, be important, for example, for pre-operative risk assessment and prognosis estimation. The Clinical Frailty Scale (CFS) has been developed to help operationalize frailty in the individual patient. However, the inter-rater reliability of retrospective CFS scoring through patient records by health care personnel is currently unknown in patients over 80 years of age undergoing emergency abdominal surgery.

Methods: Retrospective review of electronic patient journal of 112 patients over 80 years of age undergoing emergency abdominal surgery between 2015 and 2016. Three researchers individually assigned each patient a CFS score. The inter-rater reliability was assessed using Cohen’s weighted kappa for the comparison of pairs of assessors, as well as Kendall’s coefficient of concordance for the comparison of all three raters simultaneously.

Results: The agreement across raters was strong, with Cohen’s kappa values ranging between 0.74 and 0.85 and a Kendall’s coefficient of concordance of 0.86.

Conclusions: The inter-rater reliability of assigned CFS from patient journals seems acceptable. This could permit retrospective research utilizing CFS measures from several raters and across centers.

Keywords: clinical frailty scale, frailty, interrater reliability, octogenarian, perioperative medicine

Editorial Comment

This analysis found that when assessors with different degrees of clinical experience score patients according to the Clinical Frailty Scale, a strong agreement in their scoring was found. This suggests possible usefulness of the Clinical Frailty Scale in the perioperative period to support an informed assignment of healthcare resources, to try to improve patient outcomes.
1 | INTRODUCTION

A growing global population and an increasing share of older individuals generate a rise in the number of surgeries performed in older patients, including emergency procedures. Older patients undergoing emergency surgery have a high risk of adverse outcomes and mortality.

Surgical patients are traditionally evaluated using the American Society of Anesthesiologist Physical Status Classification system (ASA). ASA focuses on the presence of one or more diseases or smoking and does not evaluate the general health and functional status of the patient. Therefore, a need has risen for more accurate risk assessment. One suggested method for improved risk identification in older patients is the use of a frailty score, which has been shown to be a more accurate independent predictor of mortality and morbidity in older patients than ASA.

Frailty can be defined as a reduced ability to maintain homeostasis and therefore an increased vulnerability to stressors such as infections, surgery, and traumas, as well as an increased risk of adverse outcomes. There is a multitude of definitions of frailty and a corresponding amount of screening tools. One of the most cited frailty scoring tools is the "Clinical Frailty Scale" (CFS) by Rockwood et al. which is based on the accumulation of deficits and loss of functions, and consists of nine categories from "very fit" to "terminally ill". The CFS is designed to be simple and quick in use and does not require any equipment. Therefore, it is well suited for pre-operative frailty assessment in elective and emergency procedures.

Older patients undergoing emergency surgery are often extremely unwell and delirious and thus often unable to accurately account for their own physical health status. A reliable method of retrospective scoring could assist clinicians to identify patients too frail for many emergency procedures, as well as identify patients in need of individualized planning and follow-up. In addition, this would permit the use of CFS scores across centers and between researchers. Our study aims to evaluate the inter-rater reliability of retrospective CFS-scoring based on previous medical history in patients >80 years of age undergoing acute abdominal surgery.

2 | MATERIALS AND METHODS

2.1 | Sample

In this single-center study conducted at a tertiary University Hospital in Norway, the hospital's electronic operation booking system was searched for all patients above 80 years of age undergoing emergency laparotomy excluding vascular surgery, between January 2015 and December 2016.

2.2 | Design and methods

All patients included in the study were independently scored by three researchers, one consultant anesthesiologist, and two medical students. The scoring of patients was based on a retrospective review of the written medical history. The CFS had no further instructions or examples on how the different categories should be interpreted and there were no internal discussions on how the CFS should be applied.

The patients were scored using the nine-point CFS ranging from "Very Fit" to "Terminally Ill" developed by Rockwood et al. The scale consists of a single-page index describing each score with both a short text and an illustration.

Each patient was assigned three CFS scores, one by each researcher, hereafter called CFS-A (consultant anesthesiologist, EKA), CFS-B (medical student, KMF), and CFS-C (medical student, PLN). These three scores were compared to assess the reliability of the CFS.

2.3 | Ethics approval

This study was conducted as an extension of a published study. The Ethical Board approved the study and waived informed consent (REK2017/610).

| Variable | Count or (median) | Percentage |
|----------|------------------|------------|
| Age (80–94) | (85) |          |
| Female | 66 | 59% |
| ASA classification | | |
| ASA 1 | 0 | – |
| ASA 2 | 11 | 10% |
| ASA 3 | 63 | 58% |
| ASA 4 | 34 | 31% |
| ASA 5 | 0 | – |

| Comorbidities | Count | Percentage |
|---------------|-------|------------|
| Hypertension | 58 | 52% |
| COPD, emphysema or asthma | 28 | 25% |
| Ischemic heart disease | 28 | 25% |
| Arrhythmia | 28 | 25% |
| Diabetes mellitus | 18 | 16% |
| Cerebrovascular disease | 16 | 14% |
| Cogestive heart failure | 14 | 13% |
| Kidney failure | 15 | 13% |
| Dementia or Cognitive failure | 13 | 12% |
| Peripheral vascular disease | 7 | 6% |

Abbreviations: ASA, American Society of Anesthesiologist Physical Status Classification system; COPD, chronic obstructive pulmonary disease.

Four patients have missing information on ASA category.
2.4 | Statistical analyses

For each pair of CFS raters (CFS-A – CFS-B, CFS-A – CFS-C, CFS-B – CFS-C) inter-rater reliability of the CFS measures was calculated using quadratic weighted Cohen’s kappa. The 95% confidence intervals (95%CI) were calculated based on 10,000 bootstrap replicates and the percentile method. Using quadratic weights, large variations are emphasized, whereas small variations (e.g., a difference of 1 point on the ordinal scale) are less influential. The kappa value is presented as a numerical value between −1 and 1, where −1 represents negative agreement, 0 represents no or random concordance, and 1 represents perfect agreement.

We computed the arithmetic mean of all pairs of assessors as a measure of inter-rater reliability, in line with the current literature.

As an additional measure of inter-rater reliability across all three assessors, Kendall’s coefficient of concordance was used to calculate the degree of concordance between all three researchers simultaneously. 95% CIs were calculated with bootstrapping as described above. Kendall’s coefficient of concordance does not utilize the CFS scores themselves, but rather uses the rank of the CFS given to each observation by each assessor, that is, a non-parametric method. The Kendall’s coefficient of concordance can be a numerical value between 0 and 1, where 0 represents no or random agreement between raters, and 1 represents perfect agreement. However, the numeric values can be to some degree likened to the kappa, for example, a Kendall’s coefficient of concordance of 0.85 can be defined as strong.

All statistical analyses were conducted using R v. 4.0.3.

3 | RESULTS

A total of 112 patients were identified and scored. Patients were between 80 and 96 years old and of these 66 (59%) were female. Patient characteristics are presented in Table 1.

The distribution of CFS scores set by each researcher is depicted in Figure 1. The median (and range) of assigned CFS scores were 4 (2–8), 5 (1–9), and 4 (2–7) for CFS-A, CFS-B, and CFS-C respectively. When comparing CFS-A to CFS-B, 42 scores were identical, 45 differed by one point and 21 differed by two points. Only four differed by more than two points. Comparing CFS-B to CFS-C, 53 scores were identical, 51 differed by one point and 6 differed by two points. Only two differed by more than two points. Comparing CFS-A to CFS-C, 39 were identical, 58 differed by one point and 9 differed by two points. Six differed by more than two points. This is illustrated in Figure 2.

Weighted Cohen’s kappa was used to compare all three CFS scores to each other, yielding kappa values ranging from 0.74 to 0.84. The arithmetic mean of the weighted Cohen’s kappa from all pairs of assessors was 0.78. In addition, Kendall’s coefficient of concordance across all three CFS was calculated to 0.86. These results are summarized in Table 2.

4 | DISCUSSION

In this observational study of octo- and nonagenarians undergoing emergency abdominal surgery, we investigated the inter-rater reliability of CFS scored retrospectively by three independent assessors.
The agreement across raters was strong, with Cohen’s kappa values ranging between 0.74 and 0.85 and a Kendall’s coefficient of concordance of 0.86. It should be noted that we demonstrate a strong agreement across all three raters using both the arithmetic mean of the quadratic weighted Kappa measures (a parametric method) and Kendall’s coefficient of concordance (a non-parametric method).

This indicates that our results cannot easily be explained away by, for example, model-misspecification. Additionally, we demonstrate a high degree of agreement despite a wide range in the CFS scores set. This indicates that the CFS tool is well suited for assessing frailty.

Our study is currently the only study to compare retrospectively assigned CFS scores of real patients set by three researchers per patient, thus increasing the validity of our data. Other studies have compared scores set by pairs of researchers. Moreover, to our knowledge, this is the first study to investigate CFS from electronic patient journals among older individuals undergoing emergency surgery. Other studies have investigated retrospective assignment of CFS and inter-rater reliability in intensive care unit (ICU) patients and in the geriatric outpatient population. A recent large multi-center study demonstrated a strong inter-rater reliability of the CFS for older ICU patients (Cohen’s weighted kappa 0.86). In this study, one researcher scored based on all available information and one scored solely based on written medical history. Interestingly they noted that scores set based on information from the patient were likely to be lower (less frail) than the corresponding score based on information from relatives or medical records. This demonstrates the unreliability of health status information provided by critically ill patients, hence the need for a standardized scoring system.

One strength of our study is that the three researchers had varying degrees of clinical experience and familiarity with the CFS tool, indicating that the CFS tool is reliable independent of prior experience. The two medical students had slightly higher agreement, suggesting that equal clinical experience could further improve reliability. It is possible that retrospective scoring would be an even more reliable method if official guidelines for interpretation were implemented to counter the subjective nature of the scoring system. This may be the case with the recently updated version of the CFS (CFS version 2.0). In this version, the level names are revised and supplementary guidance for interpretation is available.

This study also has limitations. It was conducted as a single-center study. Also, there was no prospective scoring of the patients, hence we could not compare retrospective scores to prospective scores. This has been examined in a small study by Davies et al. where they demonstrated a moderate agreement (Cohen’s kappa 0.64) between CFS-scoring from clinical examination with retrospective scoring.

### Table 2

Overview of the results of weighted Cohen’s kappa between three pairs of raters, as well as the arithmetic mean of these values. Kendall’s coefficient of concordance for all three raters simultaneously.

|                | Cohen’s coefficient | CI          |
|----------------|---------------------|-------------|
| CFS-A          | 0.75                | 0.66–0.82   |
| CFS-A          | 0.74                | 0.63–0.82   |
| CFS-B          | 0.85                | 0.77–0.90   |
| Arithmetic mean of Cohen's coefficients | 0.78       | 0.70–0.83   |
| Kendall's coefficient of concordance | 0.86       | 0.82–0.90   |

FIGURE 2 Scatterplots (A, B, and C) visualizing the agreement between CFS* scores set by pairs of raters; CFS-A, CFS-B, and CFS-C. The intersection between the x- and y-axis represents the number of times both researchers scored a particular CFS score. The size of the circle represents the number of times this combination occurred. N as a reference for interpreting the size of each circle. *CFS, clinical frailty scale.
from electronic patient journals. However, we believe they may have used unweighted kappa, rather than weighted Cohen's kappa, yielding a falsely low estimate of agreement. Possible bias in our study include prior knowledge of the patient's age. Also, upon opening the electronic journal system the researcher was alerted when the patient in question was dead. This is due to that the Norwegian electronic journal system is directly connected to the national population register. This could influence the researcher to perceive the patient in question as more frail and thus assign a higher CFS score which again could lead to falsely high reliability. In addition, the quality of medical records and the amount of information varies greatly, which could result in different interpretations of the patient's health condition by each researcher; however, this is likely a reflection of daily clinical work and thus speaks to the strengths of the CFS as a clinical tool in daily clinical practice.

Retrospective CFS scoring has great potential for both clinical use and for research purposes. It could allow for improved risk evaluation and risk stratification, as well as permitting researchers to utilize frailty measures from several raters, and across centers. Our study found that the inter-rater reliability of retrospective clinical frailty scoring based on patient records was strong. Future studies should focus on investigating CFS as a routine preoperative assessment.

CONFLICTS OF INTEREST
None to declare.

ORCID
Elin Kismul Aakre https://orcid.org/0000-0002-3131-8914
Tor-Arne Hegvik https://orcid.org/0000-0002-8045-6643
Ib Jammer https://orcid.org/0000-0001-6759-3929

REFERENCES
1. Wan He DG, Kowal P. An Aging World: 2015 - International Population Reports. 2016.
2. Partridge JS, Harari D, Dhesi JK. Frailty in the older surgical patient: a review. Age Ageing. 2012;41(2):142-147.
3. Saklad M. Grading of patients for surgical procedures. Anesthesiology. 1941;2(3):281-284.
4. Makary MA, Segev DL, Pronovost PJ, et al. Frailty as a predictor of surgical outcomes in older patients. J Am Coll Surg. 2010;210(6):901-908.
5. Bandeen-Roche K, Seplaki CL, Huang J, et al. Frailty in older adults: a nationally representative profile in the United States. J Gerontol, 2015. 70:1427-1434.
6. Sioutas G, Tsoufas G. Frailty assessment and postoperative outcomes among patients undergoing general surgery. Surgeon. 2020;18(6):e55-e66.
7. Rockwood K. A global clinical measure of fitness and frailty in elderly people. CMAJ. 2005;173(5):489-495.
8. Hewitt J, Williams M, Pearce L, et al. The prevalence of cognitive impairment in emergency general surgery. Int J Surg. 2014;12(10):1031-1035.
9. Rockwood K, Theou O, Mitriltski A. What are frailty instruments for? Age Ageing. 2015;44(4):545-547.
10. Church S, Rogers E, Rockwood K, Theou O. A scoring review of the clinical frailty scale. BMC Geriatr. 2020;20(1):393.
11. Young RL, Smithard DG. The clinical frailty scale: do staff agree? Geriatrics. 2020;5(2):40. 10.3390/geriatrics5020040
12. Wallis SJ, Wall J, Biram R, Romero-Ortuno R. Association of the clinical frailty scale with hospital outcomes. QJM. 2015;108(12):943-949.
13. Aakre EK, Utlvik A, Hufthammer KO, Jammer IB. Mortality and complications after emergency laparotomy in patients above 80 years. Acta Anaesthesiol Scand. 2020;64(7):913-919.
14. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm. 2013;9(3):330-338.
15. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282.
16. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34.
17. Team, R.C. A language and environment for statistical computing: R Core Team. Accessed January 25, 2021.
18. Davies J, Whitlock J, Gutmanis I, Kane S-L. Inter-rater reliability of the retrospectively assigned clinical frailty scale score in a geriatric outreach population. Can Geriatr J. 2018;21(1):1-5.
19. Flaatten H, Guidet B, Andersen FH, et al. Reliability of the clinical frailty scale in very elderly ICU patients: a prospective European study. Ann Intensive Care. 2021;11(1):22.
20. Darvall JN, Boonstra T, Norman J, et al. Retrospective frailty determination in critical illness from a review of the intensive care unit clinical record. Anaesth Intensive Care. 2019;47(4):343-348.
21. Rockwood K, Theou O. Using the clinical frailty scale in allocating scarce health care resources. Can Geriatr J. 2020;23(3):210-215.

How to cite this article: Fornæss KM, Nome PL, Aakre EK, Hegvik T-A, Jammer I. Clinical frailty scale: Inter-rater reliability of retrospective scoring in emergency abdominal surgery. Acta Anaesthesiol Scand. 2022;66:25–29. https://doi.org/10.1111/aas.13974