Evaluation of Frailty Assessment Tools and their Measurement Properties in Chronic Kidney Disease

by

Alisha Puri

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Science

Department of Medicine

University of Alberta

Abstract

Background:

Frailty is three to seven times more common in people with chronic kidney disease (CKD) than in those with normal kidney function. Although frailty and its impact in CKD is well-recognized, the measurement properties of the tools used to assess this syndrome are not known. The aim of this systematic review was to evaluate frailty assessment tools and their measurement properties in CKD.

Methods:

The study was conducted using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) guidelines and Preferred Reporting Items for Systematic reviews and Meta-Analyses for Protocols (PRISMA-P 2015). We searched ten electronic databases and screened studies as per the inclusion criteria: peer-reviewed original research, adults with CKD (non-dialysis, dialysis or kidney transplant (KT)), examines at least one established multidimensional tool used for the assessment of frailty, and presents information to evaluate the measurements properties of the tool. Methodological quality assessment and data synthesis were performed as per COSMIN guidelines. This review was registered with PROSPERO (CRD42021234558).

Results:

We retrieved 648 unique citations with 52 eligible studies of which a majority (n = 37, 71.1%) were prospective cohort studies. A large proportion (n = 12, 23%) of the data was retrieved from prevalent dialysis patients. There was limited data (n = 4, 7.7%) available for KT

recipients due to population overlap. Across all studies, the Fried Frailty Tool (original: n = 27, 51.9%; modified: n = 8, 15.4% of studies) was used most frequently. Only three measurement properties were evaluated for the frailty measurement tools: construct validity (discriminative (n = 42, 80.8%); convergent (n = 9, 17.3%), criterion validity (n = 2, 3.8%), and responsiveness (n =2, 3.8%). Studies using the Fried Frailty Tool most commonly evaluated the tool's discriminative validity. Only in the CKD non-dialysis population, the Fried Frailty Tool demonstrated good discriminative ability (pooled adjusted hazard ratio (aHR): 2.00 (95% CI: 1.51, 2.64, $p < 0.001$) in estimating the risk of death. Upon assessing the methodological quality of all 52 studies, 51 (98%) had inadequate methodological quality. Only one study comparing the Fried Frailty Tool to the Geriatric Assessment (GA) (criterion validity) was assessed as doubtful methodological quality. Due to the inadequate methodological quality, when rating the studies against the "Updated Criteria for Good Measurement Properties", all studies presented "indeterminate" overall quality. The quality of evidence per single study or pooled result was graded and resulted in "very low" quality of evidence. Finally, due to the lack of data, we cannot comment on the interpretability and feasibility of the frailty assessment tools used in each study.

Conclusions:

We aimed to identify the frailty tools validated in CKD and provide a recommendation for a tool(s) for use in clinical research and practice. Although a number of frailty tools exist, only a number have been validated in CKD populations, such as the Fried Frailty Tool, Clinical Frailty Scale (CFS), Frailty Index (FI), Comprehensive Geriatric Assessment (CGA)/GA, FRAIL Scale, and Groningen Frailty Indicator (GFI). The Fried Frailty Tool was used most frequently across all CKD subpopulations and provided sufficient data for construct (discriminative and

convergent) validity and criterion validity. However, due to low study quality, we cannot recommend the Fried Frailty Tool with confidence. Additionally, this review also did not identify any studies evaluating the frailty tools' reliability, measurement error, structural validity, internal consistency, and content validity. Hence, we cannot provide a recommendation for a tool(s) as per the COSMIN guidelines for use in clinical research and practice.

Preface

This thesis is an original work by Alisha Puri. No part of this thesis has been previously published.

Acknowledgements

I would like to express my deepest appreciation to my principle investigator, Dr. Stephanie Thompson (S.T.), who gave me the opportunity to pursue graduate studies and supported me throughout my journey to Medical School. Without her guidance and belief in my ability to succeed, this project would not have been possible. Given all the obstacles we made it.

I would like to thank my committee members, Dr. Aminu Bello and Dr. Puneeta Tandon, whose continuous feedback and support allowed me to succeed and fulfill my requirements on a short timeline.

In addition, a big thank you to Ms. Anita Lloyd (A.L.) who spent many hours helping me with each step of my thesis, from extracting data to late meetings discussing details. I am grateful to have met you and hope to work with you again soon. Finally, thank you to Ms. Sandra Campbell (S.C.) for taking the time to be a part of my project.

Funding

Funding provided from Department of Medicine and Division of Nephrology, University of Alberta.

Table of Contents

List of Tables

List of Figures

1.0 Introduction to the Concept of Frailty

Frailty is a multidimensional construct characterized by a decline in multiple domains of human function: physiological, nutritive, cognitive, social, and psychological.[1] Subsequently, frailty confers a high risk for hospital-related outcomes and mortality in those populations vulnerable to a decline in physical, social, psychological, cognitive, sensory, and nutritive domains.[2] Often frailty is defined synonymously with aging, comorbidity, multimorbidity and disability.[3] However, frailty is distinct from these conditions and more specifically shares the underlying pathophysiology. Due to the multifaceted nature of frailty, the concept of frailty is defined variably in clinical practice and research. Fried and colleagues (2001) operationalized a physical phenotype of frailty characterized by shrinking (unintentional weight loss and sarcopenia), weakness, poor endurance or exhaustion, slowness, and low activity.[2] While Fried presented the Frailty Phenotype, Mitniski and Rockwood (2001) developed the Frailty Index (FI), which describes frailty as a state of age-related deficit accumulation reflective of impairments in multiple systems.[4,5] A greater number of deficits poses a greater risk for frailty and adverse outcomes.[6] Both the Frailty Phenotype and the Frailty Index aim to define frailty and the criteria used to measure the syndrome; however, both differ in how frailty is operationalized. To date, there is no consensus on what frailty really is but rather what is known is the multifactorial nature of the condition and the vulnerability to adverse clinical outcomes.[6]

1.1 Assessment and Diagnosis of Frailty

The assessment and diagnosis of frailty is variable, due to the presence of several instruments and operational definitions of frailty. In the literature, over 67 unique frailty

assessment tools have been identified.[7,8] To help categorize these tools, four frailty models have been identified: physical, cumulative deficits, geriatric syndrome, and multidimensional.[9]

The physical model of frailty is one that defines frailty based on physical performance. As per Chowdhury et al (2017), a majority of the literature in the general elderly population identifies physical frailty using the Frailty Phenotype (FP) (n=23, 72%), using both the original or modified model.[10] The original definition of the Frailty Phenotype consists of five criteria (1 point each): 1) slowness – measured by gait speed, 2) weakness – measured by grip strength, 3) exhaustion – measured by the centre for epidemiological studies depression scale, 4) shrinkage – measured by > 10 pounds of unintentional weight loss in 12 months, and 5) low physical activity – measured by an estimation of kilocalories per week.[2] Modifications to the Frailty Phenotype consist of self-report measures using the 36-Item Short Form Health Survey (SF-36) of slowness/weakness (SF-36 physical function score < 75; 2 points) and poor endurance/exhaustion (SF-36 Vitality Score < 55; 1 point).[11] A score of ≥ 3 is considered as frail, a score of 1-2 is considered pre-frail, and a score of zero is considered robust. Although the original model of the Frailty Phenotype is short and a good representation of underlying pathophysiology, the adaptations to the tool alter the reliability and validity by introducing ceiling effects in cases where self-reported measures are adopted.[12,13] Conversely, the questionnaire-based assessment increases the feasibility of administration in large populations, disabled populations, and permits use across virtual platforms, which are needed at the time of the COVID-19 pandemic.

Within the physical model, the short-performance physical battery (SPPB) is commonly used as a marker for frailty across elderly populations.[14,15] The SPPB solely assesses lower limb function by three different tests: balance (side-by-side, semi-tandem, and tandem balance), 4-

metre gait speed and chair stand test. Each test is scored from zero to 4 possible points with a total possible of 12 points overall.[7] A score less than 2 is indicative of disability, 3 to 9 indicates frailty, and a score greater than 10 indicates robustness.[14] The SPPB is not designed to measure frailty, rather, it is a surrogate measure of the physical domain to assess vulnerability in the elderly to stressors.[16] It can be argued that assessing frailty only by the physical domain is not truly indicative of frailty, as it is a multifactorial condition.

The Frailty Index, a tool categorized within the cumulative deficits model, is a quantitative assessment of frailty, operationalizing the condition as a collection of symptoms, behaviours, functional limitations, clinical conditions, and diseases.[5,17] In this approach, over 30 variables are evaluated for deficits.[4, 6] The variables must meet the following criteria to generate a frailty index: 1) the item must be acquired, 2) age-associated, and 3) associated with an adverse outcome.[18] Each included variable is dichotomized (present or absent) from which the frailty index can be calculated as a fraction of one (number of deficits present divided by total number of variables assessed). The distinguishing feature of the Frailty Index is its continuous nature represented by the frailty score ranging from 0 to 1; a cut-off score > 0.25 indicating the presence of frailty.[4] The Frailty Index is categorized in both the cumulative deficits and multidimensional model. Within the multidimensional model is the FI-CGA, where the comprehensive geriatric assessment (CGA) is used to measure baseline variables and > 15 variables are assessed for deficits.[19]

The third frailty model presented by Montgomery (2021) is the 'geriatric syndrome' model–a model that is not well defined in the literature.[9] Frailty is categorized as a geriatric syndrome because it is a multifactorial condition where impairment in multiple systems results in vulnerability to stressors. The geriatric syndrome model, like the cumulative deficits model, is

multidimensional, however, differs by mode of assessment. To assess this geriatric syndrome, the comprehensive geriatric assessment (CGA) is considered the closest to a gold-standard for the diagnosis of frailty in clinical care.[19] The CGA is a multidisciplinary diagnostic process that encompasses the core domains of functional status, cognition, emotional status, nutritional status, comorbidities, polypharmacy, fall risk, sleep, pain, urinary incontinence, and social history.[19,20] Each domain is scored variably based on the criteria included. The CGA is designed to account for each individual's unique history, however, due to the numerous criteria, the CGA is a resource intensive process that also requires clinical expertise to conduct. Although the CGA provides a comprehensive measure of multiple domains, it is difficult to interpret and use for frailty assessments. The CGA and any tool derived from this have been recommended for the management and follow-up of frailty.[20] Thus, the CGA has been recommended for frailty assessment in the elderly, those clinically high risk, and hospitalized individuals.[20] As discussed previously, the CGA has been incorporated within other frailty assessment tools, such as the FI specifically termed the FI-CGA. Previously collected data for the CGA can be incorporated into the FI to assign a quantitative frailty score and status. Although this enhances the multidimensional nature of the FI, like the CGA it is even more resource-intensive. However, performing the FI after the CGA improves the interpretability and increases the accuracy of frailty classifications as the measures are standardized by a score. Nonetheless, the CGA remains a clinical standard for what a frailty tool should assess to accurately identify an individual's frailty status.[19]

From a Delphi study aimed at operationalizing a definition for frailty, it was recommended that frailty should be identified within the multidimensional model.[21] The multidimensional model as the name suggests, assesses the different domains of function such as

physical performance, nutritional status, mental health, and cognition. A comprehensive definition of frailty should include assessment of these domains.[21,22] The multidimensional prognostic index (MPI) derived from the CGA has been identified as a common tool for defining multidimensional frailty.[23] The MPI includes information on eight domains: functional status (ADL), independence in activities of daily living (IADL), cognitive status (measured through the Short Portable Mental Status Questionnaire (SPMSQ), comorbidity examined through the Cumulative Illness Rating Scale (CIRS), nutritional status (Mini Nutritional Assessment (MNA), mobility (risk of developing pressure sores evaluated through the Extron Smith Scale (ESS), polypharmacy, and co-habitation.[24] For each domain, the criteria is scored as follows: 0 = no problems, 0.5 = minor problems, and 1 = major problems. The total scores of the eight domains are aggregated and expressed as a single score from zero to one, where 0.0–0.33 is low risk (non-frail), 0.34–0.66 is moderate risk (mildly frail), and 0.67–1.0 is severe risk (severely frail). As compared to the CGA, the MPI reduces the time for administration, eliminating the concern of feasibility.[25] However, with the emerging concern for a multidimensional approach to frailty, there still remains an absence of data establishing the multiple domains contributing to frailty.

1.2 Measurement Properties of Frailty Assessment Tools

In selection of the appropriate frailty assessment tools for clinical application, the adequacy and quality of existing tools in literature must be identified and assessed. The COSMIN taxonomy has identified three domains: reliability, validity, and responsiveness, of which each domain contains one or more measurement properties, ‘quality aspects of the measurement instrument’.[26]

1.2.1 Validity

To first assess a tool we must evaluate its validity. Within this domain, we ask 'Does this tool measure the construct it claims to measure and for what purpose?'[26] It is crucial to define the construct as we want to ensure the tool indeed measures what we are interested in and is tailored for the target population. When assessing validity, we must consider i) content, ii) construct, and iii) criterion validity.

i) Content validity

Content validity is defined as the degree to which the content of the tool is representative of the construct it aims to measure with respect to relevance and comprehensiveness and includes face validity. Assessment of content validity in frailty tools will then depend on how frailty is conceptualized and is typically judgement based. To assess relevance, we must consider three factors: '1) Do all items refer to relevant aspects of the construct to be measured, 2) Are all items relevant for the study population, and 3) Are all items relevant for the purpose of the application of the measurement instrument?[27]

ii) Construct validity

Construct validity is used when a tool(s) are being compared for accuracy or correlation, with neither being a gold standard. Within construct validity are hypothesis testing, structural validity, and cross-cultural validity. For hypothesis testing, the tool(s) must present scores consistent with the hypotheses to have construct validity. When comparing a single tool, we are evaluating the differences within groups, classified as discriminative validity. However, when comparing the degree to which the scores of two or more instrument are consistent with the hypotheses, we are evaluating convergent validity. Structural validity is the degree to which the

scores of the instrument are an adequate reflection of the dimensionality of the construct being measured. As frailty is multidimensional construct, a tool with structural validity ensures our scores accurately classify those frail and non-frail. Cross-cultural validity is the degree to which a translated/adapted tool compares to the original version. An example of this would be the self-report Fried Frailty Tool. Fried has been modified in many different contexts with alternating names, however, the five criteria used to assess frailty remain the same.

iii) Criterion validity

Criterion validity is the 'degree to which the score of the instrument are an adequate reflection of a 'gold standard''.[26] Within criterion validity are 1) concurrent validity and 2) predictive validity. The purpose of assessing concurrent validity is to ensure both the measurement and the 'gold standard' score the same during the onset of disease. Thus, a tool used for evaluative and diagnostic purposes must have concurrent validity. To assess the similarity between the tool and gold standard, area under the curve (AUC), Pearson's correlation, sensitivity, specificity, PPV, and/or NPV are measured. Predictive validity, as the name suggests is the ability of the tool to predict the 'gold standard' in the future.[27] For a tool to predict the gold standard in the future, the disease has not yet manifested. To predict future outcomes a hazards ratio (HR), odds ratio (OR), AUC, or c-statistic can be used. However, for both concurrent and predictive validity, a criterion ('gold standard') must be established. As discussed previously, there is no 'gold standard' defined in literature for the assessment of frailty. The multidisciplinary Comprehensive Geriatric Assessment (CGA) has been identified as closest to a 'gold standard' for clinical care in frailty.[19] The CGA then can be used as a criterion ('gold standard') when assessing criterion validity of a frailty tool.

1.2.2 Reliability

Once the construct of a tool has been defined, we must consider reliability. Reliability as defined by Mokkink (2018), is "the degree to which scores remain unchanged for repeated measurements under different conditions".[26] Within the domain of reliability, is internal consistency, test-retest, intra-rater reliability, inter-rater reliability, and measurement error.[26] As there is no single tool to assess frailty across populations, when generalizing the use of a tool within a population we must consider its reliability.

i) Internal consistency

Internal consistency is the degree to which measurements are interrelated. A multidimensional tool must consist of items that measure the same construct. Thus, the scores or responses of these items are related and must be consistent across the sample. To have adequate internal consistency three requirements must be met (1) subscales should be shown to be unidimensional; (2) high Cronbach's alphas should be found in a number of studies of good methodological quality; and (3) results should be consistent.[27] A positive rating for internal consistency is assigned if factor analysis has been applied and Cronbach's alpha is between 0.70 and 0.95.[27]

ii) Test-retest, Intra-rater, and Inter-rater Reliability

Test-retest is defined as the consistency of the measurements over time.[26] Test-retest is important in the assessment of a longitudinal disease, such as frailty. The ability of a tool to reproduce the same results over a course of time allows for an accurate detection of the

progression and/or improvement in disease status. Intra-rater reliability is the consistency in measurements when the tool is administered by the same individual on different occasions.[26] When assessing frailty in individuals, the accuracy by which a primary provider measures frailty over time defines the frailty status at the given time point. This type of reliability is crucial when assessing frailty, as it is more common for an patient or participant to be evaluated by the same administrator. Inter-rater reliability is defined as the consistency in measurements when the tool is administered by different individuals on the same occasion. The parameters for assessing for the different types of reliability are Cohen's kappa for measurements on a nominal scale (unweighted kappa) or ordinal scale (weighted kappa) and the ICC for measurements with continuous outcomes.[27]

iii) Measurement Error

Measurement error is defined as the systematic and random error in scores not attributable to changes in the construct itself. [26] To measure the error, SEM or % agreement is calculated. Both measurement error and reliability are interrelated but distinct concepts. When measurement error is increased, reliability is decreased. Measurement error can be introduced by systematic error, also known as bias.[28] Systematic error can be information bias or selection bias. Systematic error is measured by the mean difference between the measured and the true value. Secondly, measurement error can be introduced by random error, also known as imprecision. Random error is measured by the standard deviation (SD) of mean values.

1.2.3 Significance of Measurement Properties

Finally, defining the purpose for which the tool is designed is necessary. To assess frailty and predict future outcomes of this syndrome, we need a tool(s) that has diagnostic, evaluative,

and has predictive abilities.[27] For diagnosis, a tool must be able to discriminate between individuals at a single point in time, which is described by COSMIN as construct – discriminative validity. Discriminative validity of a tool allows us to determine whether there is indeed a difference in disease status between healthy and affected subgroups. Further, to evaluate the condition of frailty and its dynamic nature over time, we must consider a tool that has to ability to detect changes longitudinally. COSMIN has defined this measurement property as responsiveness. Finally, for a tool to be able to predict future outcomes, it must possess criterion-predictive validity. Although, the goal of a frailty assessment is to target those populations affected with frailty, being able to predict progression of the condition or onset would be a factor in selection of a tool for clinical practice. A tool that possesses the appropriate measurement properties allows for increased confidence in whether the tool is truly measuring frailty in our population of interest. Additionally, we can ensure high quality research in this field.

1.3 Epidemiology of frailty in CKD

Chronic kidney disease (CKD) is defined by an estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73 m^2 for over 3 months.[29] The severity of CKD is defined by Stages I-V with Stage V representing end-stage kidney disease (eGFR < 15 mL/min/1.73m^2) with or without dialysis dependence. Progression through the stages of CKD is measured via eGFR decline and magnitude of albuminuria.[30] Frailty is a condition that is highly prevalent in those with all stages of kidney disease. The prevalence of frailty in non-CKD elderly populations is 11% versus 43% in pre-dialysis patients and 73% in dialysis patients.[31] Additionally, many studies have found an inverse correlation between eGFR and frailty regardless of the population age.[32] An eGFR decline of > 4.1 ml/min per 1.73 m^2 per year has been associated with incident

frailty as compared to a normal decline of eGFR in the elderly of 6.3 mL/min/1.73 m^2 per decade.[33,34] As CKD progresses, frailty is expected to progress to more severe stages resulting in adverse outcomes as observed in non-CKD populations.[34] Zhang and colleagues have demonstrated this in CKD by systematically reviewing existing literature assessing the association between frailty and all-cause mortality.[35] Overall, frailty was reported to increase the risk of all-cause mortality by about two-fold in CKD and dialysis populations (HR 1.95; 95% CI 1.50, 2.53). It is evident that the increased prevalence frailty in CKD poses a greater risk for adverse outcomes than in the general population.[31,35]

An increased prevalence of frailty in CKD populations than the general population can be explained by the mechanisms shared between frailty and CKD. In CKD, due to reduced kidney function and the requirement of kidney replacement therapy (hemodialysis, peritoneal dialysis), multiple physiological changes can occur. The physiological changes include but are not limited to cardiovascular disease, inflammation, malnutrition, anemia, and bone mineral metabolism.[36] CKD and renal failure are associated with increased inflammation due to renal replacement therapy and reduced kidney function in itself.[36,37] Factors contributing to inflammation include pro-inflammatory cytokines, uremia-related complications (metabolic acidosis, sarcopenia), and oxidative stress.[38,39] Inflammation plays a key role in reduction of muscle mass via the increased circulation of pro-inflammatory cytokines.[40] The cytokines function to inhibit muscle anabolism by reducing the rate of protein synthesis paralleled by enhanced protein breakdown. Thus, the decreased muscle mass contributes to decreased muscle contractility, resulting in dynapenia (loss of muscle strength).[41,42] Further, dynapenia contributes to physical inactivity, muscle wasting, sarcopenia, fatigue, and disability; all characteristics of frailty.[36,42] In addition, chronic inflammation is associated with cardiovascular disease (CVD) and atherosclerosis, leading to a

greater risk of CVD and related diseases in CKD.[43] Due to the progressive nature of CKD, individuals with more severe CKD are at a greater risk of frailty.[31] Of the CKD population, Stage V patients and those seeking renal replacement therapy are at greatest risk of frailty and adverse outcomes.[10] Further, those categorized in Stage V (ESKD) are likely to be assessed at this stage for their eligibility for renal transplantation options.[31] Being at severe risk for frailty reduces the likelihood of receiving a kidney transplant, further deteriorating renal function in ESKD and/or increasing their risk for complications upon transplant success.[31] Therefore, it is crucial to assess and identify these CKD frail populations to detect frailty, predict clinical outcomes, and intervene with the aim of altering the progressive course of frailty.

1.4 Objectives

Frailty is a multidimensional construct dichotomized as a physiological syndrome or as a heightened state of vulnerability across multiple systems leading to poor health outcomes.[9] Frailty is three to seven times more common in people with chronic kidney disease (CKD) than in those with normal kidney function.[31] Previous systematic reviews have thoroughly studied the manifestation of frailty in CKD. However, despite the significance of frailty in CKD, there is limited research to inform which frailty tools have optimal diagnostic and predictive accuracy in this population. The aim of this systematic review was to evaluate frailty assessment tools and their measurement properties in CKD with the following objectives:

1. To identify and summarize the established multidimensional frailty screening tools that have been evaluated in Chronic Kidney Disease (CKD) populations.
2. To evaluate the measurement properties of these tools using the COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments) Checklist.

3. To make recommendations for application of tools in clinical practice and research based on their diagnostic and predictive accuracy.

2.0 Methods

The systematic review was conducted in accordance with the COSMIN recommendations and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-DTA).[44,45] This review is registered in PROSPERO (CRD42021234558).

2.1 Search strategy

The search strategy was developed and executed in collaboration with an expert searcher/health librarian (S.C.) on the following databases: PROSPERO, OVID Medline, OVID EMBASE, OVID Health and Psychosocial Instruments, Cochrane Library (CDSR and Central), EBSCO CINAHL, Proquest Dissertations and Theses Global and SCOPUS. Controlled terms (MeSH, Emtree) and key words representing the concepts "chronic kidney disease" and "frailty" and "measures or instruments" were used. No limits were applied to the search. Databases were searched from inception to January 23, 2021. The results obtained were exported to Covidence, a review management software platform where any duplicates were removed[46]. Identified records were screened by A.P. and A.L. using a three-step process: 1) title-abstract screening, 2) a calibration exercise of 10 randomly selected studies to ensure both reviewers were familiar with the inclusion/exclusion criteria, and 3) citations considered to be relevant by both reviewers were retrieved for full-text screening. A full description of the search strategy can be found in Appendix I: Search Strategy.

2.2 Eligibility criteria

Studies were included in the review if: 1) the study was peer-reviewed and published as a full original article in English; 2) the study participants were human ≥ 18 years of age; 3) the target population or a subgroup of the population were individuals with Chronic Kidney Disease (CKD), defined as stage 3-5 CKD, GFR < 60 mL/min/1.73 m^2, on dialysis or non-dialysis, or renal transplant recipients; 4) frailty was assessed in the population using at least one established multidimensional tool, defined as a tool assessing two or more domains of function[47,48,49] the study evaluated at least one of the following measurement properties (content validity, reliability, structural validity, responsiveness, internal consistency, measurement error, cross-cultural validity, construct validity, criterion validity) of frailty assessment tool(s) as defined by the COSMIN taxonomy (Appendix II: Relevance Form). [26]

Studies were excluded from the review if: 1) the study participants belonged to a pediatric population; 2) the sample population was non-CKD; 3) frailty was not assessed by an established multidimensional tool; and 4) the study was presented as a letter to the editor, narrative review, case report or case series.

2.3 Data extraction

A standardized method was used to extract and record relevant data for all eligible studies. Data from eligible studies were extracted independently by one reviewer (A.P.) and checked by a second reviewer (A.L.) using a standardized data extraction sheet. The following parameters were extracted from each study: study identification (year, setting, country, language), study characteristics (study design, study aim, number of groups, group types, participant type, formal study name), participant demographics (total sample size, sample size,

age, sex, BMI, comorbidities (diabetes, cardiovascular disease (CVD), hypertension, heart failure, stroke, chronic obstructive pulmonary disease (COPD), dementia, peripheral vascular disease (PVD)), renal replacement therapy (RRT) duration, time since transplant, chronic kidney disease (CKD) stage, eGFR, serum albumin, serum creatinine, creatinine clearance, dialysis dependence, dialysis modality (home hemodialysis (HHD), conventional HD, peritoneal dialysis (PD)), previous transplant recipient), instrument administration (study referent frailty assessment tool, frailty assessment tool(s) administered, frailty scale used, tool adaptation, tool validation, and multidimensional nature (physical, nutritive, social, psychological, cognitive, and sensory domains). Assessed outcomes included: tool comparisons, tool accuracy, all-cause mortality, hospital-related outcomes, and measurement tool properties as per COSMIN.[26] Data were extracted from graphs, if required.

2.4 COSMIN

The COSMIN Checklist was designed by Mokkink (2018) and committee to evaluate the methodological quality of studies on measurement properties of health status measurement instruments.[26] The checklist allows for an accurate quality assessment of studies while comparing measurement properties of multiple validated and non-validated instruments. Thus, the use of the COSMIN Checklist provides a systematic and reproducible approach with which to evaluate studies on measurement properties.

2.4.1 Risk of Bias Assessment/ Methodological Quality Assessment

We performed the Risk of Bias assessment as outlined by the COSMIN manual.[26] Two reviewers A.P. and A.L. independently performed this step. Consensus was achieved by discussion upon completion of the checklist. We followed the COSMIN Risk of Bias Checklist consisting of 10 boxes and 10 measurement properties (Appendix III: Table 4). Of the 10 boxes,

only nine boxes were assessed as follows: content validity (box 2), structural validity (box 3), internal consistency (box 4), cross-cultural validity/ measurement invariance (box 5), reliability (box 6), measurement error (box 7), criterion validity (box 8), hypotheses testing for construct validity (box 9), and responsiveness (box 10). PROM development (box 1) was not assessed as we did not assess the construction of a frailty assessment tool. Per discussion with the review team (A.P., A.L., and S.T.) and search of existing literature, the CGA/GA was selected as our gold standard or "criterion".[9] For each eligible study, the checklist was completed for each of the frailty assessment tool(s). In an instance where more than one measurement property was evaluated in a study, the respective measurement property was assessed independently under its corresponding box. The COSMIN four-point rating system was used to evaluate each standard within each of boxes: 'very good', 'adequate', 'doubtful', or 'inadequate'(Appendix III: Table 3). To standardize the Risk of Bias Assessment, both A.P. and A.L. searched relevant literature to create a thorough guide to the four-point rating system which can be found in Appendix IV. The purpose of the guide was to set a number of criteria to answer the following question posed by COSMIN in each of the boxes '*Were there any other important flaws in the design or statistical methods of the study*? Based on the criteria defined by both reviewers, each standard was rated independently. Consensus of the ratings per standard were achieved by discussion between A.P. and A.L.. The overall rating of quality was assigned by taking the lowest rating of any standard in the box.[26]

2.4.2 Quality Criteria Assessment

Following the Risk of Bias assessment, one reviewer (A.P.) independently rated each single study against the 'Updated Criteria for Good Measurement Properties' (Appendix III: Table 4).[49,50] This step is to evaluate the overall rating of the study's findings for each

measurement property. A second reviewer (A.L.) reviewed the results. Each result was rated as 'sufficient (+)', 'insufficient (-)' or 'indeterminate (?)'.

2.4.3 Data Synthesis

Analyses were performed using Stata/MP, version 17 (StataCorp, LLC). Due to expected diversity between studies, we decided a priori to combine results (e.g. hazard ratios, odds ratios) using random effects models.[51] Outcomes were pooled by population, frailty assessment tool and measurement property, given sufficient data. Populations were categorized as kidney transplant (KT) recipients, CKD non-dialysis, incident dialysis (< 3 months on dialysis), prevalent dialysis (> 3 months on dialysis), and CKD mixed dialysis/non-dialysis. Adjusted results were chosen over unadjusted if both presented; unadjusted odds ratios were calculated for studies where only counts of outcomes of interest were presented (and no other useable results were available). Adjusted results were pooled separately from unadjusted results. Statistical heterogeneity was quantified the I^2 statistic.[52] For studies with overlapping populations that presented the same type of result, the outcome of the larger cohort was used for pooling. For outcomes that could not be pooled, we narratively summarized the findings.

2.4.4 Summary and Grading Quality of Evidence

The quality of evidence for the pooled or summarized results of each tool was graded using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach (Appendix III: Table 5).[53] The quality of evidence was graded as high, moderate, low, or very low. To determine the quality of evidence, the GRADE approach uses four factors: i) risk of bias, ii) inconsistency, iii) imprecision, and iv) indirectness. Each study's pooled or summarized results were evaluated as high quality and downgraded a level(s) per the individual factor evaluations. The evaluation per factor is as follows:

- 0 - none (multiple studies of adequate quality or one study with very good quality)
- -1 - serious (multiple studies of doubtful quality or one study of adequate quality)
- -2 - very serious (multiple studies of inadequate quality or one study of doubtful quality)
- -3 - extremely serious (only one study of inadequate quality).

The first factor, Risk of Bias, was assessed using the COSMIN Risk of Bias checklist during the methodological quality assessment. Next, the consistency of results was determined during the assessment for overall quality as described above. In the case of unexplained inconsistency, three scenarios were taken into account: i) the single study results can be pooled and can be rated as sufficient or insufficient, ii) the single study results cannot be pooled and will be rated as inconsistent, or iii) the single study results are indeterminate. In the first scenario, the results would be graded, however, will be downgraded due to serious (-1) or very serious (-2) inconsistency. In the second and third scenario, the evidence would not be graded and rated as indeterminate, respectively. The third factor, imprecision, refers to the total sample size of the pooled or summarized studies. If the total sample size was $n < 50$ to 100, imprecision would be categorized as serious (-1), however, if the total sample size was $n < 50$, imprecision would be categorized as very serious (-2). Finally, indirectness was determined by evaluating whether the evidence gathered was indeed from the population of interest. In the case of study inclusion from a different population or a partially eligible population, the quality of evidence was downgraded due to serious (-1) or very serious indirectness (-2). The final grade for each individual study or pooled resulted was added to provide a recommendation(s) for a frailty assessment tool.

3.0 Results

3.1 Search Results

Of the 1106 unique records identified, 52 studies were included in this systematic review as per our inclusion criteria. A summary of the PRISMA diagram illustrates the process by which articles were deemed appropriate for inclusion (Figure 1). After the initial screening, 348 articles were retrieved of which 296 articles were excluded. Of the studies excluded, 138 studies were not peer-reviewed or published, 75 had no outcome of interest, 37 had no original research (narrative reviews, reviews, case report), 16 did not assess frailty by at least one established multidimensional tool, 10 were in non-English language, nine were of the non-CKD population eight were systematic reviews, eight had no original research – other (systematic review), two studies were duplicates, and one consisted of a pediatric population.

3.2 Characteristics of Included Studies

Of the 52 studies, we had 46 cohort studies (37 prospective, 9 retrospective), 5 cross-sectional studies, and one secondary analysis of a randomized clinical trial (Table 1). A majority of the studies were conducted in the United States (n = 17, 33%), the Netherlands (n = 7, 13%) and Canada (n = 7, 13%). The median year of publication was 2019 (min: 2013, max: 2021). The studies in our systematic review were categorized by five unique subpopulations: kidney transplant (KT) recipients, CKD non-dialysis, incident dialysis, prevalent dialysis, and mixed stage 5 CKD. Within the unique subpopulations are the following non-overlapping studies (n = 38): kidney transplant (KT) recipients (n = 4, 11%), CKD non-dialysis participants (n = 8, 21%). incident dialysis (n = 7, 18%), prevalent dialysis (n = 12, 32%), and mixed stage 5 CKD (dialysis and non-dialysis) (n = 7, 18%). End stage kidney disease (ESKD) was the most commonly

reported CKD status across all subpopulations. The mean age of the KT recipients was 50.7 years, with 62.1% being male. In the KT recipient population, an average of 77.2% participants were on dialysis, however, dialysis modality was not reported. The mean age of CKD non-dialysis participants was 71.3 years with 60.9% being male. The mean age of incident dialysis participants was 69.7 years with 59.1% being male. All participants were on dialysis, where 90.5% were on conventional hemodialysis (HD), 9.1% on peritoneal dialysis (PD), and 0.4% on home HD (HHD). The subpopulation with the greatest number of non-overlapping studies was the prevalent dialysis subpopulation with a mean age of 60.3 years and 60.3% male participants. All the prevalent dialysis participants were on dialysis, where 72.7% were on conventional HD, 23.3% were on PD, and 4% were on HHD. The mixed dialysis and non-dialysis subgroup had a mean age of 62.9 years, with 57.4% being male and only 51.7% of the participants on dialysis (Table 1).

3.3 Frailty Assessment Tools in CKD Populations

In the 52 studies, six frailty assessment tools were used most frequently across the studies: Fried Frailty Tool (original: n = 27, 51.9%; modified: n = 8, 15.4% of studies), Clinical Frailty Scale (CFS) (n = 10, 19.2%), Frailty Index (FI) (n = 4, 7.7%), Geriatric Assessment (GA) or Comprehensive Geriatric Assessment (CGA) (n =5, 9.6%), Groningen Frailty Indicator (GFI) (n = 7, 13.5%), and the FRAIL scale (n = 4, 7.7%) (Table 2). Within the studies including KT recipients, the Fried Frailty Tool was used most frequently (n = 5, 71.4%) to assess frailty, where 35.4% of the participants were assessed as frail overall. In the CKD non-dialysis subpopulation, the Fried Frailty Tool was used most frequently (n = 3, 30%), followed by the CFS and FRAIL Scale (each n = 2, 20%). The CGA was used to assess frailty in one

study; however, this study did not report the proportion frail using the CGA, as the Fried Frailty Tool was used instead. The prevalence of frailty in the CKD non-dialysis subpopulation was 44.3% using the Fried Frailty Tool; 23.5% frail by the CFS; and the FRAIL scale primarily assessed the participants as non-frail (98.5%). In the incident dialysis subpopulation, the Fried Frailty Tool was the most commonly used tool (n = 7, 63.6%), however, both the original Fried Frailty Tool (n = 5, 45.5%) and modified Fried Frailty Tool (n = 2, 18.2%) were used. The next most commonly used frailty assessment tools in studies that included incident dialysis patients were the CFS (n = 5, 45.5%) and CGA (n=4, 36.4%). The prevalence of frailty in incident dialysis participants was assessed as 43% frail by Fried and 65.8% frail by the CGA. Due to the heterogeneity in the scale cut-offs used for the CFS, the prevalence of frailty overall from the incident dialysis studies cannot be determined. In the prevalent dialysis subpopulation, primarily the modified Fried Frailty Tool (n = 6, 40%) was used to assess frailty. In the prevalent dialysis populations, the modified Fried Frailty Tools' scoring criteria were subjective to each study, hence, an overall frailty status cannot be determined. Finally, in the mixed dialysis and non-dialysis studies, again the Fried Frailty Tool was used most frequently (n =7, 77.8%). In the studies using the Fried, with a dichotomous scale (n = 3) (frail and non-frail), 18% of the participants were frail. In the remaining studies using the Fried (n = 4), 25.3% of the participants were frail, 48% were intermediately/pre-frail, and 26.8% were non frail.

3.4 Measurement Properties

Of the nine measurement properties defined by the COSMIN taxonomy within the domains of validity (content validity, structural validity, cross-cultural validity/measurement invariance, criterion validity, hypotheses testing for construct validity), reliability (internal

consistency, measurement error, and reliability: test-retest, intra- and inter-rater) and responsiveness, only three properties were evaluated for the frailty measurement tools in the 52 studies: construct validity (discriminative (n = 42, 80.8%); convergent (n = 9, 17.3%), criterion validity (n = 2, 3.8%), and responsiveness (n =2, 3.8%). The remaining measurement properties, content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, and measurement error were not assessed in any of the included studies. Studies evaluating more than one measurement property are reported independently under their respective outcome(s) (Table 2). Due to the clinical heterogeneity in studies, only a select number of studies that were quantitatively pooled are reported.

3.4.1 Construct Validity

3.4.1.1. Discriminative Validity

As reported in Section 3.4, a majority (n = 42, 80.8%) of the studies in this systematic review evaluated and presented the discriminative validity of the six commonly used frailty assessment tools. All the KT recipient studies (n =7) evaluated the discriminative validity of the frailty tool administered. Among the two tools used in the KT recipient studies, the discriminative ability of the Fried Frailty Tool was evaluated in five studies (71.4%) to measure the difference in hospital-related outcomes between frail and non-frail subgroups. In this subpopulation, only two studies, Konel et al[54] and dos Santos Mantovani et al[58], presenting an unadjusted odd ratio (uOR) of death for frail versus non-frail based on the Fried Frailty Tool were pooled. The pooled result, uOR 1.15 (95% CI: 0.50, 2.66, p = 0.74) refutes the hypothesis that the Fried Frailty Tool has adequate discriminative ability to detect the risk of death between frailty groups (Figure 2).

Of the non-dialysis CKD studies, a majority (n = 9, 90%) evaluated the discriminative validity of the following frailty tools: The Fried Frailty Tool (n = 2, 22%), CFS (n = 2, 22%), GFI (n = 1, 11%), and the FRAIL scale (n = 2, 22%). In the non-dialysis CKD subpopulation, the pooled result of two studies[101,104] presented an adjusted HR (aHR) 2.00 (95% CI: 1.51, 2.64, $p < 0.001$) of death for frail versus non-frail based on the Fried Frailty Tool.

In the incident dialysis studies (n = 11), discriminative validity was the most frequently evaluated measurement property (n = 9, 81.8%). Of the tools commonly used in the incident dialysis studies, the Fried Frailty Tool (n = 4, 36%), CFS (n = 4, 36%), GA (n = 3, 27%), and GFI (n = 2, 18%) all assessed discriminative validity. In this subpopulation, two pooled results were quantified. First, Fitzpatrick et al[85] and van Loon et al[87] presented an aHR for death based on the Fried Frailty Tool for frail versus non-frail (aHR: 1.58, 95% CI 1.04, 2.42, p = 0.03). This finding is consistent with the hypothesis that the Fried Frailty Tool can accurately discriminate risk of death between frailty groups in incident dialysis subpopulations. Second, Alfaadhel et al[89] and Yoshida et al[94] also presented an aHR for death but based on the CFS (score = 4) versus CFS (score of 1 to 3) (aHR: 3.18, 95% CI 0.76, 13.40, p = 0.12).

In the mixed dialysis studies (n = 9), discriminative validity (n = 5, 55.6%) was primarily evaluated by the Fried Frailty Tool (n = 4, 80%). Due to the heterogeneity in the statistics presented and frailty category definition in the studies, no pooled results were obtained. One study performed by McAdams-DeMarco et al[62] evaluated the discriminative validity of the Fried Frailty Tool by predicting the risk of death with and without frailty in the model and comparing c-statistics (c = 0.646 and 0.642, respectively (p = 0.65). The authors also presented an aHR death for intermediately frail (score of 1-2) vs. non-frail (score of zero) as 1.72 (95% CI 1.03, 2.88) and an aHR for death frail vs. non-frail as 2.18 (95% 1.25, 3.78).

In the prevalent dialysis studies (n = 15), the Fried Frailty tool was most commonly administered and evaluated discriminative validity (n = 5, 33.3%). Two pooled results were quantified. First, a pooled aHR of 1.28 (95% CI 0.48, 3.39; p = 0.62) was presented of death for frail versus non-frail based on the Fried Frailty Tool[74, 79]. Second, a pooled uOR of 2.81 (95% CI 1.67, 4.73; p < 0.001) was presented of death for frail versus non-frail based on the Fried Frailty Tool[70, 81].

3.4.1.2 Convergent Validity

All studies evaluating convergent validity were not pooled due to heterogeneity, thus, are qualitatively summarized in Table 2. Convergent validity was evaluated to a lesser extent within the concept of hypothesis testing for construct validity (n = 9, 17.3%). The Frailty Index (FI) and the Fried Frailty Tool were commonly compared against the GFI and CFS to assess convergent validity. In the KT recipient subpopulation, no studies evaluated the convergent validity of the frailty assessment tools. In the non-dialysis CKD studies (n =10), one study evaluated the convergent validity of the Fried Frailty tool[104]. Similarly, in the incident dialysis studies (n = 11), one study (Clark, 2017) evaluated the convergent validity of the FI vs CFS, FI vs FACT-CFS, and FI vs DMMS-FRAIL; and the CFS vs FACT-CFS and the CFS vs DMMS-FRAIL[91]. When comparing the tool accuracy between the FI and CFS (FI > 0.45 versus CFS ≥ 4), using the FI as the comparator, the CFS presented a sensitivity of 90 (95% CI: 68, 99) and a NPV of 95 (95% CI: 84, 99). When comparing the FI and CFS (FI > 0.21 versus CFS ≥ 5), the CFS presented a specificity of 100 (95% CI: 88, 100) and a PPV of 100 (95% CI: 81, 100). When comparing the FI and CFS (FI > 0.45 versus CFS ≥ 5), the CFS presented a NPV of 91 (95% CI: 84, 95). Of the prevalent dialysis studies (n = 15), four studies evaluated the convergent validity of the frailty

assessment tools. The Fried Frailty Tool (n = 3, 20%) , GFI (n = 1, 7%) and FRAIL scale (n = 1, 7%) were several of the tools commonly compared with the nurse and physician impression and the SF-36, GFI, G8, EFS, and TFI, respectively. Finally, in the mixed dialysis studies (n = 9), 33.3% of the studies evaluated the convergent validity of the frailty assessment tools. Of these tools the FI commonly presented convergent validity (n = 3, 33%), followed by the Fried Frailty Tool (n = 2, 22%), CFS and GFI (n =1, 11%).

3.4.2 Criterion Validity

Criterion validity was assessed in 3.9% of the total studies (n =2) (Table 2). As our defined criterion was the GA or CGA, only the Fried Frailty Tool was compared against the GA/CGA in both the CKD non-dialysis (n = 1, 50%) and incident dialysis (n =1, 50%) populations [86, 102]. In the non-dialysis CKD population, when compared with the GA (>1 impairment), the sensitivity of the Fried Frailty Tool was 66 (95% CI: 55, 77) for classifying participants as frail with a specificity of 85 (95%: CI 74, 96), PPV of 88 (95% CI: 83, 93), and NPV of 56 (95% CI 50, 62)[102]. Additionally, when compared with the GA (> 2 impairments), the sensitivity of the Fried Frailty Tool was 83 (95% CI: 71, 95), with a specificity of 76 (95% CI: 66, 86), a PPV of 89 (95% CI: 85, 93), and a NPV of 66 (95% CI 59, 73)[102]. In the incident dialysis population, when compared with the GA (≥ 2 impairments), the sensitivity of the Fried Frailty Tool (score ≥ 3) was 59 (95% CI: 48, 70), with a specificity of 85 (95% CI: 66, 96), a PPV of 92 (95% CI: 83, 97), and a NPV of 41 (95% CI: 34, 39)[86].

3.4.3 Responsiveness

Responsiveness was assessed in 3.9% of the included studies. Due to the limited nature of data, the results are qualitatively summarized (Table 2). In the incident dialysis studies (n = 11), only one study evaluated the responsiveness of the GA[73]. The study measured the change in frailty scores at baseline and at 12 months (median score baseline and 12 months respectively: 8 [IQR 7.0 - 9.5), 6 [IQR 5.0, 7.5], $p<0.001$. In the mixed dialysis studies (n = 9), one study evaluated the responsiveness of the Fried Frailty Tool. Lorenz et al[67] evaluated the difference between baseline (median 1 [IQR 1 – 3]) and post-intervention (exercise) (median 1 [IQR 0.5 – 2]) frailty scores ($p = 0.13$).

3.5 COSMIN Quality Checklist

Upon assessing the methodological quality of all 52 studies, 51 (98%) had inadequate methodological quality (Table 2). The most common methodological flaws across studies that evaluated construct validity and responsiveness was the absence of a clearly defined hypothesis with magnitude and direction of association specified as well as a lack of information on the expected level of agreement. The absence of one or both criteria downgraded the respective boxes to inadequate methodological quality. A single study performed by Lee et al[73] comparing the Fried Frailty Tool to the GA (criterion validity) was assessed as doubtful methodological quality. Further, due to the inadequate methodological quality, when rating the studies against the "Updated Criteria for Good Measurement Properties", all studies presented indeterminate overall quality. Finally, the quality of evidence per single study or pooled result was graded and resulted in very low quality of evidence. The detailed approach to the COSMIN methodology used to evaluate each step of the risk of bias checklist is summarized in Appendices IV.

Additionally, due to the lack of data on the floor and ceiling effects, minimal important change (MIC) or minimal important difference (MID), response shift, type and ease of administration of the tool, cost, and time associated use of tool, we cannot comment on the interpretability and feasibility of the frailty assessment tools used in each study.

Figure 1: Search Strategy - PRISMA Flow Diagram

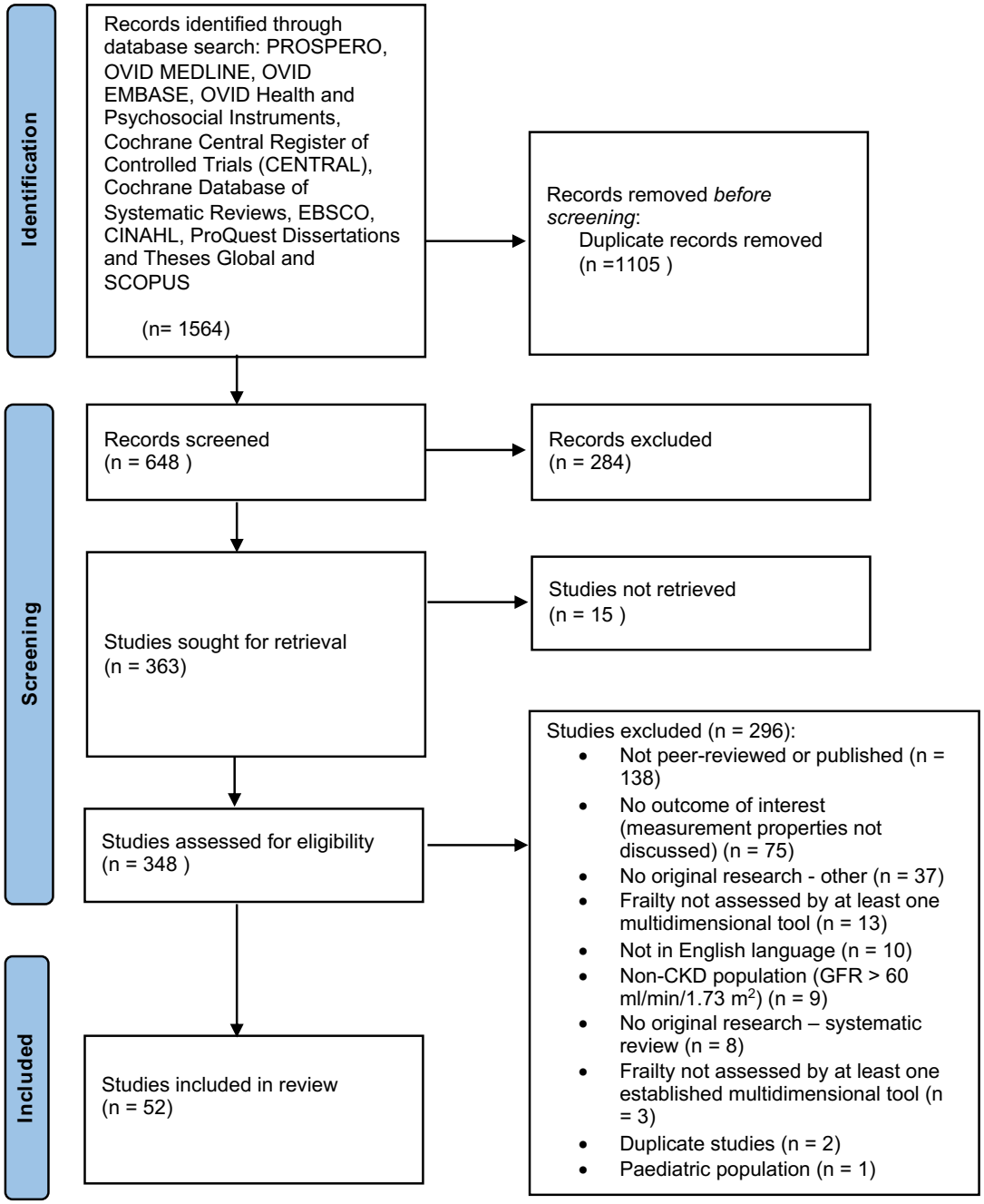
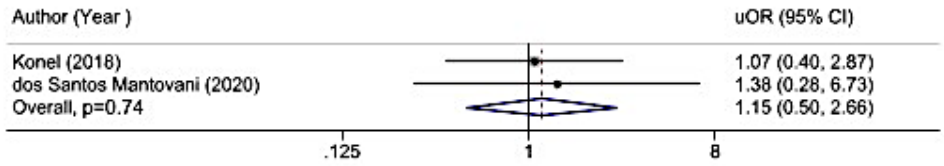
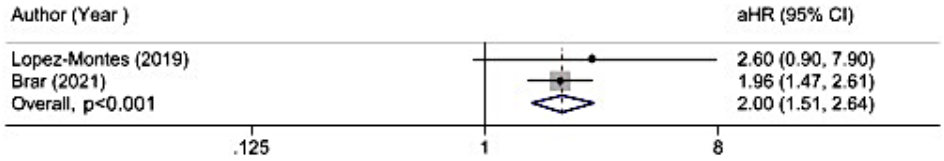


Figure 2: Pooled Outcomes for Risk of Death Based on Frailty Status

Kidney Transplant recipients, unadjusted OR of death for frail versus non-frail based on the Fried Frailty Tool

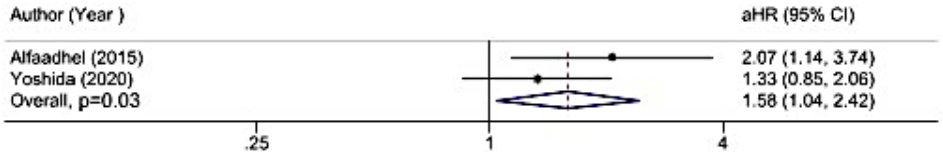


CKD non-dialysis, adjusted HR of death for frail versus non-frail based on the Fried Frailty Tool



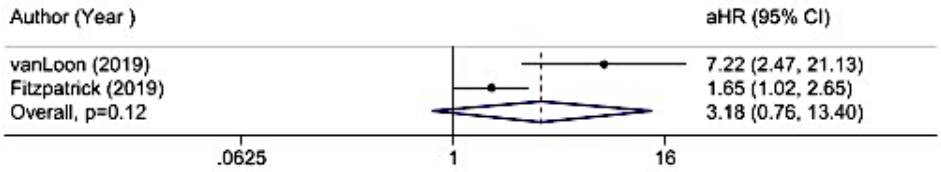
Lopez-Montes adjusted for frailty, age, sex, Charlson index, BMI; Brar adjusted for age, sex, and comorbidity count.

Incident dialysis, adjusted HR of death for CFS (= 4) versus CFS (1 to 3)



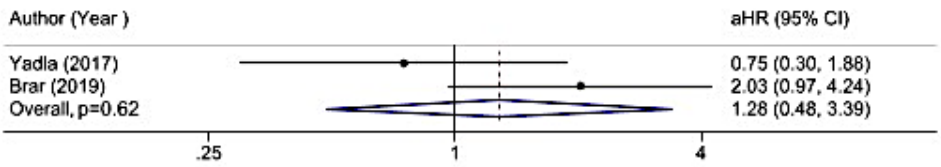
Alfaadhel adjusted for age, race, sex, CCI ≥ 5, diabetic ESKD, GFR, albumin, dialysis modality, location of dialysis start; Yoshida adjusted for CONUT score, CCI, and SPICES score.

Incident dialysis, adjusted HR of death for frail versus non-frail based on the Fried Frailty Tool



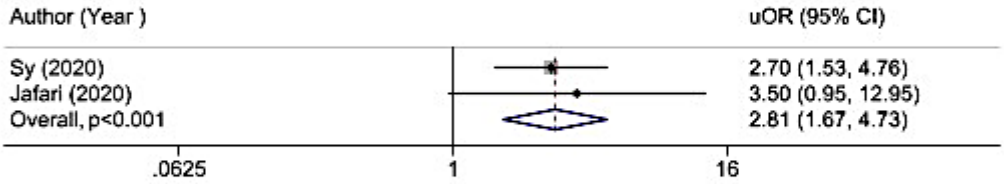
vanLoon adjusted for age, sex, CIRS-G comorbidity burden, smoking, residual renal function and dialysis modality; Fitzpatrick adjusted for age, sex, race, BMI, WHR, CCI, serum albumin, and dialysis vintage.

Prevalent dialysis, adjusted HR of death for frail versus non-frail based on the Fried Frailty Tool



Yadla adjusted for factors unknown; Brar adjusted for age, sex, albumin, hemoglobin, and comorbidity count.

Prevalent dialysis, unadjusted OR of death for frail versus non-frail based on the Fried Frailty Tool



uOR: unadjusted odds ratio; aHR: adjusted hazard ratio; CFS: Clinical Frail Scale; BMI: body mass index; CCI: Charlson Comorbidity Index; ESKD: end-stage kidney disease; GFR: glomerular filtration rate; CONUT: Controlling Nutritional Status Score; SPICES score: skin integrity; problems eating; incontinence; confusion; evidence of falls; and sleep disturbance; CIRS-G: Cumulative Illness Rating Scale – Geriatric; WHR: waist – hip ratio; CAD: coronary artery disease; PTH: parathyroid hormone; p: p-value indicating level of significance.

Table 1: Study Characteristics

Author	Country	Study Design	Number of participants (n)	Mean age, years	Sex, Male %	CKD Status	Dialysis Modality %
KT Recipients							
characteristics presented at admission for KT							
Konel[a] (2018)[54]	United States	prospective longitudinal study	773	54	62.2	NR	overall (74)
Haugen[a] (2020)[55]	United States	prospective longitudinal cohort study	378	55.5	70.1	NR	NR
McAdams-DeMarco[a] (2013)[56]	United States	prospective longitudinal cohort study	383	53.4	60.3	NR	NR
McAdams-DeMarco[a] (2015)[57]	United States	prospective longitudinal cohort study	537	53	60	NR	NR
dos Santos Mantovani (2020)[58]	Brazil	prospective longitudinal cohort study	87	44.7	58.6	NR	HD (81.6), PD (14.9)
Schopmeyer (2018)[59]	Netherlands	prospective cohort study	139	51.81*	62.6	NR	overall (58.3)
Schaenman (2019)[60]	United States	retrospective cohort study	60	52.2	65	NR	overall (80)
Mixed Stage 5 CKD: dialysis and non-dialysis							
Drost[b] (2016)[61]	Netherlands	cross-sectional study	95	65.2	57	ESKD	conventional HD (44), PD (15)

McAdams-DeMarco[c] (2018)[62]	United States	prospective, longitudinal cohort study	1975	53.7	59.5	ESKD	conventional HD (67), PD (14.5)
Nixon (2019)[63]	United Kingdom	prospective cohort study	90	69.1	50	CKD (G4-5) and CKD G5D	conventional HD (33.3)
Haugen[c] (2020)[64]	United States	prospective longitudinal cohort study	3255	54	60	ESKD	overall (54.7)
Lorenz (2019)[65]	United States	retrospective cohort study	272	61.8	62.1	NR	overall (57.4)
van Munster[b] (2016)[66]	Netherlands	prospective cohort study	95	65.2	56.8	ESKD	conventional HD (44.2), PD (14.7)
Lorenz (2020)[67]	United States	prospective cohort study	21	62*	57.1	CKD stage 4: 6 (28.6); stage 5: 15 (71.4)	conventional HD (52.4), PD (14.3)
Nixon (2020)[68]	United Kingdom	prospective cohort study	450	76.3	55.1	NR	overall (18.7)
Chu[a] (2019)[69]	United States	prospective cohort study	569	51.7	60.8	NR	HD (58.3), PD (14.2)
Prevalent Dialysis							
Sy[d] (2020)[70]	United States	retrospective cohort study	425	56.8	57.6	ESKD	conventional HD (100)
Johansen[d] (2014)[11]	United States	cross-sectional study	731	57.3	58.7	ESKD	conventional HD (100)
Chao[e] (2015)[71]	Taiwan	prospective cohort study	46	67.3	43	ESKD	conventional HD (100)
Salter (2015)[72]	United States	cross-sectional study	146	61*	53.4	ESKD	conventional HD (100)

Lee[f] (2017)[73]	South Korea	prospective cohort study	1658	55.9	55.7	ESKD	conventional HD (75.7), PD (24.3)
Yadla (2017)[74]	India	prospective cohort study	205	44.95	69	NR	conventional HD (100)
Kang[f] (2017)[75]	Korea	prospective cohort study	1616	55.9	55.9	NR	conventional HD (77.4), PD (22.6)
Kamijo (2018)[76]	Japan	prospective cohort study	119	66.8	70.6	NR	PD (100)
Garcia-Canton (2019)[77]	Spain	prospective, observational, longitudinal study	277	65*	65.7	NR	overall (100)
Haugen[c] (2019)[78]	United States	prospective, longitudinal cohort study	5423	54	60	ESKD	overall (100)
Brar[g] (2019)[79]	Canada	prospective cohort study	109	54.7	67	advanced CKD	home HD (30.3), PD (69.7)
Bancu (2017)[80]	Spain	retrospective cross-sectional observational study	320	70.26	59.4	CKD stage 5	conventional HD (100)
Jafari (2020)[81]	Canada	prospective cohort study	100	62.86	58	NR	conventional HD (100)
Jegatheswaran (2020)[82]	Canada	cohort study	261	63.3	63	ESKD	HHD (10) , conventional HD (51), PD (39)
Chao[e] (2020)[83]	Taiwan	prospective cohort study	33	69.5	45	ESKD	conventional HD (100)
Incident Dialysis							

McAdams-DeMarco[h] (2015)[84]	United States	prospective longitudinal cohort study	324	54.8	56.5	ESKD	conventional HD (100)
Fitzpatrick[h] (2019)[85]	United States	prospective cohort study	370	54.9	58	ESKD	conventional HD (100)
van Loon[i] (2017)[86]	Netherlands	prospective cohort study	123	76	64	ESKD	conventional HD (76), PD (24)
vanLoon[i] (2019)[87]	Netherlands	prospective cohort study	192	75	67	ESKD	HD (77), PD (23)
Goto[i] (2019)[88]	Netherlands	prospective cohort study	187	75	67	ESKD	conventional HD (77), PD (23)
Alfaadhel[j] (2015)[89]	Canada	prospective cohort study	390	63	67	ESKD	conventional HD (77), PD (23)
Vinson[j] (2020)[90]	Canada	retrospective cohort study	455	62	66	NR	HD (75), PD (25)
Clark (2017)[91]	Canada	prospective cohort study	98	61	58	ESKD	HHD (3), conventional HD (82), PD (15)
Lee (2017)[92]	Korea	prospective cohort study	46	72.3	63	ESKD	conventional HD (100)
Hwang (2019)[93]	Korea	retrospective cohort study	219	79.9*	48.4	ESKD	conventional HD (100)
Yoshida (2020)[94]	Japan	prospective cohort study	310	83.1*	53.5	CKD Stage 4-5	conventional HD (100)
CKD non-dialysis							
Chao[k] (2019)[95]	Taiwan	retrospective cohort study	165,461	61.6	55	NR	NAP
Lee[k] (2020)[96]	Taiwan	retrospective cohort study	52058	62.9	51.8	NR	NAP

Pugh[l] (2016)[97]	United Kingdom	prospective cohort study	283	74	56	advanced CKD	NAP
Pyart[l] (2020)[98]	United Kingdom	retrospective cohort study	1216	78	61.7	NR	NAP
Meulendijks (2015)[99]	Netherlands	prospective cohort study	63	75	65	ESKD	NAP
Ali (2018)[100]	United Kingdom	cross-sectional study	104	77.1	51	NR	NAP
Lopez-Montes (2020)[101]	Spain	prospective cohort study	117	78.1	63.2	Stage 5	NAP
Vettoretti (2020)[102]	Italy	cross-sectional study	112	80	69.6	NR	NAP
Delgado (2015)[103]	United States	cooperative clinical trial	812	52*	60.5	CKD stage 3 to 5	NAP
Brar[g] (2021)[104]	Canada	prospective, observational cohort study	603	68.4	NR	advanced CKD	NAP

* Median age; CKD: Chronic Kidney Disease; ESKD: End Stage Kidney Disease; HD: hemodialysis; HHD: home HD; PD: Peritoneal dialysis; NR: not reported, NAP: not applicable; a-l: overlapping populations

Table 2: Summary of Findings

Fried Frailty Assessment Tool (original and modified)					
Construct Validity - discriminative					
	Risk of Bias	**Quality Criteria**	**Overall Quality**	**GRADE**	**Summarized or Pooled Result**
dos Santos Mantovani (2020)	inadequate	?	?	very low	N = 2 studies; pooled uOR of death frail vs. non-frail: 1.15 (95% CI: 0.50, 2.66), p = 0.74
Konel (2018)	inadequate	?			
Chu (2019)	inadequate	?	?	very low	aHR of death based on frailty transitions, non-frail to frail vs stable non-frail: 1.60 (95% CI: 0.72, 3.56), frail to non-frail vs stable non frail: 1.24 (95% CI: 0.54,2.88), stable frail vs. stable non-frail: 1.58 (95% CI: 0.65, 3.81); aHR LOS for frailty transitions: 1.43 (95% CI: 0.71, 2.89), 0.71 (95% CI: 0.32, 1.57), and 1.68 (95% CI: 0.79, 3.55), respectively
Haugen (2020)	inadequate	?	?	very low	risk of death with frailty in the model - c-statistic (c = 0.7)
McAdams-DeMarco (2013)	inadequate	?	?	very low	aRR early hospital readmission, frail vs. non-frail: 1.59 (95% CI: 1.17, 2.17), p = 0.003; AUC early hospital readmission model with frailty: 0.7
McAdams-DeMarco (2015)	inadequate	?	?	very low	aHR death, intermediately frail vs non-frail: 1.44 (95% CI: 0.69, 3.02); aHR death, frail vs. non-frail: 2.22 (95% CI: 1.03, 4.81); c-statistic death model with frailty: 0.751
McAdams-DeMarco (2018)	inadequate	?	?	very low	with and without frailty in the model and comparing c-statistics (c = 0.646 and 0.642, respectively, p = 0.65); aHR death, intermediately frail vs. non-frail: 1.72 (95% CI 1.03, 2.88), aHR death, frail vs. non-frail: 2.18 (95% 1.25, 3.78)
Haugen (2020)	inadequate	?	?	very low	c-statistic death model with frailty: 0.7

Lorenz (2019)	inadequate	?	?	very low	aHR for death, frail vs non-frail: 7.1 (95% CI: 1.6, 32.4) {adjusted for age}, 11.5 (95% CI: 2.5, 53.6) {adjusted for female sex}, 6.6 (95% CI: 1.5, 29.4) {adjusted for diabetes}, 6.2 (95% CI: 1.4, 28.3) {adjusted for history of CVD}
Sy (2020)	inadequate	?	?	very low	N = 2 studies; pooled uOR of death, frail vs non-frail: 2.81 (95% CI:1.67, 4.73), p < 0.001
Jafari (2020)	inadequate	?			
Lee (2017)	inadequate	?	?	very low	aHR for hospitalization, frail vs. non-frail: 1.83 (95% CI: 1.41, 2.37), prefrail vs non-frail: 1.29 (95% CI: 1.00, 1.67); aHR for death: frail vs. non-frail: 2.08 (95% CI: 1.04, 4.16), prefrail vs. non-frail: 1.01 (95% CI: 0.48, 2.12)
Kang (2017)	inadequate	?	?	very low	aHR of death, frail vs. non-frail: HD participants: 2.35 (95% CI: 1.36, 4.06), p = 0.002; PD participants: 1.75 (95% CI: 0.68, 4.50), p = 0.243)
Yadla (2017)	inadequate	?	?	very low	N = 2 studies, aHR of death, frail vs non-frail: 1.28 (95% CI: 0.48, 3.39), p = 0.62
Brar (2019)	inadequate	?			
Haugen (2019)	inadequate	?	?	very low	adjusted SHR of death, frail vs non-frail: 1.7 (95% CI: 1.36, 2.14)
Bancu (2017)	inadequate	?	?	very low	uOR death, frail vs. non-frail: 2.05 (95% CI: 0.64, 6.55); rate of hospitalization, frail: 0.78, non-frail: 0.28 (p = 0.005)
McAdams-DeMarco (2015)	inadequate	?	?	very low	uOR death, frail vs non-frail: 1.02 (95% CI 0.47, 2.2)

Fitzpatrick (2019)	inadequate	?	?	very low	N = 2 studies, aHR of death, frail vs non-frail: 3.18 (95% CI 0.76, 13.40), p = 0.12
vanLoon (2019)	inadequate	?			
vanLoon (2019)	inadequate	?	?	very low	aOR hospitalization, frail vs. non-frail: 2.31 (95% CI: 1.24, 4.32)
Goto (2019)	inadequate	?	?	very low	aOR functional decline/all-cause mortality, frail vs. non-frail: 1.46 (95% CI: 0.80, 2.68)
Lopez-Montes (2019)	inadequate	?	?	very low	N= 2 studies, pooled aHR of death, frail vs non-frail: 2.00 (95% CI: 1.51, 2.64), p < 0.001
Brar (2021)	inadequate	?			
Delgado (2015)	inadequate	?	?	very low	aHR for death, intermediately frail vs. non-frail: 1.45 (95% CI: 1.13, 1.87), frail vs. non-frail: 1.57 (95% CI: 1.15, 2.10)
Construct validity - convergent					
Drost (2016)	inadequate	?	?	very low	FI vs. Fried: Sens: 90.1, Spec: 100, PPV: 100, NPV 46.7

Nixon (2019)	inadequate	?	?	very low	CFS vs. Fried: Spearman Correlation: 0.77 (95% CI: 0.66, 0.85), AUC: 0.9 (95% CI: 0.84, 0.97), CFS Score ≥ 5 – Sens: 0.79 (95% CI: 0.57, 0.91), Spec: 0.87 (95% CI: 0.78, 0.93), PPV: 0.63 (95% CI: 0.43, 0.79), NPV: 0.94 (95% CI: 0.85, 0.98); PRISMA-7 vs. Fried: corr: 0.64 (95% CI: 0.50, 0.75), AUC: 0.83 (95% CI: 0.73, 0.93), PRISMA Score ≥ 3 – Sens: 0.89 (95% CI: 0.69, 0.97), Spec: 0.61 (95% CI: 0.49, 0.71), PPV: 0.38 (95% CI: 0.25, 0.52), NPV: 0.96 (95% CI: 0.85, 0.99); CKD FI vs. Fried: corr: 0.75 (95% CI: 0.65, 0.81), AUC: 0.88 (95% CI: 0.81, 0.96), CKD FI Score > 0.21 - Sens: 1 (95% CI: 0.83, 1.00), Spec: 0.37 (95% CI: 0.26, 0.48), PPV: 0.3 (95% CI: 0.20, 0.42), NPV: 1 (95% CI: 0.87, 1.00)
Johansen (2014)	inadequate	?	?	very low	self-reported Fried: Sens: 88 (95% CI: 82, 94), Spec: 63 (95% CI: 57, 69), PPV: 52 (95% CI: 45, 59), NPV: 92 (95% CI: 88, 96), Accuracy: 71 (95% CI: 64, 78); modified self-report Fried: Sens: 73 (95% CI: 65, 82), Spec: 89 (95% CI: 85, 93), PPV: 75 (95% CI: 67, 83), NPV: 88 (95% CI: 84, 92), Accuracy: 84 (95% CI: 79, 89)
Salter (2015)	inadequate	?	?	very low	Nephrologist perceived vs. Fried: % agreement: 64.1, kappa: 0.24, corr of scores: 0.32; NP perceived vs. Fried: % agreement: 67, kappa: 0.27, corr of scores: 0.35; Patient perceived vs. Fried: % agreement: 55.5, kappa: 0.07, corr of scores: 0.09
Brar (2019)	inadequate	?	?	very low	Kappa score: SPPB vs. Fried: 0.55, Physician impression vs. Fried: 0.46, Nurse impression vs. Fried: 0.38
Responsiveness - construct					
Lorenz (2020)	inadequate	?	?	very low	median score baseline and post-intervention, respectively: 1 [IQR 1 – 3] and 1[IQR 0.5 – 2], p = 0.13

Clinical Frailty Scale (CFS)					
Construct Validity - Discriminative					
Nixon (2020)	inadequate	?	?	very low	aHR death (frailty continuous): 2.15 (95% CI: 1.63, 2.85); aHR hospitalization (frailty continuous): 1.35 (95% CI: 1.20, 1.53)
Kamijo (2018)	inadequate	?	?	very low	aHR death (subgroups unknown): 9.83 (95% CI: 1.80, 53.7)
Alfaadhel (2015)	inadequate	?	?	very low	N=2 studies, pooled aHR death (score of 4 vs. 1-3): 1.58 (95% CI: 1.04, 2.42), p = 0.03
Yoshida (2020)	inadequate	?			
Hwang (2019)	inadequate	?	?	very low	uOR for death frail vs. non-frail: 4.32 (95% CI: 2.02, 9.23)
Vinson (2020)	inadequate	?	?	very low	SHR time to first EMS-ED, CFS score 3-4 vs. 1-2: 1.89 (95% CI: 1.17, 3.05), CFS score ≥5 vs. 1-2: 2.28 (95% CI: 1.30, 3.98); SHR time to recurrent EMS-ED: CFS score 3-4 vs. 1-2: 1.88 (95% CI: 1.17, 3.01) , CFS score ≥5 vs. 1-2: 2.73 (95% CI: 1.54, 4.84); SHR first EMS-ED (frailty continuous): 1.23 (95% CI:1.10, 1.36)
Pugh (2016)	inadequate	?	?	very low	aHR death (frailty continuous): 1.35 (95% CI:1.16, 1.57), p < 0.001
Pyart (2020)	inadequate	?	?	very low	aHR death (frailty continuous): 1.29 (95% CI: 1.15, 1.45)
Frailty Index (FI)					

Construct Validity - Convergent					
van Munster (2016)	inadequate	?	?	very low	GFI vs. FI: Sens: 89, Spec: 57, AUC: 0.83 (95% CI: 0.74, 0.91), PPV: 54.4, NPV: 89.5; VMS vs. FI: Sens: 77, Spec: 67, AUC: 0.76 (95% CI: 0.65, 0.86), PPV: 57.4, NPV: 83.3
Clark (2017)	inadequate	?	?	very low	CFS ≥4 vs FI >0.21: Sens: 71 (95 % CI: 58, 81), Spec: 72 (95% CI: 53, 87), PPV: 86 (95% CI: 77, 92), NPV: 51 (95% CI: 41, 62); CFS ≥4 vs FI >0.45: Sens: 90 (95% CI: 68, 99), Spec: 51 (95% CI: 39, 62), PPV: 31 (95% CI: 26-37), NPV: 95 (95% CI: 84-99); CFS ≥ 5 vs FI >0.21: Sens: 47 (95% CI: 35, 60), Spec: 100 (95% CI: 88, 100), PPV: 100 (95% CI: 81, 100), NPV: 45 (95% CI: 39, 50); CFS ≥ 5 vs FI >0.45: Sens: 70 (95% CI: 46, 88), Spec: 77 (95% CI: 66, 86), PPV: 43 (95% CI: 31, 55), NPV: 91 (95% CI: 84, 95); FACT-CFS ≥4 vs FI >0.21: Sens: 94 (95% CI: 86, 98), Spec: 48 (95% CI: 29, 68), PPV: 81 (95% CI: 75, 86), NPV: 78 (95% CI: 56, 91); FACT-CFS ≥4 vs FI >0.45: Sens: 95 (95% CI: 75, 100), Spec: 21 (95% CI: 13, 32), PPV: 23 (95% CI: 21, 26), NPV: 94 (95% CI: 70, 99); FACT-CFS ≥ 5 vs FI >0.21: Sens: 62 (95% CI: 50, 74), Spec: 100 (95% CI: 87, 100), PPV: 100 (95% CI: 84, 100), NPV: 53 (95% CI: 45, 60); FACT-CFS ≥ 5 vs FI >0.45: Sens: 85 (95% CI: 62, 97) Spec: 66 (95% CI: 54, 76), PPV: 38 (95% CI: 30, 47), NPV: 95 (95% CI: 86, 98);

					DMMS-Frail vs FI >0.21: Sens: 97 (95% CI: 90-100), Spec: 70 (95% CI: 50, 86), PPV: 88 (95% CI: 81, 93), NPV: 91 (95% CI: 72, 98); DMMS-Frail vs FI >0.45: Sens: 100 (95% CI: 83, 100), Spec: 28 (95% CI: 18, 39), PPV: 26 (95% CI: 23, 28), NPV: 100 (95% CI: 75-100)
Geriatric Assessment (GA)/ CGA					
Construct Validity - Discriminative					
van Loon (2019)	inadequate	?	?	very low	aHR death (≥3 impairments vs. fit): 2.97 (95% CI: 1.19, 7.45); uOR hospitalization (≥3 impairment vs. fit): 1.50 (95% CI: 0.84, 2.65)
Goto (2019)	inadequate	?	?	very low	aOR functional decline/all-cause mortality, frail vs. non-frail: 1.65 (95% CI: 0.81, 3.35)
Lee (2017)	inadequate	?	?	very low	aHR composite, frail vs. non-frail: 23.58 (95% CI: 1.61, 346.03)
Criterion Validity					
Vettoretti (2020)	doubtful	?	?	low	Fried vs. GA (>1 impairment): Sens: 66 (95% CI: 55, 77), Spec: 85 (95% CI: 74, 96), PPV: 88 (95% CI: 83, 93), NPV: 56 (95% CI: 50, 62); Fried vs. GA (> 2 impairments): Sens: 83 (95% CI: 71, 95), Spec: 76 (95% CI: 66, 86), PPV: 89 (95% CI: 85, 93), NPV: 66 (95% CI: 59, 73)

van Loon (2017)	inadequate	?	?	very low	Fried vs. GA: Sens: 59 (95% CI: 48, 70), Spec: 85 (95% CI: 66, 96), PPV: 92 (95% CI: 83, 97), NPV: 41 (95% CI: 34, 39); GFI vs GA: Sens: 74 (95% CI: 64, 83), Spec: 52 (95% CI: 33, 70), PPV: 82 (95% CI: 76, 87), NPV: 40 (95% CI: 29, 52); G8 vs GA: Sens: 92 (95% CI: 85, 97), Spec: 26 (95% CI: 12, 45), PPV: 79 (95% CI 75, 82), NPV: 53 (95% CI: 31, 74); VMS vs GA: Sens: 90 (95% CI: 79, 96), Spec: 38 (95% CI: 19, 59), PPV: 78 (95% CI: 72, 83), NPV: 60 (95% CI: 37, 79)
Responsiveness					
Lee (2017)	inadequate	?	?	very low	median score baseline and 12 months respectively: 8 [IQR 7.0, 9.5], 6 [IQR 5.0, 7.5], p<0.001
Groningen Frailty Indicator (GFI)					
Construct Validity - Discriminative					
Schopmeyer (2019)	inadequate	?	?	very low	uOR death, frail vs. non-frail: 5.13 (95% CI: 0.1, 265.52)
van Loon (2019)	inadequate	?	?	very low	uHR death, frail vs. non-frail: 1.71 (95% CI: 0.76, 3.86); uOR hospitalization, frail vs. non-frail: 1.27 (95% CI: 0.71, 2.67)
Goto (2019)	inadequate	?	?	very low	aOR functional decline/all-cause mortality, frail vs. non-frail: 1.97 (95% CI:1.05, 3.68)
Meulendijks (2015)	inadequate	?	?	very low	uOR death, frail vs. non-frail: 4.18 (95% CI: 1.03, 17.03); uOR hospitalization, frail vs. non-frail: 7.82 (95% CI: 1.61, 37.96)

FRAIL Scale					
Construct Validity - Discriminative					
Jegatheswaran (2020)	inadequate	?	?	very low	OR death (frail + prefrail vs. non-frail): 1.36 (95% CI: 0.65, 2.85); OR hospitalization (frail + prefrail vs. non-frail): 1.92 (95% CI: 1.12, 3.29)
Chao (2019)	inadequate	?	?	very low	aHR death frailty continuous: 1.16 (95% CI: 1.14, 1.19); aHR hospitalization frailty continuous: 1.14 (95% CI: 1.13, 1.15); aHR ICU frailty continuous: 1.17 (95% CI: 1.15, 1.19), $p < 0.001$
Lee (2020)	inadequate	?	?	very low	aHR death, FRAIL score 1 vs FRAIL 0: 1.05 (95% CI: 0.97, 1.13), FRAIL 2 vs FRAIL 0: 1.18 (95% CI: 1.08, 1.29), FRAIL 3 vs FRAIL 0: 1.20 (95% CI: 1.03, 1.39)
Other					
Construct Validity - Discriminative					
Schaenman (2019)	inadequate	?	?	very low	FRS: AUC hospitalization (model with frailty): 0.57
Garcia-Canton (2019)	inadequate	?	?	very low	EFS: aHR death, frail vs. non-frail: 2.34 (95% CI: 1.39, 3.95), vulnerable vs. non-frail: 1.45 (95% CI: 0.75, 2.79); aHR hospitalization, frail vs. non-frail: 1.78 (95% CI: 1.15, 2.77), vulnerable vs. non-frail: 1.82 (95% CI: 1.13, 2.92)
Chao (2020)	inadequate	?	?	very low	CHS: OR death, frail vs. non-frail: 8.33 (95% CI: 1.28, 54.42)
Ali (2018)	inadequate	?	?	very low	PRISMA+ TUG: aHR death, frail vs. non-frail: 4.28 (95% CI: 1.22, 12.98)

Construct Validity - Convergent					
Chao (2015)	inadequate	?	?	very low	Pearson's correlation coefficient : EFS vs. SF: 0.5, p <0.01; FRAIL vs. SF: 0.66, p < 0.01; GFI vs. SF: 0.7, p < 0.01; G8 vs. SF: -0.37, p = 0.01; SF vs TFI: 0.56, p < 0.01; EFS vs. FRAIL: 0.53, p < 0.01; EFS vs. GFI: 0.64, p < 0.01; EFS vs. G8: -0.04, p = 0.81; EFS vs. TFI: 0.45, p < 0.01; FRAIL vs. GFI: 0.49, p < 0.01; FRAIL vs. G8: -0.27, p=0.06; FRAIL vs. TFI: 0.43, p < 0.01; GFI vs G8: -0.22, p =0.13; GFI vs. TFI: 0.59, p < 0.01; G8 vs. TFI: -0.03, p =0.86

?: indeterminate, uOR: unadjusted odds ratio, aOR: adjusted odds ratio; aHR: adjusted hazard ratio; LOS: length of hospital stay; aRR: adjusted relative risk; AUC: area under the curve; CVD: cardiovascular disease; SHR: sub hazard ratio; corr: correlation; Sens: sensitivity; Spec: specificity; PPV: positive predictive value; NPV: negative predictive value; EMS-ED: emergency medical services – emergency department; EFS: Edmonton Frail Scale; FRS: Frailty Risk Score; CHS: Cardiovascular Health Study; SF: Short Form survey; VMS: Hospital Safety Management (VeiligheidsManagementSysteem; CI: confidence interval; HD: hemodialysis; PD: peritoneal dialysis; FI: frailty index; CFS: Clinical Frailty Scale; NP: nurse practitioner; SPPB: short physical performance battery; IQR: inter-quartile range; GFI: Groningen Frailty Indicator; FACT-CFS: Frailty Asessment for Care Planning Tool – Clinical Frailty Scale; DMMS-FRAIL: Dialysis Morbidity and Mortality Study-FRAIL; GA: geratric assessment; aHR composite: consideration of several outcomes

4.0 Discussion

In this systematic review we identified the measurement properties of the existing multidimensional frailty tools in chronic kidney disease (CKD) populations. Fifty-two studies were included, evaluating six common multidimensional frailty tools in the following subpopulations: kidney transplant (KT) recipients, CKD non-dialysis, incident dialysis, prevalent dialysis, and CKD mixed (dialysis and non-dialysis). A large proportion of the data available was from prevalent dialysis patients. In KT recipients, however, only four studies were considered unique due to overlapping participants in the studies, limiting the data available for KT recipients. Across all studies, the following six frailty assessment tools were used most frequently: Fried Frailty Tool, CFS, FI, GA/CGA, GFI, and FRAIL Scale. Overall, the Fried Frailty Tool was used most commonly. From the data presented in these studies, we could only evaluate construct validity (discriminative and convergent), criterion validity (concurrent), and responsiveness. No studies provided data to evaluate reliability, measurement error, structural validity, internal consistency, and content validity. Studies using the Fried Frailty Tool commonly assessed construct validity, criterion validity, and responsiveness. We found that the Fried Frailty Tool demonstrated good discriminative ability in estimating the risk of death in CKD non-dialysis, but not in the incident dialysis and prevalent dialysis populations. The Fried Frailty Tool (score ≥ 3) had adequate criterion (concurrent validity) compared with the GA. Concurrent validity is supported by the high specificity and PPV when compared with more than two impairments on the GA. Further, the GA was the only assessment presenting responsiveness to change in frailty score over twelve months.[96] When assessing the quality of the studies included in this review per the COSMIN Checklist, all but one study[102] presented with inadequate methodological quality. Due to the low methodological rating, when rating all the

studies against the 'Updated Table of Quality Criteria,' a rating of “indeterminate” was selected. As the final step, the GRADE rating of all the included studies was “very low” because of the inadequate methodological quality. Due to the lack of data reported, we could not report on the interpretability and feasibility of the studies in this review.

The majority of the studies included in this review were primarily prospective and to a lesser extent, retrospective cohort studies. We found only one secondary analysis of a randomized clinical trial (RCT). Observational studies, such as the cohort study designs included in this review are advantageous in assessing the association between multiple exposures and outcomes over time in unselected populations.[105] Thus, cohort studies can effectively assess the effect of frailty in CKD on hospital-related outcomes (i.e., risk of death, risk of hospitalization), in turn, evaluating the construct validity and criterion validity of the frailty tools. To assess responsiveness, a longitudinal study design is required to determine whether the measurement tool is responsive to change. Although, cohort studies are longitudinal, they lack precision due to confounding and random variation.[105] RCTs are better able to randomize participants, minimize error, assess the efficacy of the intervention, and measure the effect of the exposure over time.[105] RCTs are also effective at differentiating between minimally important change or difference (MIC/MID) and responsiveness, thus can be useful for assessing measurement error of the measurement tools.[106]

The Fried Frailty Tool, although used most commonly, was also the primary tool with evidence to support construct (discriminative) validity in the CKD non-dialysis subpopulation. In the KT subpopulation, the Fried Frailty Tool failed to provide sufficient evidence of discriminative ability to detect the mortality risk between frailty groups. Due to the limited data available for KT recipients, it is not known whether the results are representative of the

discriminative ability of the Fried Frailty Tool. In the CKD non-dialysis[89, 94] and incident dialysis subpopulations, [101,104] the pooled results were consistent with the hypothesis that the Fried has "sufficient" discriminative ability to predict the risk of death between frailty groups. Finally, in the prevalent dialysis subpopulation, we found conflicting evidence of the Fried Frailty Tool's discriminative validity. First, when assessing the pooled adjusted hazards ratio of two studies, [74,79] we found "insufficient" evidence for the Fried Frailty tool demonstrating discriminative validity. However, when assessing the pooled unadjusted odds ratio of two other studies,[70,81] we found that the results provided "sufficient" evidence for discriminative validity. As compared to the unadjusted results, in the adjusted results the effect size was attenuated and not significant. However, both adjusted HR and unadjusted OR results presented with a magnitude in the same direction. Given that both pooled results are not consistent with each other, we cannot determine whether the Fried Frailty Tool indeed can discriminate between frailty groups in prevalent dialysis patients. Overall, there was limited quality evidence to support the discriminative ability of the Fried tool to predict mortality risk in the CKD non-dialysis, incident dialysis, and prevalent dialysis subpopulations. Additionally, the quality of evidence presented in the studies was limited as all studies were of "inadequate" methodological quality, limiting recommendations about which tests have good construct validity.

The Fried Frailty Tool was the only tool that was evaluated for criterion (concurrent validity) against our clinical gold standard, the Comprehensive Geriatric Assessment (CGA) or Geriatric Assessment (GA). In the CKD non-dialysis subpopulation, the Fried Frailty Tool score ≥ 3 had high specificity and PPV for frailty as measured by more than one impairment on the GA.[102] High specificity and low sensitivity suggests that the Fried Frailty Tool is good for diagnosis. However, when compared to more than two impairments on the GA, a score ≥ 3 on

the Fried Frailty Tool had good sensitivity and PPV. This suggests that the Fried Frailty Index becomes more sensitive with increasing impairments and may be a better tool for frailty screening in higher risk, elderly CKD populations. However, when dysfunction in a single domain is assessed, the Fried Frailty Tool cannot identify those with frailty. This indicates that the Fried Frailty Tool is not an accurate tool to screen for frailty in pre-frail or intermediate frail states in the CKD non-dialysis subpopulation as compared to GA. Based on the results, we can conclude that the Fried Frailty Tool and GA present with "sufficient" evidence for concurrent validity. However, the quality of the studies was limited to "doubtful" as there was no information reported regarding the validation of the tools in CKD non-dialysis populations, level of agreement between the two tools, and whether administration was performed independently. In the incident dialysis subpopulation, similar to our findings in the CKD non-dialysis subpopulation, the Fried Frailty Tool score ≥ 3 presents with good specificity (%) and PPV (%) as compared to more than two impairments on the GA.[86] Thus, qualitatively, the Fried Frailty Tool and the GA had sufficient evidence for concurrent validity; however, the quality of the evidence was "inadequate" due to methodological flaws.

In the incident dialysis subpopulation, the single study[73] that evaluated the responsiveness of the GA provides evidence that the GA was indeed responsive to change in frailty status throughout 12-months. However, the quality of the evidence was limited to "inadequate" due to the absence of a defined hypothesis. In the mixed dialysis studies, the single study performed by Lorenz et al[67] evaluating the responsiveness of the Fried Frailty Tool, did not detect a difference between baseline and post-intervention (exercise) frailty scores. The results suggest that there is limited evidence available for responsiveness of the Fried Frailty Tool in mixed dialysis populations.

While assessing the methodological quality of all fifty-two studies included in this review, all but one study had inadequate methodological quality.[102] The most common reason for inadequate quality was that the authors did not state a hypothesis with direction and magnitude of association. Though studies were designed to study measurement properties, no hypothesis was stated. To remain consistent, we rated each study according to the criteria provided by COSMIN itself for each separate box and the COSMIN-recognized text, Practical Guide to Biostatistics.[27] Our team suggested a set of criteria to improve the consistency of rating for methodological quality (Appendix IV). When rating the overall quality of all the pooled and single studies against the 'Updated Criteria for Good Measurement Properties,' the rating was indeterminate due to the absence of a hypothesis. According to the strobe checklist for observational (cohort) studies, prespecified hypotheses should be included in the report[107]. Due to the absence of hypotheses, when grading the quality of evidence, the increased risk of bias, resulted in a “very low” quality of evidence. The remaining factors such as inconsistency, imprecision, and indirectness were not responsible for the low quality of evidence. Thus, for a tool to have good construct validity and responsiveness, there must be a defined hypothesis by the study a priori to evaluate whether the results correspond to the construct that the tool aimed to measure.

Similar to previous systematic reviews, we aimed to review the literature for all the existing frailty tools and their measurement properties to provide a recommendation for use in clinical practice and research.[108, 109, 110] Previous reviews have focused on evaluating frailty tools in the general or elderly populations.[108, 109, 110] In this review, we aimed to identify the measurement properties of frailty tools in CKD populations. Similar to the approach of Pialoux et al[109] and Sutton et al,[110] only studies evaluating frailty by a multidimensional frailty tool(s)

were included. Alternatively, in the review performed by de Vries and colleagues,[108] unidimensional frailty tools were included and further assessed for their multidimensional nature under content validity. Additionally, de Vries et al[108] and Pialoux et al[109] used an adapted version of methodological quality rating derived from Terwee et al[49] 'Quality criteria for measurement properties of health status questionnaires', to which we cannot directly compare our results as the rating systems differ. Within the measurement properties assessed in both studies, evaluation of structural validity, criterion validity, and measurement error were absent. Similar to this review, de Vries et al[108] also reported that for most frailty tools, only information on construct validity was available. The authors recommended the Frailty Index to be suitable for evaluative (screening) purposes based on quality evidence for construct validity, multidimensionality (content validity), and continuous scoring. Pialoux et al[109] and Sutton et al[110], also reported construct validity to be the most frequently evaluated measurement property. Sutton and colleagues[110] reported that the Tilburg Frailty Indicator (TFI) and FI-CGA were the frailty tools with adequate reliability, validity, and had evidence of fair quality. The Fried Frailty Tool only presented with fair to excellent quality for construct validity. However, in this review the evidence-base was limited for the Fried Frailty Tool under construct validity. Likewise, Pialoux and colleagues[109] also recommended the TFI and a modified version of the FI, the SHARE-FI, as relevant tools for frailty based on the evidence for multiple measurement properties.

4.1 Strengths and Limitations

This review was a complete investigation into all the multidimensional frailty assessment tools used across a broad spectrum of CKD subpopulations. We identified the tools used most frequently and ensured they were multidimensional and adequate for an accurate assessment of the frailty syndrome. We adhered to the PRISMA standards for reporting in systematic reviews and carefully performed each relevant step with two reviewers (A.P. and A.L.). However, there are a few limitations to be addressed. The first limitation of this review is the lack of RCTs available for inclusion. There was only one secondary analysis of an RCT included in this review. RCTs are regarded as the best choice of study for assessing responsiveness. We only had two prospective cohort studies evaluating responsiveness in CKD, which is not enough data to conclude whether a tool would be responsive to change. Measuring the responsiveness of a frailty tool to change is crucial in selecting a research tool to evaluate the effect of interventions on frailty. Second, the lack of data available for KT recipients may have introduced errors in our evaluation of the quality of frailty tools in KT. A majority of the studies had overlapping populations, thus limiting our scope of evaluating existing frailty tools in KT recipients. Third, the COSMIN checklist has many limitations. To assess hypothesis testing for construct validity and responsiveness, COSMIN had recommended to formulate a set of generic hypotheses for evaluating all study results. We believed this would introduce inconsistency and bias into our evaluation of quality as a cohort study should present prespecified hypotheses.[107] Although we did not expect to have inadequate quality for all the studies included, using the standardized COSMIN approach to evaluating measurement properties is a strength of this review. Additionally, for the methodological quality assessment, taking the lowest rating of the items assessed per measurement property, many studies were rated poorly despite adequate ratings for

other items. This is a limitation as the poor rating was primarily a result of other 'methodological flaws' and does not accurately represent all the presented evidence. Finally, the greatest limitation of this review is that we are unable to report on the reliability, measurement error, structural validity, internal consistency, and content validity; all properties crucial for selecting a tool. Due to this, we cannot provide a recommendation for frailty assessment tool(s) for use in clinical research and practice.

4.2 Future Work

There are many gaps in the existing literature for the evaluation of the measurement properties on frailty in CKD. After reviewing all the studies per our inclusion criteria, we could not identify any studies evaluating reliability, measurement error, structural validity, internal consistency, and content validity in this review. The lack of data available on measurement properties indicates that additional studies are required that are specifically designed to evaluate measurement properties. Further work is required to assess these measurement properties of the existing frailty tools in CKD to select and recommend a tool(s) in research and practice. Additionally, within the literature, a gold standard to assess the criterion validity of frailty tools has not been agreed upon. However, provided the CGA's multidimensional nature and identification as a "clinical standard", this suggests that it be used as a gold standard in future work.

Not only are we interested in the application of the frailty assessment tools within CKD but also the utility of the tools. Within the existing literature, studies failed to report on aspects of interpretability and feasibility. It is important to consider what a tool's scores mean qualitatively and how this applies clinically but also the ease, cost, and resources associated with tool

administration. Without this data, it is difficult to select a tool that may be suitable in research or clinical practice. Future work should therefore include evaluating both interpretability and feasibility of the frailty assessment tools in addition to their measurement properties.

In addition to evaluating measurement properties, we must first consider the appropriate study design to assess the property. Observational studies are appropriate when assessing measurement properties such as construct validity and criterion validity or longitudinal outcomes. However, due to confounding in measuring responsiveness in observational studies, this suggests the use of an RCT. RCTs control for confounding ensuring the responsiveness measured by the frailty tool is indeed due to change in frailty status dependent or independent of intervention. RCTs are also appropriate for assessing measurement error and capable of differentiating between error and responsiveness.[106] Further research is needed using randomized controlled trials (RCTs) to assess the role of frailty tools in changing environments.

The evaluation of a frailty assessment tool is only applicable if the study is performed on a sample of the target population. When selecting studies within a specific target population, like CKD, often many studies are excluded or those included contain overlapping participants. The KT recipient subpopulation is one where studies on measurement properties of frailty assessment tools are limited. To accurately assess which frailty tools are effective in assessing KT patients, more unique studies are required to compare tools across studies.

The Fried Frailty Tool has been most extensively examined in terms of its measurement properties across all CKD subpopulations. However, studies evaluating this tool do not report adequate quality evidence supporting its reliability and validity. For a frailty assessment tool to be recommended for application, it must possess high methodological quality and evidence for measurement properties pertaining to its use as diagnostic, evaluative, or predictive. Further

research of adequate quality, specifically in CKD, is needed before the Fried Frailty can be recommended for use in a clinical or research setting.

Based on previous systematic reviews, the Frailty Index (FI) or a modified version of the FI has been identified and recommended for use in the general and elderly populations.[109,110] The FI is a multidimensional tool with a continuous scale, which has a higher sensitivity in detecting frail individuals. Previous studies have assessed the FI in the general population and found evidence for the tool's multiple measurement properties. Further validation of the FI in CKD populations would add to the limited data available for quality frailty tools in CKD. Considering all of the presented gaps in literature, further work encompassing these recommendations will add to the limited data available regarding frailty in CKD.

5.0 Conclusion

Many frailty tools exist; however, only a number have been validated in CKD populations, such as the Fried Frailty Tool, CFS, FI, CGA/GA, FRAIL Scale, and GFI. We aimed to identify the frailty tools validated in CKD and provide a recommendation for a tool(s) for use in clinical research and practice. We can conclude that although the Fried Frailty Tool provided sufficient data for construct (discriminative and convergent) validity and criterion validity, the confidence in these findings is limited by study quality. We also did not identify any studies evaluating the frailty tools' reliability, measurement error, structural validity, internal consistency, and content validity. Additionally, the interpretability and feasibility of the frailty tools used was not reported. As per COSMIN, to provide a recommendation of any tool, a tool must possess adequate structural validity, internal consistency, good interpretability, and feasibility. We therefore cannot provide a recommendation for a tool(s) for use in clinical

research and practice. However, these results highlight important gaps in the existing literature and pave the way for future research in the sphere of frailty and CKD.

6.0 References

1. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people [published correction appears in Lancet. 2013 Oct 19;382(9901):1328]. Lancet. 2013;381(9868):752-762. doi:10.1016/S0140-6736(12)62167-9
2. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci. 2001;56(3):M146-M156. doi:10.1093/gerona/56.3.m146
3. Fried L, Ferrucci L, Darer J, Williamson J, Anderson G. Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care. J Gerontol A Biol Sci Med Sci.2004;59A:255-263.
4. Mitnitski AB, Mogilner AJ, Rockwood K. Accumulation of deficits as a proxy measure of aging. ScientificWorldJournal. 2001;1:323-336. Published 2001 Aug 8. doi:10.1100/tsw.2001.58
5. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci. 2007;62(7):722-727. doi:10.1093/gerona/62.7.722
6. Rockwood K, Mitnitski A. Frailty defined by deficit accumulation and geriatric medicine defined by frailty. Clin Geriatr Med. 2011;27(1):17-26. doi:10.1016/j.cger.2010.08.008
7. Bruyère O, Buckinx F, Beaudart C, et al. How clinical practitioners assess frailty in their daily practice: an international survey. Aging Clin Exp Res. 2017;29(5):905-912. doi:10.1007/s40520-017-0806-8
8. Buta BJ, Walston JD, Godino JG, et al. Frailty assessment instruments: Systematic characterization of the uses and contexts of highly-cited instruments. Ageing Res Rev. 2016;26:53-61. doi:10.1016/j.arr.2015.12.003

9. Montgomery, C. L., Hopkin, G., Bagshaw, S. M., Hessey, E., & Rolfson, D. B. (2021). Frailty inclusive care in acute and community-based settings: a systematic review protocol. Systematic Reviews, 10(1). https://doi.org/10.1186/s13643-021-01638-0
10. Chowdhury R, Peel NM, Krosch M, Hubbard RE. Frailty and chronic kidney disease: A systematic review. Arch Gerontol Geriatr. 2017;68:135-142. doi:10.1016/j.archger.2016.10.007
11. Johansen KL, Dalrymple LS, Delgado C, et al. Comparison of self-report-based and physical performance-based frailty definitions among patients receiving maintenance hemodialysis [published correction appears in Am J Kidney Dis. 2015 Jul;66(1):178]. Am J Kidney Dis. 2014;64(4):600-607.
12. Woo J, Leung J, Morley JE. Comparison of frailty indicators based on clinical phenotype and the multiple deficit approach in predicting mortality and physical limitation. J Am Geriatr Soc. 2012;60(8):1478-1486. doi:10.1111/j.1532-5415.2012.04074.x
13. Wou F, Gladman JR, Bradshaw L, Franklin M, Edmans J, Conroy SP. The predictive properties of frailty-rating scales in the acute medical unit. Age Ageing. 2013;42(6):776-781. doi:10.1093/ageing/aft055
14. Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB (1995) Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. N Engl J Med 332:556–561 doi:10.1056/NEJM199503023320902
15. Guralnik JM, Ferrucci L, Pieper CF, et al. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. J Gerontol A Biol Sci Med Sci. 2000;55(4):M221-M231. doi:10.1093/gerona/55.4.m221

16. Cesari M, Landi F, Calvani R, et al. Rationale for a preliminary operational definition of physical frailty and sarcopenia in the SPRINTT trial. Aging Clin Exp Res. 2017;29(1):81-88. doi:10.1007/s40520-016-0716-1
17. Cesari M, Gambassi G, van Kan GA, Vellas B. The frailty phenotype and the frailty index: different instruments for different purposes. Age Ageing. 2014;43(1):10-12. doi:10.1093/ageing/aft160
18. Searle SD, Mitnitski A, Gahbauer EA, Gill TM, Rockwood K. A standard procedure for creating a frailty index. BMC Geriatr. 2008;8:24. Published 2008 Sep 30. doi:10.1186/1471-2318-8-24
19. Jones DM, Song X, Rockwood K. Operationalizing a frailty index from a standardized comprehensive geriatric assessment. J Am Geriatr Soc. 2004;52(11):1929-1933. doi:10.1111/j.1532-5415.2004.52521.x
20. Lee H, Lee E, Jang IY. Frailty and Comprehensive Geriatric Assessment. J Korean Med Sci. 2020;35(3):e16. Published 2020 Jan 20. doi:10.3346/jkms.2020.35.e16
21. Rodríguez-Mañas L, Féart C, Mann G, et al. Searching for an operational definition of frailty: a Delphi method based consensus statement: the frailty operative definition-consensus conference project. J Gerontol A Biol Sci Med Sci. 2013;68(1):62-67. doi:10.1093/gerona/gls119
22. Veronese N, Custodero C, Cella A, et al. Prevalence of multidimensional frailty and pre-frailty in older people in different settings: A systematic review and meta-analysis [published online ahead of print, 2021 Oct 23]. Ageing Res Rev. 2021;72:101498. doi:10.1016/j.arr.2021.101498

23. Dent E, Kowal P, Hoogendijk EO. Frailty measurement in research and clinical practice: A review. Eur J Intern Med. 2016;31:3-10. doi:10.1016/j.ejim.2016.03.007
24. Pilotto A, Ferrucci L, Franceschi M, et al. Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients. Rejuvenation Res. 2008;11(1):151-161. doi:10.1089/rej.2007.0569
25. Pilotto A, Veronese N, Daragjati J, et al. Using the Multidimensional Prognostic Index to Predict Clinical Outcomes of Hospitalized Older Persons: A Prospective, Multicenter, International Study. J Gerontol A Biol Sci Med Sci. 2019;74(10):1643-1649. doi:10.1093/gerona/gly239
26. Mokkink, L. B., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., Vet, H. C. de, & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs). COSMIN Manual for Systematic Reviews of PROMs. Retrieved from www.cosmin.nl
27. De Vet, H.C.W., Terwee, C.B., Mokkink, L.B., and Knol, D.L. (2011). Measurement in medicine: A practical guide (Cambridge University Press).
28. Chatburn RL. Evaluation of instrument error and method agreement. AANA J. 1996;64(3):261-268.
29. Levey AS, Eckardt KU, Tsukamoto Y, et al. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). Kidney Int. 2005;67(6):2089-2100. doi:10.1111/j.1523-1755.2005.00365.x
30. Drawz P, Rahman M. Chronic kidney disease. Ann Intern Med. 2015;162(11):ITC1-ITC16. doi:10.7326/AITC201506020

31. Worthen G, Tennankore K. Frailty Screening in Chronic Kidney Disease: Current Perspectives. Int J Nephrol Renovasc Dis. 2019;12:229-239. Published 2019 Dec 5. doi:10.2147/IJNRD.S228956
32. Johansen KL, Delgado C, Bao Y, Kurella Tamura M. Frailty and dialysis initiation [published correction appears in Semin Dial. 2015 Jul-Aug;28(4):455]. Semin Dial. 2013;26(6):690-696. doi:10.1111/sdi.12126
33. Denic A, Glassock RJ, Rule AD. Structural and Functional Changes With the Aging Kidney. Adv Chronic Kidney Dis. 2016;23(1):19-28. doi:10.1053/j.ackd.2015.08.004
34. Guerville F, de Souto Barreto P, Taton B, et al. Estimated Glomerular Filtration Rate Decline and Incident Frailty in Older Adults. Clin J Am Soc Nephrol. 2019;14(11):1597-1604. doi:10.2215/CJN.03750319
35. Zhang Q, Ma Y, Lin F, Zhao J, Xiong J. Frailty and mortality among patients with chronic kidney disease and end-stage renal disease: a systematic review and meta-analysis. Int Urol Nephrol. 2020;52(2):363-370. doi:10.1007/s11255-019-02369-x
36. Walker SR, Wagner M, Tangri N. Chronic kidney disease, frailty, and unsuccessful aging: a review. J Ren Nutr. 2014;24(6):364-370. doi:10.1053/j.jrn.2014.09.001
37. Oberg BP, McMenamin E, Lucas FL, et al. Increased prevalence of oxidant stress and inflammation in patients with moderate to severe chronic kidney disease. Kidney Int. 2004;65(3):1009-1016. doi:10.1111/j.1523-1755.2004.00465.x
38. Mihai S, Codrici E, Popescu ID, et al. Inflammation-Related Mechanisms in Chronic Kidney Disease Prediction, Progression, and Outcome. J Immunol Res. 2018;2018:2180373. Published 2018 Sep 6. doi:10.1155/2018/2180373

39. Moorthi RN, Avin KG. Clinical relevance of sarcopenia in chronic kidney disease. Curr Opin Nephrol Hypertens. 2017;26(3):219-228. doi:10.1097/MNH.0000000000000318

40. Costamagna D, Costelli P, Sampaolesi M, Penna F. Role of Inflammation in Muscle Homeostasis and Myogenesis. Mediators Inflamm. 2015;2015:805172. doi:10.1155/2015/805172

41. Biolo G, Cederholm T, Muscaritoli M. Muscle contractile and metabolic dysfunction is a common feature of sarcopenia of aging and chronic diseases: from sarcopenic obesity to cachexia. Clin Nutr. 2014;33(5):737-748. doi:10.1016/j.clnu.2014.03.007

42. Manini TM, Clark BC. Dynapenia and aging: an update. J Gerontol A Biol Sci Med Sci. 2012;67(1):28-40. doi:10.1093/gerona/glr010

43. Upadhyay A, Larson MG, Guo CY, et al. Inflammation, kidney function and albuminuria in the Framingham Offspring cohort. Nephrol Dial Transplant. 2011;26(3):920-926. doi:10.1093/ndt/gfq47

44. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535. Published 2009 Jul 21. doi:10.1136/bmj.b2535

45. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018;319(4):388-396

46. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Accessed via University of Alberta. Available at www.covidence.org

47. Veronese N, Custodero C, Cella A, et al. Prevalence of multidimensional frailty and pre-frailty in older people in different settings: A systematic review and meta-analysis [published

online ahead of print, 2021 Oct 23]. Ageing Res Rev. 2021;72:101498. doi:10.1016/j.arr.2021.101498

48. Rodríguez-Mañas L, Féart C, Mann G, et al. Searching for an operational definition of frailty: a Delphi method based consensus statement: the frailty operative definition-
49. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
50. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. Trials. 2016;17(1):449.
51. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-88
52. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414): 557-560.
53. GRADE. GRADE Handbook - Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013.
54. Konel JM, Warsame F, Ying H, et al. Depressive symptoms, frailty, and adverse outcomes among kidney transplant recipients. *Clin Transplant*. 2018;32(10):e13391. doi:10.1111/ctr.13391
55. Haugen CE, Gross A, Chu NM, et al. Development and Validation of an Inflammatory-Frailty Index for Kidney Transplantation. *J Gerontol A Biol Sci Med Sci*. 2021;76(3):470-477. doi:10.1093/gerona/glaa167
56. McAdams-DeMarco MA, Law A, Salter ML, et al. Frailty and early hospital readmission after kidney transplantation. Am J Transplant. 2013;13(8):2091-2095. doi:10.1111/ajt.12300

57. McAdams-DeMarco MA, Law A, King E, et al. Frailty and mortality in kidney transplant recipients. Am J Transplant. 2015;15(1):149-154. doi:10.1111/ajt.12992

58. Dos Santos Mantovani M, Coelho de Carvalho N, Archangelo TE, et al. Frailty predicts surgical complications after kidney transplantation. A propensity score matched study. PLoS One. 2020;15(2):e0229531. Published 2020 Feb 26. doi:10.1371/journal.pone.0229531

59. Schopmeyer L, El Moumni M, Nieuwenhuijs-Moeke GJ, Berger SP, Bakker SJL, Pol RA. Frailty has a significant influence on postoperative complications after kidney transplantation-a prospective study on short-term outcomes. Transpl Int. 2019;32(1):66-74. doi:10.1111/tri.13330

60. Schaenman J, Castellon L, Liang EC, et al. The Frailty Risk Score predicts length of stay and need for rehospitalization after kidney transplantation in a retrospective cohort: a pilot study. Pilot Feasibility Stud. 2019;5:144. Published 2019 Dec 10. doi:10.1186/s40814-019-0534-2

61. Drost D, Kalf A, Vogtlander N, van Munster BC. High prevalence of frailty in end-stage renal disease. Int Urol Nephrol. 2016;48(8):1357-1362. doi:10.1007/s11255-016-1306-z

62. McAdams-DeMarco MA, Ying H, Thomas AG, et al. Frailty, Inflammatory Markers, and Waitlist Mortality Among Patients With End-stage Renal Disease in a Prospective Cohort Study. Transplantation. 2018;102(10):1740-1746. doi:10.1097/TP.0000000000002213

63. Nixon AC, Bampouras TM, Pendleton N, Mitra S, Dhaygude AP. Diagnostic Accuracy of Frailty Screening Methods in Advanced Chronic Kidney Disease. Nephron. 2019;141(3):147-155. doi:10.1159/000494223

64. Haugen CE, Agoons D, Chu NM, et al. Physical Impairment and Access to Kidney Transplantation. Transplantation. 2020;104(2):367-373. doi:10.1097/TP.0000000000002778

65. Lorenz EC, Cosio FG, Bernard SL, et al. The Relationship Between Frailty and Decreased Physical Performance With Death on the Kidney Transplant Waiting List. Prog Transplant. 2019;29(2):108-114. doi:10.1177/1526924819835803

66. van Munster BC, Drost D, Kalf A, Vogtlander NP. Discriminative value of frailty screening instruments in end-stage renal disease. Clin Kidney J. 2016;9(4):606-610. doi:10.1093/ckj/sfw061

67. Lorenz EC, Hickson LJ, Weatherly RM, et al. Protocolized exercise improves frailty parameters and lower extremity impairment: A promising prehabilitation strategy for kidney transplant candidates. Clin Transplant. 2020;34(9):e14017. doi:10.1111/ctr.14017

68. Nixon AC, Brown J, Brotherton A, et al. Implementation of a frailty screening programme and Geriatric Assessment Service in a nephrology centre: a quality improvement project. J Nephrol. 2021;34(4):1215-1224. doi:10.1007/s40620-020-00878-y

69. Chu NM, Deng A, Ying H, et al. Dynamic Frailty Before Kidney Transplantation: Time of Measurement Matters. Transplantation. 2019;103(8):1700-1704. doi:10.1097/TP.0000000000002563

70. Sy J, McCulloch CE, Johansen KL. Depressive symptoms, frailty, and mortality among dialysis patients. Hemodial Int. 2019;23(2):239-246. doi:10.1111/hdi.12747

71. Chao CT, Hsu YH, Chang PY, et al. Simple self-report FRAIL scale might be more closely associated with dialysis complications than other frailty screening instruments in rural chronic dialysis patients. Nephrology (Carlton). 2015;20(5):321-328. doi:10.1111/nep.12401

72. Salter ML, Gupta N, Massie AB, et al. Perceived frailty and measured frailty among adults undergoing hemodialysis: a cross-sectional analysis. BMC Geriatr. 2015;15:52. Published 2015 Apr 24. doi:10.1186/s12877-015-0051-y

73. Lee SY, Yang DH, Hwang E, et al. The Prevalence, Association, and Clinical Outcomes of Frailty in Maintenance Dialysis Patients. J Ren Nutr. 2017;27(2):106-112. doi:10.1053/j.jrn.2016.11.003

74. Yadla M, John JP, Mummadi M. A study of clinical assessment of frailty in patients on maintenance hemodialysis supported by cashless government scheme. Saudi J Kidney Dis Transpl. 2017;28(1):15-22. doi:10.4103/1319-2442.198102

75. Kang SH, Do JY, Lee SY, Kim JC. Effect of dialysis modality on frailty phenotype, disability, and health-related quality of life in maintenance dialysis patients. PLoS One. 2017;12(5):e0176814. Published 2017 May 3. doi:10.1371/journal.pone.0176814

76. Kamijo Y, Kanda E, Ishibashi Y, Yoshida M. Sarcopenia and Frailty in PD: Impact on Mortality, Malnutrition, and Inflammation. Perit Dial Int. 2018;38(6):447-454. doi:10.3747/pdi.2017.00271

77. Garcia-Canton C, Rodenas A, Lopez-Aperador C, et al. Frailty in hemodialysis and prediction of poor short-term outcome: mortality, hospitalization and visits to hospital emergency services. Ren Fail. 2019;41(1):567-575. doi:10.1080/0886022X.2019.1628061

78. Haugen CE, Chu NM, Ying H, et al. Frailty and Access to Kidney Transplantation. Clin J Am Soc Nephrol. 2019;14(4):576-582. doi:10.2215/CJN.12921118

79. Brar R, Whitlock R, Komenda P, et al. The Impact of Frailty on Technique Failure and Mortality in Patients on Home Dialysis. Perit Dial Int. 2019;39(6):532-538. doi:10.3747/pdi.2018.00195

80. Bancu I, Graterol F, Bonal J, et al. Frail Patient in Hemodialysis: A New Challenge in Nephrology-Incidence in Our Area, Barcelonès Nord and Maresme. J Aging Res. 2017;2017:7624139. doi:10.1155/2017/7624139

81. Jafari M, Kour K, Giebel S, Omisore I, Prasad B. The Burden of Frailty on Mood, Cognition, Quality of Life, and Level of Independence in Patients on Hemodialysis: Regina Hemodialysis Frailty Study. Can J Kidney Health Dis. 2020;7:2054358120917780. Published 2020 May 2. doi:10.1177/2054358120917780

82. Jegatheswaran J, Chan R, Hiremath S, et al. Use of the FRAIL Questionnaire in Patients With End-Stage Kidney Disease. Can J Kidney Health Dis. 2020;7:2054358120952904. Published 2020 Sep 16. doi:10.1177/2054358120952904

83. Chao CT, Huang JW, Chiang CK, Hung KY; COhort of GEriatric Nephrology in NTUH (COGENT) Study Group. Applicability of laboratory deficit-based frailty index in predominantly older patients with end-stage renal disease under chronic dialysis: A pilot test of its correlation with survival and self-reported instruments. Nephrology (Carlton). 2020;25(1):73-81. doi:10.1111/nep.13583

84. McAdams-DeMarco MA, Tan J, Salter ML, et al. Frailty and Cognitive Function in Incident Hemodialysis Patients. Clin J Am Soc Nephrol. 2015;10(12):2181-2189. doi:10.2215/CJN.01960215

85. Fitzpatrick J, Sozio SM, Jaar BG, et al. Frailty, body composition and the risk of mortality in incident hemodialysis patients: the Predictors of Arrhythmic and Cardiovascular Risk in End Stage Renal Disease study. Nephrol Dial Transplant. 2019;34(2):346-354. doi:10.1093/ndt/gfy124

86. van Loon IN, Goto NA, Boereboom FTJ, Bots ML, Verhaar MC, Hamaker ME. Frailty Screening Tools for Elderly Patients Incident to Dialysis. Clin J Am Soc Nephrol. 2017;12(9):1480-1488. doi:10.2215/CJN.11801116

87. van Loon IN, Goto NA, Boereboom FTJ, et al. Geriatric Assessment and the Relation with Mortality and Hospitalizations in Older Patients Starting Dialysis. Nephron. 2019;143(2):108-119. doi:10.1159/000501277

88. Goto NA, van Loon IN, Boereboom FTJ, et al. Association of Initiation of Maintenance Dialysis with Functional Status and Caregiver Burden. Clin J Am Soc Nephrol. 2019;14(7):1039-1047. doi:10.2215/CJN.13131118

89. Alfaadhel TA, Soroka SD, Kiberd BA, Landry D, Moorhouse P, Tennankore KK. Frailty and mortality in dialysis: evaluation of a clinical frailty scale. Clin J Am Soc Nephrol. 2015;10(5):832-840. doi:10.2215/CJN.07760814

90. Vinson AJ, Bartolacci J, Goldstein J, Swain J, Clark D, Tennankore KK. Predictors of Need for First and Recurrent Emergency Medical Service Transport to Emergency Department after Dialysis Initiation. Prehosp Emerg Care. 2020;24(6):822-830. doi:10.1080/10903127.2019.1701157

91. Clark DA, Khan U, Kiberd BA, et al. Frailty in end-stage renal disease: comparing patient, caregiver, and clinician perspectives. BMC Nephrol. 2017;18(1):148. Published 2017 May 2. doi:10.1186/s12882-017-0558-x

92. Lee SW, Lee A, Yu MY, et al. Is Frailty a Modifiable Risk Factor of Future Adverse Outcomes in Elderly Patients with Incident End-Stage Renal Disease?. J Korean Med Sci. 2017;32(11):1800-1806. doi:10.3346/jkms.2017.32.11.1800

93. Hwang D, Lee E, Park S, et al. Validation of risk prediction tools in elderly patients who initiate dialysis. Int Urol Nephrol. 2019;51(7):1231-1238. doi:10.1007/s11255-019-02160-y

94. Yoshida, M., Takanashi, Y., Harigai, T. et al. Evaluation of frailty status and prognosis in patients aged over 75 years with chronic kidney disease (CKD). Ren Replace Ther 6, 60 (2020). https://doi.org/10.1186/s41100-020-00300-0

95. Chao CT, Wang J, Huang JW, Chan DC, Chien KL. Frailty Predicts an Increased Risk of End-Stage Renal Disease with Risk Competition by Mortality among 165,461 Diabetic Kidney Disease Patients. Aging Dis. 2019;10(6):1270-1281. Published 2019 Dec 1. doi:10.14336/AD.2019.0216

96. Lee SY, Wang J, Chao CT, Chien KL, Huang JW. Frailty modifies the association between opioid use and mortality in chronic kidney disease patients with diabetes: a population-based cohort study. Aging (Albany NY). 2020;12(21):21730-21746. doi:10.18632/aging.103978

97. Pugh J, Aggett J, Goodland A, et al. Frailty and comorbidity are independent predictors of outcome in patients referred for pre-dialysis education. Clin Kidney J. 2016;9(2):324-329. doi:10.1093/ckj/sfv150

98. Pyart R, Aggett J, Goodland A, et al. Exploring the choices and outcomes of older patients with advanced kidney disease. PLoS One. 2020;15(6):e0234309. Published 2020 Jun 10. doi:10.1371/journal.pone.0234309

99. Meulendijks FG, Hamaker ME, Boereboom FT, Kalf A, Vögtlander NP, van Munster BC. Groningen frailty indicator in older patients with end-stage renal disease. Ren Fail. 2015;37(9):1419-1424. doi:10.3109/0886022X.2015.1077315

100. Ali H, Abdelaziz T, Abdelaal F, Baharani J. Assessment of prevalence and clinical outcome of frailty in an elderly predialysis cohort using simple tools. Saudi J Kidney Dis Transpl. 2018;29(1):63-70. doi:10.4103/1319-2442.225175

101. López-Montes A, Martínez-Villaescusa M, Pérez-Rodríguez A, et al. Frailty, physical function and affective status in elderly patients on hemodialysis. Arch Gerontol Geriatr. 2020;87:103976. doi:10.1016/j.archger.2019.103976

102. Vettoretti S, Caldiroli L, Porata G, Vezza C, Cesari M, Messa P. Frailty phenotype and multi-domain impairments in older patients with chronic kidney disease. BMC Geriatr. 2020;20(1):371. Published 2020 Sep 29. doi:10.1186/s12877-020-01757-8

103. Delgado C, Grimes BA, Glidden DV, Shlipak M, Sarnak MJ, Johansen KL. Association of Frailty based on self-reported physical function with directly measured kidney function and mortality. BMC Nephrol. 2015;16:203. Published 2015 Dec 9. doi:10.1186/s12882-015-0202-6

104. Brar RS, Whitlock RH, Komenda PVJ, et al. Provider Perception of Frailty Is Associated with Dialysis Decision Making in Patients with Advanced CKD. Clin J Am Soc Nephrol. 2021;16(4):552-559. doi:10.2215/CJN.12480720

105. Carlson MD, Morrison RS. Study design, precision, and validity in observational studies. J Palliat Med. 2009;12(1):77-82. doi:10.1089/jpm.2008.9690

106. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102-109. doi:10.1016/j.jclinepi.2007.03.012

107. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-349. doi:10.1016/j.jclinepi.2007.11.008

108. de Vries NM, Staal JB, van Ravensberg CD, Hobbelen JS, Olde Rikkert MG, Nijhuis-van der Sanden MW. Outcome instruments to measure frailty: a systematic review. Ageing Res Rev. 2011;10(1):104-114. doi:10.1016/j.arr.2010.09.001

109. Pialoux T, Goyard J, Lesourd B. Screening tools for frailty in primary health care: a systematic review. Geriatr Gerontol Int. 2012;12(2):189-197. doi:10.1111/j.1447-0594.2011.00797.x

110. Sutton JL, Gould RL, Daley S, et al. Psychometric properties of multicomponent tools designed to assess frailty in older adults: A systematic review. BMC Geriatr. 2016;16:55. Published 2016 Feb 29. doi:10.1186/s12877-016-0225-2

7.0 Appendices

APPENDIX I: Search Strategy

Ovid MEDLINE(R) ALL <1946 to January 21, 2021>

#	Search Statement	Results
1	exp Renal Insufficiency, Chronic/ or exp Kidney Failure, Chronic/ or (((chronic or "end stage") adj3 (kidney or kidneys or renal)) or CKD or ESKD or ESRD or ((renal or kidney*) adj3 (dialysis or transplant* or replace* or failure or insufficien*))).mp.	392912
2	("Strawbridge questionnaire" or "Multidisciplinary prognostic index" or (Fried adj3 (scale* or phenotype* or criteria)) or "Edmonton frail scale" or "Frail Elder Functional Assessment" or "Groningen frailty indicator" or ("Clinical Frailty" adj3 (Scale or score* or index)) or "Clinical Global impression of Change in Physical Frailty" or CGIC-PF or "Geriatric Functional Evaluation" or "Modified Functional Independence Measure" or "Tilburg frailty indicator" or "G8 questionnaire" or "FRAIL scale" or "Frail index" or "Vulnerable Elders Survey" or (winograd adj3 instrument*)).mp.	1723
3	Psychometrics/ or "Surveys and Questionnaires"/	531109
4	Frailty/ or frailty.mp.	16486
5	3 and 4	785
6	Frailty/ and (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or surve* or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or observation forms or tally sheets or sociometric device* or clinimetric*).mp.	2702
7	(frailty adj4 (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or surve* or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or observation forms or tally sheets or sociometric device* or clinimetric*)).mp.	4281
8	2 or 5 or 6 or 7	6409
9	1 and 8	275

Embase <1974 to 2021 January 21>

#	Search Statement	Results
1	exp chronic kidney failure/ or end stage renal disease/ or renal replacement therapy-dependent renal disease/	137003

2	(((chronic or "end stage") adj3 (kidney or kidneys or renal)) or CKD or ESKD or ESRD or ((renal or kidney or kidneys) adj3 (dialysis or transplant* or replace* or failure or insufficien*))).ti,ab.	433143
3	kidney transplantation/ or kidney allograft/ or kidney autotransplantation/ or kidney graft/	157262
4	1 or 2 or 3	499160
5	frailty.ti,ab,kw. or exp frailty/	26379
6	exp *questionnaire/ or *checklist/ or exp *psychometry/	58445
7	5 and 6	147
8	exp *frailty/ and (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or surve* or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or observation forms or tally sheets or sociometric device* or clinimetric*).ti,ab.	4945
9	(frailty adj3 (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or surve* or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or observation forms or tally sheets or sociometric device* or clinimetric*)).ti,ab.	6420
10	("Strawbridge questionnaire" or "Multidisciplinary prognostic index" or (Fried adj3 (scale* or phenotype* or criteria)) or "Edmonton frail scale" or "Frail Elder Functional Assessment" or "Groningen frailty indicator" or ("Clinical Frailty" adj3 (Scale or score* or index)) or "Clinical Global impression of Change in Physical Frailty" or CGIC-PF or "Geriatric Functional Evaluation" or "Modified Functional Independence Measure" or "Tilburg frailty indicator" or "G8 questionnaire" or "FRAIL scale" or "Frail index" or "Vulnerable Elders Survey" or (winograd adj3 instrument*)).ti,ab,kw.	3009
11	7 or 8 or 9 or 10	9249
12	4 and 11	493

Health and Psychosocial Instruments <1985 to October 2020>

#	Search Statement	Results
1	(((chronic or "end stage") adj3 (kidney or kidneys or renal)) or CKD or ESKD or ESRD or ((renal or kidney or kidneys) adj3 (dialysis or transplant* or replace* or failure or insufficien*))).mp.	186
2	(frailty adj3 (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or surve* or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or observation forms or tally sheets or sociometric device* or clinimetric*)).mp.	9
3	("Strawbridge questionnaire" or "Multidisciplinary prognostic index" or (Fried adj3 (scale* or phenotype* or criteria)) or "Edmonton frail scale" or "Frail Elder Functional Assessment" or "Groningen frailty indicator" or ("Clinical Frailty" adj3 (Scale or score* or index)) or "Clinical Global impression of Change in Physical Frailty" or CGIC-PF or "Geriatric Functional Evaluation" or "Modified Functional Independence Measure" or "Tilburg frailty indicator" or "G8 questionnaire" or "FRAIL scale" or "Frail index" or "Vulnerable Elders Survey" or (winograd adj3 instrument*)).mp.	14
4	2 or 3	20
5	1 and 4	0

CINAHL (EBSCO) Searched January 28, 2021
Limiters/Expanders Search modes - Find all my search terms

#	Query	Results
S1	(MH "Renal Insufficiency, Chronic+") OR (MH "Kidney Failure, Chronic+")	28,482
S2	(MH "Kidney Transplantation+")	11,412
S3	(chronic or "end stage") n3 (kidney or kidneys or renal)) or CKD or ESKD or ESRD or ((renal or kidney*) n3 (dialysis or transplant* or replace* or failure or insufficien*)))	75,599
S4	S1 OR S2 OR S3	75,602
S5	"Strawbridge questionnaire" or "Multidisciplinary prognostic index" or (Fried n3 (scale* or phenotype* or criteria)) or "Edmonton frail scale" or "Frail Elder	1,064

	Functional Assessment" or "Groningen frailty indicator" or ("Clinical Frailty" n3 (Scale or score* or index)) or "Clinical Global impression of Change in Physical Frailty" or CGIC-PF or "Geriatric Functional Evaluation" or "Modified Functional Independence Measure" or "Tilburg frailty indicator" or "G8 questionnaire" or "FRAIL scale" or "Frail index" or "Vulnerable Elders Survey" or (winograd n3 instrument*))	
S6	(MH "Psychometrics") OR (MH "Measurement Issues and Assessments+")	283,089
S7	(MH "Surveys+") OR (MH "Structured Questionnaires") OR (MH "Open-Ended Questionnaires")	244,368
S8	S6 OR S7	498,475
S9	"Frailty"	8,491
S10	S8 AND S9	968
S11	frailty n4 (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or surve* or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or "observation form*" or "tally sheet*" or psychometr* orsociometric or clinimetric*))	2,409
S12	S5 OR S10 OR S11	3,244
S13	S4 AND S12	81

Scopus Searched January 22, 2021 Results=227

(TITLE-ABS-KEY(((chronic or "end stage") w/3 (kidney or kidneys or renal)) or CKD or ESKD or ESRD or ((renal or kidney or kidneys) w/3 (dialysis or transplant* or replace* or insufficien*)))) and ((TITLE-ABS-KEY(frailty w/3 (instrument or instruments or indicies or index* or inventory or inventories or scale or scales or screen or screened or screening or survey or surveys or checklist* or questionnaire* or protocol* or assessment* or evaluat* or tool or tools or "observation form*" or "tally sheet*" or psychometr* or sociometric or clinimetric*))) or (TITLE-ABS-KEY("Strawbridge questionnaire" or "Multidisciplinary prognostic index" or (Fried w/3 (scale* or phenotype* or criteria)) or "Edmonton frail scale" or "Frail Elder Functional Assessment" or "Groningen frailty indicator" or ("Clinical Frailty" w/3 (Scale or score* or index)) or "Clinical Global impression of Change in Physical Frailty" or CGIC-PF or "Geriatric Functional Evaluation" or "Modified Functional Independence Measure" or "Tilburg frailty indicator" or "G8 questionnaire" or "FRAIL scale" or "Frail index" or "Vulnerable Elders Survey" or (winograd w/3 instrument*))))

Proquest Dissertations and Theses Global Searched January 22, 2021 Result=17

noft("Strawbridge questionnaire" OR "Multidisciplinary prognostic index" OR "Fried scale*" OR "Fried phenotype*" OR "Fried criteria" OR "Edmonton frail scale" OR "Frail Elder Functional Assessment" OR "Groningen frailty indicator" OR "Clinical Frailty Scale*" OR "clinical frailty score*" OR "clinical frailty index" OR "Clinical Global impression of Change in Physical Frailty" OR CGIC-PF OR "Geriatric Functional Evaluation" OR "Modified Functional Independence Measure" OR "Tilburg frailty indicator" OR "G8 questionnaire" OR "FRAIL scale" OR "Frail index" OR "Vulnerable Elders Survey" OR "winograd instrument*") OR noft(frailty NEAR/3 instrument* OR frailty NEAR/3 index* OR frailty NEAR/3 indicies OR frailty NEAR/3 inventor* OR frailty NEAR/3 scale* OR frailty NEAR/3 score* OR frailty NEAR/3 screen* OR frailty NEAR/3 survey* OR frailty NEAR/3 checklist* OR frailty NEAR/3 questionnaire* OR frailty NEAR/3 protocol* OR frailty NEAR/3 assessment* OR frailty NEAR/3 evaluat* OR frailty NEAR/3 forms OR frailty NEAR/3 "tally sheet*" OR frailty NEAR/3 sociometric* OR frailty NEAR/3 psychometr* OR frailty NEAR/3 clinimetric*) AND noft("chronic kidney*" or "chronic renal" or "end stage renal" or "end stage Kidney*" or ckd or eskd or esrd or "renal replac*" or dialysis or "renal insufficien*" or "renal transplant*" or "kidney* transplant*")

PROSPERO Searched January 22, 2021 Result =12

Line	Search for	Hits
#1	"Strawbridge questionnaire" OR "Multidisciplinary prognostic index" OR "Fried scale*" OR "Fried phenotype*" OR "Fried criteria" OR "Edmonton frail scale" OR "Frail Elder Functional Assessment" OR "Groningen frailty indicator" OR "Clinical Frailty Scale*" OR "clinical frailty score*" OR "clinical frailty index" OR "Clinical Global impression of Change in Physical Frailty" OR CGIC-PF OR "Geriatric Functional Evaluation" OR "Modified Functional Independence Measure" OR "Tilburg frailty indicator" OR "G8 questionnaire" OR "FRAIL scale" OR "Frail index" OR "Vulnerable Elders Survey" OR "winograd instrument*"	66
#2	"frailty inventor*" or "frailty index*" or "frailty indicies" or "frailty inventor*" or "frailty scale*" or "frailty screen*" or "frailty survey*" or "frailty checklist*" or "frailty questionnaire*" or "frailty protocol*" or "frailty assessment or "frailty evaluat*" or "frailty tool*" or "frailty observation form*" or "frailty tally sheet*"	172
#3	(psychhometr* or sociometric* or clinimetric*) and frailty	2
#4	#1 OR #2 OR #3	183
#5	ckd or eskd or esrd	1007
#6	"chronic kidney*" or "chronic renal*" or "end stage kidney*" or "end stage renal*"	1981
#7	dialysis or "kidney* transplannt*" or "renal transplant*" or "renal replace*" or "renal insufficien*"	1915

#8	#5 OR #6 OR #7	3022
#9	#4 and #8	12

Cochrane Library Searched January 22, 2021 Results =0

ID	Search	Hits
#1	MeSH descriptor: [Renal Insufficiency, Chronic] explode all trees	6780
#2	MeSH descriptor: [Kidney Failure, Chronic] explode all trees	4705
#3	(ckd or eskd or esrd):ti,ab,kw	6958
#4	MeSH descriptor: [Kidney Transplantation] explode all trees	3566
#5	("chronic kidney"):ti,ab,kw	8378
#6	("chronic renal"):ti,ab,kw	2935
#7	("end stage kidney"):ti,ab,kw	524
#8	("end stage renal"):ti,ab,kw	4256
#9	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8	20966
#10	MeSH descriptor: [Psychometrics] explode all trees	2822
#11	MeSH descriptor: [Surveys and Questionnaires] explode all trees	54547
#12	("Strawbridge questionnaire" or "Multidisciplinary prognostic index"):ti,ab,kw	0
#13	("Edmonton frail scale" or "Frail Elder Functional Assessment" or "Groningen frailty indicator" or "Fried scale" or "Fried phenotype" or "fried criteria"):ti,ab,kw	96
#14	("Clinical Global impression of Change in Physical Frailty" or CGIC-PF or "Geriatric Functional Evaluation" or "Modified Functional Independence Measure" or "Tilburg frailty indicator" or "G8 questionnaire" or "FRAIL scale" or "Frail index" or "Vulnerable Elders Survey"):ti,ab,kw	79
#15	("clinical frailty scale" or "clinical frailty score" or "clinical frailty index" or "winograd instrument"):ti,ab,kw	70
#16	("frailty instrument" or "frailty indicies" or "frailty index" or "frailty inventory" or "frailty scale" or "frailty screen" or "frailty survey" or "frailty questionnaire" or "frailty checklist" or "frailty protocol" or "frailty assessment" or "frailty evaluation" or "frailty tool" or "frailty observation form" or "frailty tally sheet"):ti	16
#17	(frailty and (psychometric* or sociometric* or clinimetric*)):ti,ab,kw	3
#18	MeSH descriptor: [Frailty] explode all trees	132
#19	(frailty):ti,ab, kw	1863
#20	#18 or #19	1863
#21	#10 or #11	56365
#22	#20 and #21	50
#23	#12 or #13 or #14 or #15 or #16 or #17 or #22	281
#24	#9 and #22	0

Appendix II: Relevance Form

Reviewer: _____
Study ID: _____

Please assess each screened-in article according to the criteria below.

Preliminary:

1. Is the article written in English, peer-reviewed and published? (no abstracts, no letters, no posters) **Yes** ☐ **No** ☐

2. Does this article contain original research? **Yes** ☐ **No** ☐
If no, does it contain a relevant systematic review?

Population:

3. Were the study participants human?
If so were the participants largely adults (≥18 years)? **Yes** ☐ **No** ☐

4. Did the study population or a subgroup of the population have CKD? **Yes** ☐ **No** ☐
Defined as stage 3-5 CKD (GFR < 60ml/min/1.73 m²), on dialysis or non-dialysis, or renal transplant.

Screening measure:

5. Was frailty assessed in the population using atleast one multidimensional screening tool? **Yes** ☐ **No** ☐
Frailty is defined as a decline in one or more domains of function (physical, social, or psychological). A multidimensional tool studies > 1 domain.

Outcomes:

6. Is any one of the following an outcome in the study? **Yes** ☐ **No** ☐
- ☐ Evaluation of clinimetric properties in frailty assessment tool(s)
  - [ ] Content validity
  - [ ] Reliability
  - [ ] Structural validity
  - [ ] Responsiveness
  - [ ] Internal consistency
  - [ ] Measurement error
  - [ ] Cross-cultural validity
  - [ ] Construct validity
  - [ ] Criterion validity

Study Design:

7. Was the study design any of the following? **Yes** ☐ **No** ☐
- [ ] RCT
- [ ] retrospective cohort
- [ ] prospective cohort
- [ ] ambidirectional cohort
- [ ] cross-sectional
- [ ] case-control

Note: exclusions
a) case reports
b) case series

Final decision:
Should this study be included in the next stage? **Yes** ☐ **No** ☐
(Answer yes if all the above are yes)

☐
Unsure

Consensus decision:

Put into Unsure group for consensus

Yes ☐ **No** ☐ **3rd Party** ☐

APPENDIX III: List of Tables

Table 1: Measurement properties defined by COSMIN for health-related patient reported outcomes

Clinimetric Properties	Definition	How it was measured?	Statistic
Reliability	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions.	**Internal consistency:** The degree of the interrelatedness among the measures.	**Cronbach's α** 0.8 – 0.95 = excellent 0.70 – 0.80 = good 0.60 – 0.70 = satisfactory < 0.60 = suspect
		Test-retest: Over time	**Intraclass coefficient (ICC)**
		Inter-rater reliability: By different persons on the same occasion.	**Cohen's kappa (κ)**
		Intra-rater reliability: By the same persons on different occasions.	**Cohen's kappa (κ)**
Validity	The degree to which an assessment tool measures the construct(s) it is designed to measure	**Content validity: T**he degree to which the scores of an instrument are consistent with the hypotheses based on the assumption that the instrument validly measures the construct to be measured. - **Face validity:** The degree to which instrument measures what it claims to; subjective measure.	**Content validity ratio (CVR)**

		Construct validity: The degree to which the scores of an instrument are consistent with the hypotheses based on the assumption that the instrument validly measures the construct to be measured. - **Structural validity:** The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured. - **Hypothesis testing** - **Convergent validity:** The degree to which two constructs that are related are indeed related. - **Divergent validity:** The test of the absence of a relationship between two constructs with indeed no relationship. - **Cross-cultural validity:** The degree to which the performance of the items adapted or translated from the instrument reflect the original version of the instrument.	**Pearson correlation coefficient (r)** **Pearson correlation coefficient (r)**

		Criterion validity: The degree to which the scores of an instrument are an adequate reflection of a 'gold standard'**.** - **Concurrent validity:** The level of agreement between two assessment tools. - **Predictive validity:** The evaluation of a condition in the present to predict an event in the future.	**Pearson correlation coefficient (r)**
Responsiveness	The ability of an assessment tool to detect change over time in the construct to be measured.	By the measure of the smallest detectable change.	**Mean change in scores ± SD**

Table 2: Boxes for COSMIN Risk of Bias Checklist

Mark the measurement properties that have been evaluated in the article.	
Content validity	
	Box 1. PROM development
	Box 2. Content Validity
Internal structure	
	Box 3. Structural validity
	Box 4. Internal consistency
	Box 5. Cross-cultural validity
Remaining measurement properties	
	Box 6. Reliability
	Box 7. Measurement error
	Box 8. Criterion validity
	Box 9. Hypothesis testing for construct validity
	Box 10. Responsiveness

Table 3: Standards for the assessment of methodological quality of frailty assessment tools

Box 1. PROM development **1a. PROM design**					
General design requirements	**Very good**	**Adequate**	**Doubtful**	**Inadequate**	**N/A**
1. Is a clear description provided of the frailty condition to be measured?	Construct clearly described			Construct not clearly described	
2. Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided to define the construct to be measured?	Origin of the construct is clear		Origin of the construct is not clear		
3. Is a clear description provided of the target population for which the PROM was developed?	Target population clearly described			Target population not clearly described	
4. Is a clear description provided for the context of use?	Context of use clearly described		Context of use not clearly described		
5. Was the PROM development study performed in a sample representing the target population (Frail CKD patients) for which the PROM was developed?	Study is performed in a sample representing the target population	Assume that study was performed in a sample representing the target population, but not clearly described	Doubtful of whether study was performed in a sample representing the target population	Study was not performed in a sample representing the target population **(SKIP ITEMS 6 – 12)**	

Concept elicitation (relevance and comprehensiveness)	**Very good**	**Adequate**	**Doubtful**	**Inadequate**	**N/A**
6. Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative method used, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable for the construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population	
7. Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators /interviewers had limited experience or were trained specifically for the study	Not clear if group moderators /interviewers were trained or group moderators /interviewers not trained and no experience		N/A
8. Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		N/A

9. Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings of interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	N/A
10. Was an appropriate approach used to analyze the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate	
11. Was at least part of the data coded independently?	At least 50% of the data was coded by at least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the coding or only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involved in coding or no coding	N/A
12. Was data collection continued until saturation was reached?	Evidence was provided that saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	N/A
13. For quantitative studies (surveys): was the sample size appropriate?	≥100	50-99	30-49	<30	N/A

Subtotal quality concept elicitation (lowest score of items 6-13)					
Total quality of the PROM (lowest score of items 1-13)					

1b. Cognitive interview study or other pilot test			
Ratings: *V = very good; A= adequate; D= doubtful; I = inadequate; N/A= not applicable*			
	Rater 1	**Rater 2**	**Consensus**
14. Was a cognitive interview study or other pilot test performed? If NO skip items 15-35			
General design requirements			
15. Was the cognitive interview study or other pilot test performed in a sample representing the target population?			
Comprehensibility			
16. Were patients asked about the comprehensibility of the PROM? If NO or not clear, skip items 17-25			
17. Were all items tested in their final form?			
18. Was an appropriate qualitative method used to assess the comprehensibility of the PROM instructions, items, response options, and recall period?			
19. Was each item tested in an appropriate number of patients?			
20. Were skilled interviewers used?			

21. Were the interviews based on an appropriate interview guide?			
22. Were the interviews recorded and transcribed verbatim?			
23. Was an appropriate approach used to analyse the data?			
24. Were at least two researchers involved in the analysis?			
25. Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?			
Subtotal quality of comprehensibility (lowest score of items 15-25)			
Comprehensiveness			
26. Were patients asked about the comprehensiveness of the PROM? If NO or not clear, skip items 27-35			
27. Was the final set of items tested?			
28. Was an appropriate method used for assessing the comprehensiveness of the PROM?			
29. Was each item tested in an appropriate number of patients?			
30. Were skilled interviewers used?			
31. Were the interviews based on an appropriate interview guide?			
32. Were the interviews recorded and transcribed verbatim?			
33. Was an appropriate approach used to analyze the data?			
34. Were at least two researchers involved in the analysis?			

35. Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?			
Subtotal quality of comprehensiveness study (lowest score of items 15, 26-35)			
Total quality of the pilot study (lowest score of items 14-35)			
TOTAL QUALITY OF THE PROM DEVELOPMENT STUDY (Items 1-35)			

Table 4: Quality criteria for good measurement properties. Adapted from COSMIN guidelines

Measurement Property	Rating	Criteria
Structural validity	+	**CTT:** CFA: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08 **IRT/Rasch:** No violation of unidimensionality: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08 *AND* no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 *AND* no violation of monotonicity: adequate looking graphs OR item scalability >0.30 *AND* adequate model fit: IRT: $\chi 2$ >0.01 Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z- standardized values > -2 and <2
	?	CTT: Not all information for '+' reported IRT/Rasch: Model fit not reported
	-	Criteria for '+' not met
Internal consistency	+	At least low evidence for sufficient structural validity[5] AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale
	?	Criteria for "At least low evidence[4] for sufficient structural validity" not met
	-	At least low evidence for sufficient structural validity[5] AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	-	ICC or weighted Kappa < 0.70
Measurement error	+	SDC or LoA < MIC
	?	MIC not defined

	-	SDC or LoA > MIC
Hypothesis testing for construct validity	+	The result is in accordance with the hypothesis
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis
Cross-cultural validity	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's $R^2 < 0.02$)
	?	No multiple group factor analysis OR DIF analysis performed
	-	No multiple group factor analysis OR DIF analysis performed
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard < 0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis OR AUC < 0.70

Legend:

AUC = area under the curve, **CFA** = confirmatory factor analysis, **CFI** = comparative fit index, **CTT** = classical test theory, **DIF** = differential item functioning, **ICC** = intraclass correlation coefficient, **IRT** = item response theory, **LoA** = limits of agreement, **MIC** = minimal important change, **RMSEA:** Root Mean Square Error of Approximation, **SEM** = Standard Error of Measurement, **SDC** = smallest detectable change, **SRMR:** Standardized Root Mean Residuals, **TLI** = Tucker-Lewis index

" + " = sufficient, **" – "** = insufficient, **"?"** = intermediate

Table 5: GRADE approach for grading the quality of evidence. Adapted from the GRADE approach and COSMIN guidelines

Quality of Evidence	Downgrade if
High ➢ **Confident that the true measurement property lies close to the estimate (summary/pooled result) of the measurement property**	**Risk of bias is** • (-1) Serious • (-2) Very Serious • (-3) Extremely serious **Inconsistency is** • (-1) Serious • (-2) Very serious **Imprecision** • (-1) if total sample (n) = 50-100 • (-2) if total sample n < 50 **Indirectness is** • (-1) Serious • (-2) Very serious
Moderate ➢ **Moderately confident that the true measurement property lies close to the estimate (summary/pooled result) of the measurement property***	
Low ➢ **Low confidence in that the true measurement property lies close to the estimate (summary/pooled result) of the measurement property***	
Very Low ➢ **Very little confidence in that the true measurement property lies close to the estimate (summary/pooled result) of the measurement property***	

***** The true result of the measurement property is likely to differ from the estimate of the measurement property.

Table 6: Results of included studies on measurement properties. Adapted from the COSMIN guidelines.

Tool	Country	Structural Validity			Internal Consistency			Cross-Cultural Validity			Reliability		
		n	Meth Quality	Rating/Result	n	Meth Quality	Rating/Result	n	Meth Quality	Rating/Result	n	Meth Quality	Rating/Result
Tool A (ref)													
Overall Pooled or summary result													

Tool	Country	Measurement error			Criterion validity			Hypotheses testing			Responsiveness		
		n	Meth Quality	Rating/Result	n	Meth Quality	Rating/Result	n	Meth Quality	Rating/Result	n	Meth Quality	Rating/Result
Tool A (ref)													
Overall Pooled or summary result													

Table 7: Summary of findings table per measurement property. Adapted from COSMIN guidelines

Structural validity	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Internal consistency	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Cross-cultural validity	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Reliability	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Measurement error	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Criterion validity	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Hypotheses testing	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Responsiveness	Summary/pooled result	Overall rating	Quality of evidence
Tool A			

Appendix IV: Guide to COSMIN

OUTCOMES OBSERVED AND CORRESPONDING BOX TO ASSESS RISK OF BIAS

Outcome	Tool A is a gold standard (ie. CGA) (*See Notes on "gold standard" at the end of this document*)	No gold standard present
Correlation between scores of 2 instruments (Pearson, Spearman, kappa)	Criterion validity – concurrent (box 8)	Construct validity – convergent (box 9a)
Measure of tool accuracy Dichotomous scale: sensitivity, specificity, PPV, NPV, % accuracy; Continuous scale: ROC, AUC	Criterion validity – concurrent (box 8)	Construct validity – convergent (box 9a)
Risk of all-cause mortality or hospitalization (HR, OR, RR, AUC, ROC, c-stat)	Criterion validity – predictive (box 8)	Construct validity – discriminative (box 9b) – 1 tool
Frailty scores or change in frailty status at different time points (ie. baseline and follow-up) – continuous scores (correlation, AUC); dichotomous scale (sensitivity, specificity) *compare standardized response means*	Responsiveness (box 10 a. criterion approach)	Responsiveness (box 10b. construct approach)

****In studies where CGA is only tool being used, evaluate CGA under construct validity e.g. HR Box 9b**

CONSTRUCT VALIDITY

BACKGROUND

Construct validity → 'the degree to which the scores of a measurement instrument are consistent with hypotheses, e.g. with regard to internal relationships, relationships with scores of other instruments or differences between relevant groups'. Construct validity is assessed when there is no gold standard available (ie. criterion validity is not being assessed).

Within construct validity by hypothesis testing, COSMIN has identified the following:

1) **Convergent validity** → the extent to which 2 or more measurement tools capture a common construct (present correlation between the scores)
   Ie. comparing frailty scores between the frailty index, GFI and G8
2) **Discriminative validity** → the extent to which the tool can differentiate the construct of frailty between subgroups
   Ie. comparing scores of the frailty index between dialysis and KT populations

NOTE: When using a multidimensional instrument, each scale or each part of the instrument that measures a specific dimension should be validated, by formulating hypotheses for each dimension separately.

Face validity → at first impression, does this tool appear to adequately measure our construct of interest? (this is subjective). Ie. at first glance does the modified Fried measure our construct of interest? Does the modified Fried contain the criteria needed to assess each domain? This is a quick review of the tool, not in depth.

Content validity →is the measurement tool really measuring what we want it to. Ie. Does the frailty assessment tool accurately measure all dimensions (physical, social, psychological, etc). If the tool is only measuring the physical domain, this is not a valid tool for the multidimensional concept of frailty. We want to check 2 things here: **1) relevance of the tool and 2) comprehensiveness.** Questions to ask yourself:

- Are the items being measured relevant to the study population? (take into consideration disease (CKD, KT), age, gender)
- What is the purpose of the measurement tool being used in the study's context? (aim of the study)
- Are the items being assessed relevant to this aim/purpose (ie. discriminative, sensitive to change, predictive for adverse outcomes)?

NOTE: assessing content validity is qualitative – no statistics will be presented.

Quality Assessment using the COSMIN Risk of Bias tool

Refer to box guide on page 1 for types of outcomes to be assessed under criterion validity.

Hypotheses testing for construct validity – Box 9

9a. Comparison with other outcome measurement instruments (convergent validity)

1. Is it clear what the comparator instruments measures?
    a. very good =The methods describe the domains or criteria each tool is intended to measure; we are assuming all comparator tools are measuring frailty
    b. inadequate = There is no mention of the domains the tool is intended to mention and/or are not sure whether tool intends to measure frailty

    Note: In COSMIN, the "Adequate" and "Doubtful" ratings are greyed out

    Note: This can be tricky, one tool might provide an excellent description and the second tool may not. We usually go with the lowest rating. This is a scenario where "Doubtful" might be a better rating than "Inadequate" but not available from COSMIN.
2. Were the measurement properties of the comparator instrument(s) adequate?

    Note: We are examining 2 things to answer this question

    Population: These details should be mentioned in the Background (or methods)
    a. very good =
        Measurement properties: sufficient (At least 2 measures)
        Population: the population that the tool was validated in is very similar to study population (i.e., validation population matches study population)
        Example of non-matching: validated in CKD population does not correspond to a study that examines those on dialysis
        Example of non-matching: validated in an elderly CKD population does not necessarily correspond to a study that examines a young CKD population
    b. adequate =
        Measurement properties: sufficient (At least 2 measures)
        Population: unsure about population and if applicable to study population (e.g. instrument was validated in a population but unsure if it matches study population, this might be something like a reference provided by the study but no specific mention in the study itself)
    c. doubtful =
        Measurement properties: some information only (1 measure)
        Population: study population is any (can be a general population)
    d. inadequate =
        Measurement properties: No measures mentioned
        Population: no mention

3. Was the statistical method appropriate for the hypotheses to be tested?

    Note: Hypotheses might be close to the end of the introduction section or in the statistical methods section. For this question, we are not judging the quality of the hypothesis, we'll specifically assess the hypothesis in the flaws section.
    a. very good =Statistical method was appropriate (Hypothesis should be stated and statistical method should match the hypothesis)
    b. adequate = Assumable that statistical method was appropriate (Hypothesis not stated but statistical method seems appropriate)

***Note**: Many studies indicate a general aim or objective, which we don't consider to be a hypothesis. But, based on the aim, if we get a sense of what we think the study is trying to do, and we think the methods seems appropriate, the study would be rated "adequate".*

c. doubtful = We can't distinguish between doubtful and inadequate so we are going with the harsher assessment – see below for inadequate.

d. inadequate = Statistical method applied not appropriate (Hypothesis may or may not be stated; method used not appropriate)

4. Were there any other important flaws? (See below after 9b for how to assess this)
   a. very good = No other important methodological flaws
   b. doubtful = Other minor methodological flaws (e.g. only data presented on a comparison with an instrument that measures another construct)
   c. inadequate =Other important methodological flaws (See below after 9b)

9b. Comparison between subgroups (discriminative or known-groups validity)

5. Was an adequate description provided of important characteristics of the subgroups? (e.g. frail vs nonfrail)
   a. very good = Adequate description of the important characteristics of the subgroups
      Important characteristics include: age, sex, CKD status, diabetes, CVD or HF
      Notes: CKD status
      We ae generally expecting eGFR or stage to be presented.
      If participants are KT, but no other information (like eGFR or stage) presented, that is NOT adequate for CKD status
      If participants are on dialysis, but no other information (like eGFR or stage or comment that all were ESRD/ESKD) presented, that is NOT adequate for CKD status
      If participants are ESRD/ESKD, this on it's own is not OK. Must state whether they are on dialysis/non-dialysis AND eGFR.
   b. adequate = Adequate description of most of the important characteristics
      Where "most" = Age, sex, and CKD status must be described by frailty status
   c. doubtful = Poor or no description of the important characteristics of the subgroups
      Studies that do not meet VG or adequate as described above.

6. Was the statistical method appropriate for the hypotheses to be tested?
   Note: Hypotheses might be close to the end of the introduction section or in the statistical methods section. For this question, we are not judging the quality of the hypothesis, we'll specifically assess the hypothesis in the flaws section.

a. very good =Statistical method was appropriate (Hypothesis should be stated and statistical method should match the hypothesis)

b. adequate = Assumable that statistical method was appropriate (Hypothesis not mentioned but statistical method seems appropriate)

***Note**: Many studies indicate a general aim or objective, which we don't consider to be a hypothesis. But, based on the aim, if we get a sense of what we think the study is trying to do, and we think the methods seems appropriate, the study would be rated "adequate".*

c. doubtful = We can't distinguish between doubtful and inadequate so we are going with the harsher assessment – see below for inadequate.
d. inadequate = Statistical method applied not appropriate (Hypothesis may or may not be stated; method used not appropriate)

7. Were there any other important flaws? **(See below for how to assess this**)
   a. very good = No other important methodological flaws
   b. doubtful = Other minor methodological flaws (e.g. only data presented on a comparison with an instrument that measures another construct)
   c. inadequate =Other important methodological flaws

How to assess if there are any important flaws:

1. Has a hypothesis been stated and clearly specifies direction and magnitude of association?
   A vague/general aim or objective is not clear enough.
2. Has the level of agreement, or how large a difference is expected been specified?
   Stating the level of significance or p-value<0.05 do not count
3. If applicable, were the scores for the measurement tool(s) obtained independently?
   If there is only one tool, this point is not evaluated.
4. Were the observed findings reported (point estimate and error/CI) and explained? (i.e., Do results confirm or do not confirm the hypotheses)
   For any result, positive, association, negative, there should be an explanation or acknowledgment/comment.
   Just stating "there was an increased risk of outcome for those that were frail" or something similar, is not enough.
   Study ID 270, first paragraph discussion, is an example of adequate explanation.

3 points applicable:
Very good = All 3 points are "Yes"
Doubtful = Two points are "Yes"
Inadequate = One or less are "Yes"

4 points applicable:
Very good = All 4 points are "Yes"
Doubtful = 2 or 3 points are "Yes"
Inadequate = One or less are "Yes"

CRITERION VALIDITY
BACKGROUND
Criterion validity → 'the degree to which the scores of a measurement instrument are an adequate reflection of a gold standard'.

What does this mean in our study? We do not have a true gold standard frailty assessment tool in CKD. The comprehensive geriatric assessment (CGA) would be our gold standard in this case as it represents the true state of the construct, frailty – this is multidimensional.

Within criterion validity we have: **1) concurrent validity and 2) predictive validity.**

Concurrent validity → when assessing concurrent validity we compare the score of the measurement tool and the 'gold standard' at the **same time.**

Predictive validity → will the measurement tool predict outcomes in the **future** (ie. death, hospitalization). These outcomes have not already occurred when evaluating the tool. (statistics reported should follow the table above). Although HRs can be reported for predictive validity, percentages are preferred (ie. 17% greater risk of death in frail vs. 5% in non-frail). The statistics reported above are considered good quality vs a HR (inadequate?).

Quality Assessment using the COSMIN Risk of Bias tool
Refer to box guide on page 1 for types of outcomes to be assessed under criterion validity.

Criterion validity – Box 8
Example: ID 65, ID 72

1. For continuous scores: Were correlations, or the area under the receiver operating curve calculated?
    a. Yes = very good
    b. No = inadequate (If mentioned in methods but not presented in results, this would be rare)
    c. NA
2. For dichotomous scores: Were sensitivity and specificity determined?
   ***Note**: For something like a HR (OR, RR) where risk is being compared between frail vs non-frail, we'll assess under this question. The continuous frailty score was dichotomized to determine frail/not-frail. Example ID 72.*
    a. Yes = very good
    b. No = inadequate (If mentioned in methods but not presented in results, this would be rare)
    c. NA

3. Were there any other important flaws? (methodological or statistical) (see below for how to assess this)

How to assess if there were any important flaws:
1) **Identify a suitable criterion and method of measurement:**

The study must identify a gold standard (or reference standard) of frailty measurement (ie. This must be the CGA). Determine what level of measurement using the table below.

NOTE: in the case where the CGA is the only tool being used, the study would NOT be assessed under criterion validity. CGA is our criterion and we NEED a comparator to fulfill this box. The CGA would then be assessed under box 9b.

Table 6.2 Overview of statistical parameters for various levels of measurement for the gold standard and measurement instrument under study

Level of measurement		Same units	Statistical parameter
Gold standard	Measurement instrument		
Dichotomous	Dichotomous	Yes	Sensitivity and specificity
	Ordinal	NA	ROC
	Continuous	NA	ROC
Ordinal	Ordinal	Yes	Weighted kappa
		No	Spearman's r^a or other measures of association
	Continuous	NA	ROCs[b]/Spearman's r
Continuous	Continuous	Yes	Bland and Altman limits of agreement or ICC[c]
		No	Spearman's r or Pearson's r

[a] r = correlation coefficient; [b] ROCs: for an ordinal gold standard a set of ROCs may be used, dichotomizing the instrument by the various cut-off points; [c] ICC, intraclass correlation coefficient; NA, not applicable.

2) Is the comparison tool validated in the sample population?
   Check paragraph on tool description in study, look to see if the study reports validation information. If study mentions the tool is validated in the same population that they are studying, then this is a "YES
   Not applicable if there is only a criterion that is assessed here.
3) Has the level of agreement between the comparison tool and gold standard/criterion been identified? (ie. a correlation above 0.70)
   This must be mentioned in the methods. If a threshold is mentioned later (e.g. discussion), this doesn't count!
   ***If this is a HR, OR, RR, etc, and there is no comparison tool, assess criterion only (e.g. look for something like 10% increase or 2-fold increase in risk of outcome as an example)*
4) Were the scores for the gold standard and measurement tool obtained independently?
   Must suggest this in the methods.
   Not applicable if there is only a criterion that is assessed here.
5) Were the observed findings reported (point estimate and error/CI) and explained?

Must report an error/CI
P-values are not considered adequate (see notes below for more on p-value comments from COSMIN)

5 points applicable:
Very Good = all 5 criteria must be met
Doubtful = 3 or 4 must be met
Inadequate = 2 or less must be met

4 points applicable:
Very good = All 4 points are "Yes"
Doubtful = 2 or 3 points are "Yes"
Inadequate = One or less are "Yes"

3 points applicable:
Very good = All 3 points are "Yes"
Doubtful = Two points are "Yes"
Inadequate = One or less are "Yes"

***Note: "Adequate" is greyed out in COSMIN therefore is not an option here*

RESPONSIVENESS
BACKGROUND
Responsiveness → 'the ability of an instrument to detect change over time in the construct to be measured'. There are 2 approaches to assessing responsiveness: 1) criterion approach (where a gold standard has been identified in the study) and 2) construct approach (comparing change scores with other instruments).

What to look for when assessing responsiveness:

1. Was a longitudinal study design used?
2. Were at least 2 measurements taken?
3. Was the study designed such that at least some proportion of the participants would improve or deteriorate on the construct being measured?

Quality Assessment using the COSMIN Risk of Bias tool
Refer to box guide on page 1 for types of outcomes to be assessed under criterion validity.
Responsiveness – Box 10

10a. Criterion approach (comparison to a gold standard)

1. For continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?
    a. very good = Correlations or AUC were calculated
    b. inadequate = Correlations or AUC not calculated

c. NA

2. For dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?
    a. Sensitivity and specificity were calculated = very good
    b. Sensitivity and specificity not calculated = inadequate
    c. NA
3. Were there any other important flaws?
    a. SEE CRITERION APPROACH BOX 8

10b. Construct approach (hypotheses testing; comparison with other outcome measurement instruments)

4. Is it clear what the comparator instrument(s) measure(s)?
    a. The methods describe the domains or criteria the tool is intended to measure; assuming all comparator tools are measuring frailty = very good
    b. There is no mention of the domains the tool is intended to mention and/or are not sure whether tool intends to measure frailty = inadequate

5. Were the measurement properties of the comparator instrument(s) adequate?
    a. The instrument was validated in the CKD population, its measurement properties were discussed in the literature (ie. good predictive validity, test-retest reliability of Fried's phenotype in previous dialysis population studies) = very good
    b. Discussion of the validated tool and measurement properties but no clear mention of population it was validated in (ie. suggests Fried phenotype has been validated and presented good predictive validity) = adequate
    c. Brief mention of measurement property of the tool across all populations (ie. Fried's phenotype is predictive of adverse outcomes related to frailty across all populations) = doubtful
    d. No mention of the tool's measurement properties from previous literature = inadequate

6. Was the statistical method appropriate for the hypotheses to be tested?
    - Use of Pearson's correlation, confirmatory factor analysis, exploratory factor analysis, multitrait-multimethod matrix
    a. Statistical method was appropriate (used any of the above methods) = very good
    b. Assumable that statistical method was appropriate = adequate
    c. Statistical method applied not optimal = doubtful
    d. Statistical method applied not appropriate (none of the above methods used)= inadequate

7. Were there any other important flaws?
    a. No other important methodological flaws = very good
    b. Other minor methodological flaws (e.g. only data presented on a comparison with an instrument that measures another construct) = doubtful
    c. Other important methodological flaws = inadequate

10c. Construct approach (hypotheses testing; comparison between subgroups)
Refer to 9b

10d. Construct approach (hypotheses testing; before and after intervention)

- Ie. Measure change in frailty pre- and post dialysis by presenting change scores.

11. Was an adequate description provided of the intervention given
    Very good = adequate description of intervention
    Timing/duration, what it is
    Doubtful = poor description of intervention
    No mention of timing/duration, vague description
    Inadequate = no description of the intervention
12. Were design and statistical methods adequate for the hypotheses to be tested?
    REFER TO 9A

How to assess if there are any important flaws (applicable to all of 10)

1. Has a hypothesis been stated and clearly specifies direction and magnitude of association?
A vague/general aim or objective is not clear enough.
If hypothesis just says "improves", that is only direction, and not enough. But, ID 763 is an example of where we were okay with language like this.
2. Has the level of agreement been specified, or how large a difference is expected
Stating the level of significance or p-value<0.05 do not count
3. Were the scores for the measurement tool(s) obtained independently?
Score from "before" should not be known when assessing "after"
Scores from other tools should not be known.
4. Were the observed findings reported (point estimate and error/CI) and explained? (in the case of both results confirmed the hypotheses or did not confirm)

4 points applicable:
Very good = All 4 points are "Yes"
Doubtful = 2 or 3 points are "Yes"
Inadequate = One or less are "Yes"

3 points applicable:
Very good = All 3 points are "Yes"
Doubtful = Two points are "Yes"
Inadequate = One or less are "Yes"