J Rehabil Med 2023; 55: jrm00373

ORIGINAL REPORT

INTER-RATER RELIABILITY OF NOVICE LINKERS USING AN INNOVATIVE SEQUENTIAL ITERATIVE LINKING METHOD TO LINK PROSTHETIC OUTCOMES TO THE INTERNATIONAL CLASSIFICATION OF FUNCTIONING, DISABILITY AND HEALTH

Leigh CLARKE, BPO (Hons), MPH[1,2], Emily RIDGEWELL, BPO (Hons), PhD[1], Xia LI, BSc, MSc, PhD[3] and Michael P. DILLON, BPO (Hons), PhD[1]

From the [1]Discipline of Prosthetics and Orthotics, Department of Physiotherapy, Podiatry, Prosthetics and Orthotics, School of Allied Health, Human Services and Sport, La Trobe University, Melbourne, Victoria, [2]The Australian Orthotic Prosthetic Association, Camberwell, Victoria and [3]Office of Research and Infrastructure, La Trobe University, Melbourne, Victoria, Australia

Objective: **When linking outcomes to the International Classification of Functioning, Disability and Health (ICF), inter-rater reliability is typically assessed at the conclusion of the linking process. This method does not allow for iterative evaluation and adaptations that would improve inter-rater reliability as novices gain experience. This pilot study aims to quantify the inter-rater reliability of novice linkers when using an innovative, sequential, iterative linking method to link prosthetic outcomes to the ICF.**

Methods: **Across 5 sequential rounds, 2 novices independently linked outcomes to the ICF. A consensus discussion followed each round that informed refinement of the customized ICF linking rules. The inter-rater reliability was calculated for each round using Gwet's agreement coefficient (AC1).**

Results: **A total of 1,297 outcomes were linked across 5 rounds. At the end of round 1 inter-rater reliability was high (AC1 = 0.74, 95% confidence interval (95% CI) 0.68–0.80). At the end of round 3, inter-rater reliability (AC1 = 0.84, 95% CI 0.80–0.88) was significantly improved and marked the point of consistency where further improvements in inter-rater reliability were not statistically significant.**

Conclusion: **A sequential iterative linking method provides a learning curve that allows novices to achieve high-levels of agreement through consensus discussion and iterative refinement of the customized ICF linking rules.**

Key words: reliability; agreement; International Classification of Functioning, Disability and Health; ICF; linking; prosthetics; amputation; outcomes.

Accepted Dec 19, 2022

J Rehabil Med 2023;55: jrm00373

DOI: 10.2340/jrm.v55.2409

Correspondence address: Leigh Clarke, Discipline of Prosthetics and Orthotics, Department of Physiotherapy, Podiatry, Prosthetics and Orthotics, School of Allied Health, Human Services and Sport, La Trobe University, Melbourne, Victoria, Australia. L.Clarke2@latrobe.edu.au

LAY ABSTRACT

Outcomes are commonly used in healthcare and research to evaluate the effect of an intervention or treatment, such as the effect a prosthesis has on the ability to walk in the community or participate in activities. Cataloguing outcomes using well-established classification systems, such as the International Classification of Functioning, Disability and Health, is important, as it allows outcomes and research to be described using an internationally understood and agreed language. This study aimed to describe an innovative approach to cataloguing outcomes to the ICF, using a method that provides novices with a learning opportunity. In using this innovative method novices were able to catalogue outcomes to the ICF framework with a similar degree of reliability as experts. This will reduce the barriers to novices conducting this type of research in the future.

Linking outcomes to the International Classification of Functioning, Disability and Health (ICF) is a well-established technique for mapping existing research in a way that helps to identify areas of research focus and evidence gaps (1).

The methods used to link outcomes to the ICF are described by ICF Linking Rules (2–4). These linking rules aim to improve the accuracy and reliability with which outcomes are linked to the ICF (4). For example, the ICF linking rules recommend that researchers consider and document the *meaningful concept.* The *meaningful concept* refers to the main concept that the outcome describes (4). As an example, in the following sentence, the *meaningful concept* is probably fatigue, not meal preparation: *At the end of the day are you too tired to cook a meal*? Understanding the *meaningful concept* is important, given the context is key to accurate and reliable linking of outcomes to the ICF. In addition to identifying the *meaningful concept*, the ICF

linking rules also describe the importance of preparatory ICF training (3, 4), study-specific customization of the linking rules, and independent linking by 2 investigators followed by consensus discussion (4).

While these linking rules are an important component of a well-designed method, particularly given their intent to improve ICF linking inter-rater reliability (3), studies using these linking rules report a wide range of inter-rater reliability statistics (Cohen's kappa, $\kappa = 0.41 - 0.98$) (5–8). To some extent, the variation between studies reflects: the ICF level at which outcomes are linked (e.g. ICF component or second-level category) (5, 7), the ambiguity of the outcomes being linked (9), the expertise and familiarity with the topic area (10, 11), as well as the differing experience between expert and novice linkers (6).

Given the number of factors that influence inter-rater reliability and its importance as a quality-control mechanism, investigators routinely report agreement statistics to engender confidence in the reliability of the linking and the trustworthiness of the research conclusions. These data are an important quality control-mechanism (1) given that linking outcomes to the ICF is subjective and requires a high-degree of interpretation based on content-specific knowledge (4) with limited scope to assess accuracy (12).

Typically in ICF linking studies, all outcomes are independently linked, followed by a single consensus discussion, known as an iterative method (1). Using this method, inter-rater reliability is calculated once, after all outcomes are independently linked. This method limits the opportunity for independent linkers to regularly engage in consensus discussions throughout the linking process; discussions that aid the reconciliation of disagreement, build a shared understanding, and inform refinements to the customized linking rules that will probably improve reliability.

There is an opportunity to explore an innovative method of linking to the ICF that may improve reliability, particularly among novice linkers. This innovative method involves independent linking of a proportion of the outcomes in sequential rounds. Each round is followed by a consensus discussion that provides an opportunity to: reflect on the proportion of agreement, identify common cause of disagreement, reconcile these disagreements, and refine the customized ICF linking rules with the intent to improve inter-rater reliability in subsequent rounds. We hypothesize that this innovative approach will improve inter-rater reliability over time (i.e. across *sequential* rounds of linking) given the learning curve that it affords novice linkers.

Therefore, the aim of this pilot study was to quantify the inter-rater reliability of novice linkers over time using the *sequential* iterative ICF linking method to determine; specifically:

- a baseline agreement and inter-rater reliability after novices have completed the preparatory ICF training and the first round of linking,
- the number of outcomes that need to be independently linked to establish a consistent level of agreement and inter-rater reliability, and
- the level of agreement and inter-rater reliability once linkers have completed preparatory ICF training, established study-specific linking rules, and developed experience after all rounds of linking.

METHODS

A 2-part method was used, which included a systematic search to identify the outcomes measured in the existing prosthetics research, followed by an observational study in which participants linked these outcomes to the ICF (Fig. 1).

Ethics approval was obtained from the La Trobe University Human Research Ethics Committee (HREC number 19467).

Part 1: Identify outcomes measured in prosthetic research

A summary of the Part 1 method has been reported here; acknowledge that a more comprehensive description has previously been published (13).

A structured search was used to identify systematic reviews describing the effect of lower-limb prosthetic interventions in: MEDLINE, AMED, Embase, and PsychINFO, Cumulative Index of Nursing and Allied Health Literature (CINAHL), ProQuest Nursing and Allied Health, Web of Science, and the Cochrane Database of Systematic Reviews (CDSR). The search strategy was adopted from a previously published systematic review protocol. From the identified literature, the outcomes used to measure the effect of prosthetic interventions were extracted verbatim to establish a list of prosthetic outcomes.

PART 2: Link outcomes to the International Classification of Functioning, Disability and Health framework

The list of outcomes from Part 1 were linked to the ICF by 2 novice linkers using the recommended linking procedures (3, 4) across 5 sequential iterative linking rounds.

Participants

For the purpose of this study, novice linkers were defined as investigators who had not undertaken ICF linking training nor had prior ICF linking experience. The 2 novice linkers involved in the linking (LC; ER) were experts in the field of orthotics and prosthetics and were tertiary qualified orthotist/prosthetists, each

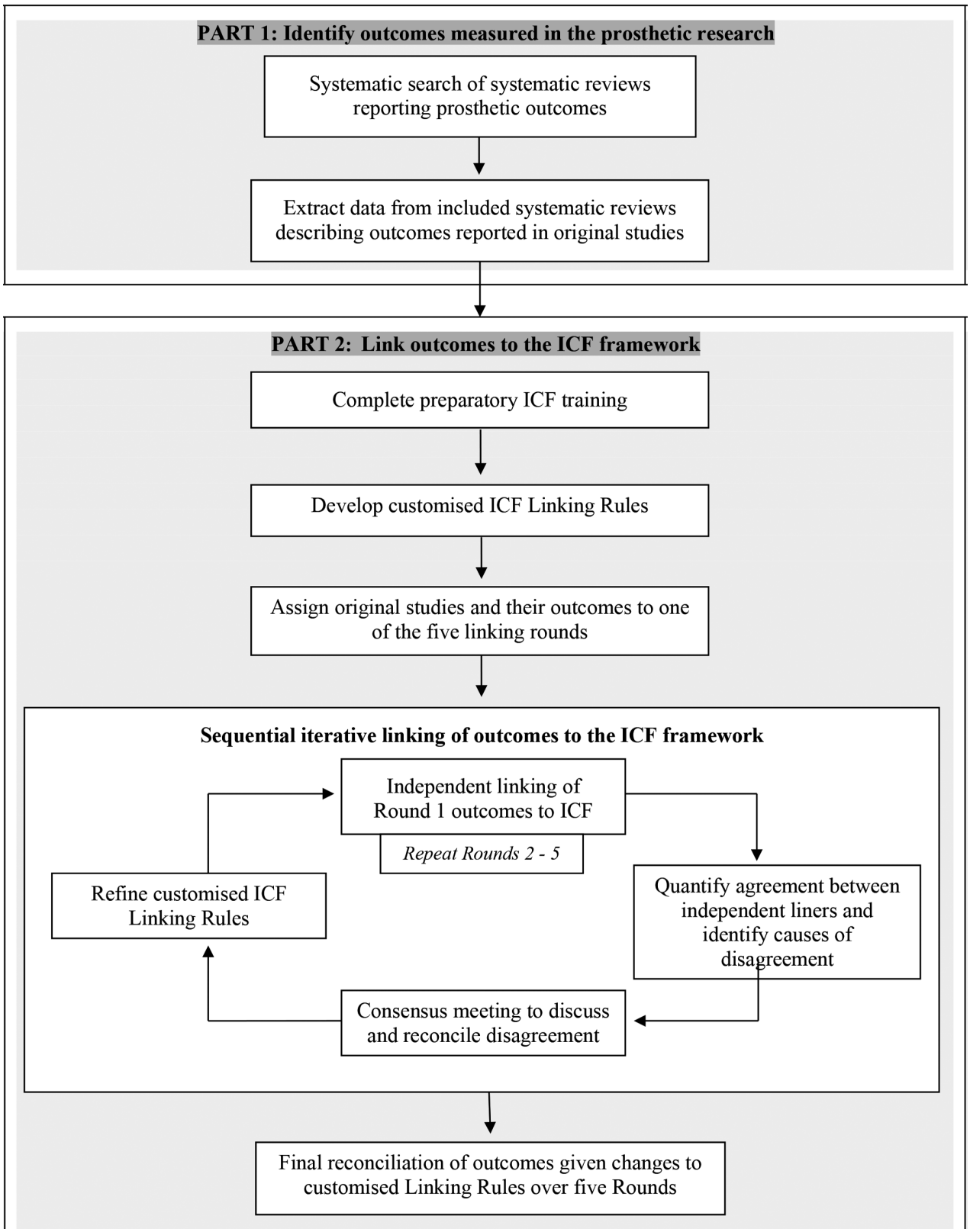

Fig. 1. Two-part method to identify and link outcomes of prosthetic interventions to the International Classification of Functioning, Disability and Health (ICF) framework.

with more than 15 years of experience working across a range of settings, including: clinical practice, research, tertiary education, association management and policy. Each had undertaken postgraduate studies at either the masters or doctoral level.

International Classification of Functioning, Disability and Health Linking training.

Both linkers (LC; ER) independently completed preparatory training (4) that aimed to improve knowledge of the ICF and linking accuracy. The training included completing and passing the ICF Introductory Module online quiz (80%) (17), studying the ICF Beginners Guide (18), Practical Manual (19), and linking rules (4), reviewing prosthetic- and amputation-related ICF linking publications (20–25), as well as familiarization with the ICF Browser Tool (26).

Customization of International Classification of Functioning, Disability and Health linking rules.

The ICF linking rules were customized for this study prior to beginning linking and throughout the iterative process. For example, the linking rules were customized to include a range of study-specific linking examples to assist with assigning outcomes to Not Covered (nc) and Not Covered Health Condition, (nc-hc) (i.e. linking rule 10) (4). Furthermore, during the iterative process the linking rules were further customized to describe an interpretative approach to identify the meaningful concept where the linkers identified the probable meaningful concept from a review of the title and abstract of each original study (i.e. linking rule 2), ensuring researchers considered the context in which the outcome was used (4). The customized ICF linking rules have been published previously (13).

International Classification of Functioning, Disability and Health linking procedure.

In accord with the best-practice ICF linking methods, a sequential iterative linking method was developed for this investigation, which included independent linking by 2 novices followed by a consensus meeting (1, 4).

Over 5 rounds, 20% of the outcomes extracted were randomly chosen and independently linked by 2 novice linkers (LC; ER) to the most specific ICF component, chapter and category possible (i.e. the linking end-point) (27). Thus, while the number of studies assigned to each round was the same, the number of outcomes reported in those studies varied; hence, the differing numbers of outcome linked to the ICF in each round.

Linkers undertook their linking independently and recorded their results in separate custom-designed Excel spreadsheets (Microsoft Corporation, Redmond, WA). One linker merged the independent linking results in a spreadsheet and identified agreement and disagreement with the linking of each outcome. A consensus meeting, attended by only the 2 novice linkers followed. During the consensus meeting each linked item with disagreement was discussed, with both novice linkers presenting their linking reasoning. Discussion followed on each linked item, until agreement on the most accurate linking for each outcome was achieved. If agreement could not be reached, review was undertaken by a third researcher (MD). Consensus meeting notes were taken and the customized ICF linking rules were refined to capture decisions made by the 2 linkers during the consensus meeting, to support improved agreement in subsequent rounds. This process was repeated for each of the 5 rounds.

A final reconciliation was conducted at the conclusion of the 5 sequential rounds of linking to ensure consistency across rounds. The number of revised linking results was recorded.

Descriptive and inferential analysis of International Classification of Functioning, Disability and Health linking reliability.

At the conclusion of each round of linking, descriptive statistics were calculated to:

1. Describe the percentage agreement between 2 independent linkers at each of the component, chapter, and category levels of the ICF.
2. Determine the agreement between 2 independent linkers at the final ICF linking end-point. The categories are described below with supporting examples to aid clarity:
   (a) Disagree: linkers disagreed on the component of the ICF including whether the outcome was "linkable" or not.
   (b) Partial agree: linkers agreed on the component of the ICF. While linkers may have agreed on the component, at some point in the linking disagreement was observed. The partial agreement was coded as either:
      - Partial agree – component: linkers agree on the component, but not the chapter,
      - Partial agree – chapter: linkers agree on the component and chapter, but disagree on the second-level category,
      - Partial agree – second-level category: linkers agree on the component, chapter, and second-level category, but disagree on the third-level category,
      - Partial agree – third-level category: linkers agree on the component, chapter, second- and third-level category, but disagree on the fourth-level category,

(c) Agree – linkers agreed on the same final ICF linking end-point.

3. Describe the proportion of outcomes where the ICF linking result was revised as part of the final reconciliation at the end of the 5 sequential rounds.

Inter-rater reliability was calculated for each of 5 sequential linking rounds. Results were reported for the final ICF linking end-point, as well as stratification by the ICF component, chapter and each category-level.

The Gwet's Agreement Coefficient 1 (AC1) (28) was used in preference to alternatives such as the Cohen's or Fleiss kappa, which are known to result in artificially low agreement statistics in the presence of strong inter-rater agreement (29–31). Using the irrCAC-package (32) in R for Windows 3.6.1 (The R Foundation, Vienna, Austria) inter-rater agreement was reported using the AC1 statistic and associated 95% confidence interval (95% CI), *p*-value, and the Ladis and Koch (33) levels of agreement (e.g. $0.80 > AC1 \leq 1$, *almost perfect*) (28). The level of agreement assigned was determined by the AC1 statistic assuming the cumulative probability was greater than 0.95. Where the cumulative probability was less than 0.95, the lower level of agreement was assigned (28). Cases of perfect agreement between raters (i.e. AC1 = 1) were assigned "not applicable, NA" given there was no possibility to calculate the cumulative probability and assign a level of agreement. While higher AC1 values indicate greater inter-rater agreement, the associated categorical level of agreement should be interpreted with caution, given that, while they are useful to present and help interpret the result, they are arbitrary.

The Cochrane-Armitage test for linear trend was used to determine whether proportionate agreement changed over sequential rounds of linking. Calculations were performed using R for Windows 3.6.1 with a Bonferroni-Holm adjustment to control for the risk of type I error.

RESULTS

A total of 1,297 outcomes were independently linked to the ICF over 5 sequential rounds. Appendix S1 includes the linking results for both novice linkers (Appendix S1, Tab 2), and the final consensus result (Appendix S1, Tab 1) for all outcomes.

Baseline level of inter-rater reliability of novice linkers

In round 1, 235 outcomes were linked to the ICF. Linkers agreed on the linking end-point 74.5% of the time (AC1 = 0.74, 0.68–0.80, $p < 0.001$) (Table I). Where there was disagreement between linkers, this most often occurred in determining the ICF Component (39 of 235, 16.6%) (Table II). For example, linkers differed in the linking of multidimensional instruments (e.g. "Activities Balance Confidence score") and ambiguous outcomes (e.g. "implant removal rate" and "stability") at the Component level (Appendix S1, Tab 3).

The point of consistency in inter-rater reliability

At the end of round 3, 787 outcomes had been linked to the ICF. Linkers agreed on the linking end-point 84.1% of the time (AC1 = 0.84, 95% CI 0.80–0.88, $p < 0.001$), which was a significant improvement over that observed at the end of round 1 (AC1 = 0.74, 0.68–0.80, $p < 0.001$) or round 2 (AC1 = 0.67, 95% CI 0.62–0.73, $p < 0.001$). Given that the inter-rater reliability of the linking end-point at the end of round 3 (84.1%, AC1 = 0.84, 95% CI 0.80–0.88, $p < 0.001$) was comparable to that observed in round 4 (88.0%, AC1 = 0.88, 95% CI 0.84–0.92, $p < 0.001$) and round 5 (92%; AC1 = 0.92, 95% CI 0.88–0.95, $p < 0.001$), round 3 marked the point of consistency in inter-rater reliability (Table I). After the point of consistency, there remained some disagreement; acknowledging that this was often a reflection of the specificity of the linking (Table II). For example, in linking "oxygen consumption", 1 linker progressed linking to the fourth level, whilst the other stopped at the third level (e.g. *b4558* and *b455*) (Appendix S1, Tab 2).

Final inter-rater reliability after completion of 5 sequential iterative linking rounds

In round 5, 218 outcomes were linked to the ICF. Linkers agreed on the linking end-point 91.7% of the time (AC1 = 0.92, 95% CI 0.88–0.95, $p < 0.001$) (Table I).

Table I. Descriptive and inferential statistics describing the inter-rater reliability of the final linking end-point for sequential rounds and total linking

	Round 1	Round 2	Round 3	Round 4	Round 5	Total of all Rounds
Total (N)*	235	269	283	292	218	1297
Missing (*n*)	0	0	0	0	0	0
Disagree (*n*)	60	86	45	35	18	244
Agreement (*n*)	175	183	238	257	200	1,053
% Agreement	74.47	68.03	84.10	88.01	91.74	81.19
AC1 (95% CI)	0.74 (0.68, 0.80)	0.67 (0.62, 0.73)	0.84 (0.80, 0.88)	0.88 (0.84, 0.92)	0.92 (0.88, 0.95)	0.81 (0.79, 0.83)
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Cochrane-Armitage test for linear trend showed significant differences in proportionate agreement between Rounds: 1 vs 3 $p = 0.046$; 1 vs 4 $p < 0.001$; 1 vs 5 $p < 0.001$; 2 vs 3 $p < 0.001$; 2 vs 4 $p < 0.001$; 2 vs 5 $p < 0.001$.

AC1: Gwet's Agreement Coefficient 1; 95% CI: 95% confidence interval).

*The number of studies assigned to each round was the same; however, the number of outcomes reported in those studies varied; hence N varies for each round.

Table II. Descriptive and inferential statistics describing the inter-rater reliability of the final linking end-point for sequential rounds of linking stratified by the ICF Component, Chapter and Second-, Third- and Fourth-level categories

	Component	Chapter	Second level	Third level	Fourth level
Round 1					
Missing	0	27	0	116	29
Disagree (N)	39	2	11	8	0
Perfect (N)	196	167	156	32	3
Total (N)*	235	196	167	156	32
% Agreement	83.40	98.82	93.41	80.00	100.00%
AC1 (95% CI)	0.82 (0.77, 0.87)	0.99 (0.97, 1)	0.93 (0.89, 0.97)	0.78 (0.64, 0.93)	1 (NA, NA)
p-value	< 0.001	< 0.001	< 0.001	< 0.001	NA
Round 2					
Missing	0	34	1	91	57
Disagree (N)	52	1	21	12	NA
Perfect (N)	217	182	160	57	NA
Total (N)*	269	217	182	160	57
% Agreement	80.67	99.45	88.40	82.61	NA
AC1 (95% CI)	0.79 (0.74, 0.84)	0.99 (0.98, 1)	0.88 (0.83, 0.93)	0.82 (0.72, 0.91)	NA
p-value	< 0.001	< 0.001	< 0.001	< 0.001	NA
Round 3					
Missing	0	30	3	160	43
Disagree (N)	30	3	9	3	0
Perfect (N)	253	220	208	45	2
Total (N)*	283	253	220	208	45
% Agreement	89.40	98.65	95.85	93.75	100.00
AC1 (95% CI)	0.89 (0.85, 0.93)	0.99 (0.97, 1)	0.96 (0.93, 0.99)	0.93 (0.86, 1)	1 (NA, NA)
p-value	< 0.001	< 0.001	< 0.001	< 0.001	NA
Round 4					
Missing	0	37	2	140	77
Disagree (N)	22	1	7	5	0
Perfect (N)	270	232	223	78	1
Total (N)*	292	280	232	273	78
% Agreement	92.47%	99.57%	96.96%	93.98%	100.00%
AC1 (95% CI)	0.92 (0.89, 0.95)	1.00 (0.99, 1.00)	0.97 (0.95, 1.00)	0.94 (0.88, 0.99)	NA
p-value	< 0.001	< 0.001	< 0.001	< 0.001	NA
Round 5					
Missing	0	35	0	108	54
Disagree (N)	11	0	4	3	0
Perfect (N)	207	172	168	57	3
Total (N)*	218	207	172	168	57
% Agreement	94.95	100.00	97.67	95.00	100.00
AC1 (95% CI)	0.95 (0.92, 0.98)	1 (1, 1)	0.98 (0.95, 1.00)	0.95 (0.89, 1)	1 (NA, NA)
p-value	< 0.001	< 0.001	< 0.001	< 0.001	NA

Cases of perfect agreement between raters (i.e. AC1 = 1) were assigned "not applicable, NA" given there was no possibility to calculate the cumulative probability and assign a level of agreement.

AC1: Gwet's Agreement Coefficient 1; 95% CI: 95% confidence interval; NA: not applicable.

*The number of studies assigned to each round was the same; however, the number of outcomes reported in those studies varied; hence, the total outcomes (N) varies for each round.

Impact of final reconciliation

The reconciliation at the conclusion of all sequential rounds resulted in revision to the linking of 35 of the 1,297 outcomes (2.7%) (Appendix S1, Tab 1).

DISCUSSION

This study investigated the baseline level of agreement of novice ICF linkers, the final agreement when a *sequential* iterative linking method was used, the number of outcomes needed to be linked to achieve consistent agreement, and the impact of final reconciliation on linking reclassification.

Baseline level of inter-rater reliability of novice linkers

The baseline agreement observed at the end of round 1 (74%) was similar to that reported in studies involving expert linkers (59–93%) (6, 10). The finding that novice linkers could achieve a similar level of agreement as experts affirms the value of a method designed in accord with ICF linking rules (4) that includes: the completion of recommended ICF training, study-specific preparatory ICF training, and the development of study-specific customized linking rules.

Study-specific ICF training and customization of the linking rules, are features of the method designed to improve the reliability with which outcomes are linked to the ICF (1, 3, 4, 34). In particular, the completion of study-specific ICF training allowed novice linkers to develop a familiarity with the outcomes and the second-level categories that would probably be linked in this study. Furthermore, customization of the study-specific linking rules over sequential rounds of linking provided linkers with a learning curve that allowed a shared understanding to be built over time. For example, the linking rules were customized to facilitate identification of the meaningful concept, as required

by linking rule 2 (4). Further customization prior to initiating linking, such as establishing an approach to link multidimensional instruments and ambiguous outcomes, would have probably led to greater inter-rater reliability at baseline. As a generalization, studies that do not incorporate these aspects in their method design have reported lower inter-rater reliability (34, 35). The degree to which rigorous method design can improve inter-rater reliability can be seen in a study in which authors re-linked after the establishment of customized linking rules showing that inter-rater reliability improved from a Cohen's kappa (κ) of 0.746 to 0.902 (34).

Refinements to the study-specific linking rules to improve inter-rater reliability

A number of refinements were made to the study-specific linking rules, particularly at the end of round 1, which helped improve agreement in subsequent rounds (13). We provide the following insights of refinements and lessons learnt from adopting the *sequential* iterative linking method.

At the end of round 1, a refinement was made to specify how multi-dimensional instruments would be linked to the ICF given the disagreement observed between independent linkers. For example, in linking the Questionnaire for Persons with Transfemoral Amputation (Q-TFA), 1 linker considered *all* of the individual concepts captured by the Q-TFA questions (e.g. Mobility, d4; Products and Technology e1; Global health, nd-gh) (36) and therefore could not link the Q-TFA to a specific component (Not Defined, nd). Disagreement occurred as the second linker considered the *majority* of individual concepts captured by the Q-TFA questions and therefore linked this outcome at the component level (Mobility, d4).

In addition, the customized ICF linking rules were refined to ensure that the *likely meaningful concept* of each outcome was considered during linking. The method did not include this rule at the outset, given that outcomes were extracted from systematic reviews that did not provide sufficient detail to confirm the context of what was being measured. At the end of round 1, there was considerable disagreement between independent linkers who had differing interpretations of the *meaningful concept* of each outcome. For example, in linking the outcome "peak pylon accelerations", 1 linker interpreted the outcome as describing the gait pattern of the prosthetic limb (e.g. motion of pylon during walking) and therefore linked to Gait Pattern Functions (b770). Disagreement occurred as the second linker interpreted the outcome as describing the mechanical function of the pylon and linked to Assistive Products and Technology for Personal Use in Daily Living (e1151). This differing interpretation highlights the importance of considering the context in which the outcome was measured (1, 4), as was captured through refinement to the ICF linking rules to consider the *likely meaningful concept* based on the title and aim of a study.

Finally, refinements were made to the customized linking rules for high-volume outcomes with multiple linking options, given these led to a high-level of disagreement. For example, in the round 1 consensus meeting, linkers noted that the outcome "skin complications", including callouses and ulcers, could be considered normal protective functions (Protective Functions of the Skin, b810) or a health condition (Not Defined – Health Condition, nc-hc). The linkers chose to specify and document rules for these high-volume and ambiguous outcomes to support consistency and reliability in subsequent linking rounds. This specified rule acknowledged that some outcomes could be correctly linked to a number of ICF categories, and therefore ongoing disagreement could be avoided in subsequent rounds.

The point of consistency in inter-rater reliability using a sequential iterative linking method

We hypothesized that linkers would become more reliable with iterative practice and that at some point, consistency would be associated with the absolute number of linked outcomes. Consistency in inter-rater reliability was achieved at the end of round 3 and was probably due to the achievement of linking saturation, rather than the absolute number of linked outcomes. For example, at the end of round 3, outcomes had been linked to 37 of the 41 (90%) ICF second-level categories that were ultimately used across the 5 rounds. As such, in rounds 4 and 5, few outcomes were linked to new ICF second-level categories, which suggests linking saturation had occurred (Appendix S1, Tab 3).

In our opinion, the point at which linking saturation occurs will vary across studies depending on the heterogeneity and volume of outcomes being linked. For example, studies with heterogeneous outcomes that link broadly across ICF second-level categories may not achieve saturation until a large proportion of the linking has been completed. In contrast, studies with homogenous outcomes that link narrowly to the ICF may achieve linking saturation using just a small proportion of the total outcomes. The point of linking saturation (37) may provide an indication that linking reliability will be maintained in subsequent rounds. Results from this study suggest that the point of linking saturation could be determined by calculating the number of new ICF second-level categories linked in each round. When few new ICF categories are linked, saturation has probably been reached.

Final inter-rater reliability using a sequential iterative linking method

Given the sequential iterative linking method used, there was a high level of agreement between linkers in

the fifth and final round (91.74%). Collectively, over the 5 rounds, a high level of agreement (81.19%) was observed between novice linkers.

The high level of agreement achieved in this pilot is similar to that of studies involving linking experts (i.e. authors of the ICF linking rules) (6, 10, 38) and is higher than that reported in a prior study of linkers with mixed experience (novices and experts) (Cohen's kappa (κ) 0.69; 55–88%) (6). This suggests that novice linkers can achieve inter-rater reliability that is similar to expert linkers when the method is designed to include: study-specific preparatory ICF training, study-specific customisation of the ICF linking rules, and use of a sequential iterative linking method. In particular, the use of a sequential iterative linking methods provides a learning curve for novice linkers to develop a shared understanding and refine the linking rules over time.

At the conclusion of the 5 sequential iterative linking rounds a reconciliation was undertaken, acknowledging that decisions and rule refinements occurred that may result in linking variation across the rounds. The reconciliation showed revision to the final linking endpoints for only a small number of outcomes. This small number of revisions suggests that the iterative method did not introduce risk of variation across rounds and therefore novice linkers should be encouraged to proceed to linking immediately post-training, where their linking is supported by a robust method, including study-specific linking rules, consensus meetings between rounds and a final reconciliation.

Recommendations

Given the results of this investigation we recommend that researchers:

- Supplement the recommended ICF linking preparations with study-specific preparatory training and customization of the ICF linking rules to achieve a high baseline level of inter-rater reliability,
- Use a sequential iterative linking method including regular consensus meetings until the point of linking saturation,
- Monitor linking saturation during the sequential iterative rounds, to ensure linking reliability is sufficiently high when the point of saturation occurs. Where reliability is not sufficiently high at the point of saturation, refinement to the customized ICF linking rules and further training to establish a shared understanding is recommended,
- Set inter-rater reliability and saturation hurdles for progression to linking the remaining outcomes. For example, our results suggest researchers could apply an a priori hurdle requirement of > 80% agreement between independent linkers, and a saturation hurdle of a just few newly linked ICF second-level categories in a round before progressing to linking all remaining outcomes.

Study limitations

The results of this pilot should be considered in light of a number of limitations.

While linking a large number of outcomes to the ICF engenders confidence in the agreement statistics that quantify inter-rater reliability, we are cognisant that the study reflects the experience of 2 novice ICF linkers with subject-matter expertise in a single discipline area. While ICF linking studies may be undertaken by pairs of novice linkers with subject-matter expertise (1, 38), we are mindful that further research is required with a larger sample of novice linkers to confirm that the results of this pilot investigation are generalizable. Should other investigators adopt the sequential iterative linking method used in this investigation, there would be opportunities to synthesize findings across studies in a form of meta-analysis to understand whether these findings are replicable across different discipline areas with independent pairs of novice linkers.

This pilot study achieved high baseline agreement between novice linkers. In this investigation, linkers did not participate in the ICF Research Branch face-to-face workshops as recommended (3, 4). The linkers did, however, complete other forms of recommended training (4), including the online ICF training modules (17) supplemented with targeted, study-specific training. The high baseline agreement between novice linkers suggests that the training undertaken was sufficient to achieve a high-level of agreement, and we encourage other researchers to conduct both the recommended ICF training, as well as study-specific training, where feasible (4).

The extraction of outcomes from systematic reviews was a variation on the approach commonly used to identify outcomes for the purpose of linking (4). Given this approach, it was necessary to determine the *likely meaningful concept* based on a review of the title and aim of each study. Identifying the *meaningful concept* is a prerequisite for any ICF linking (4), with the subsequent linking an interpretative endeavour. Given interpretations can vary based on settings (e.g. clinical or community), population groups (e.g. people with limb loss or people with spinal cord injuries) and between linkers (e.g. experts or novices) and the method used in this study, there are challenges in reliability identifying the *meaningful concept* prior to linking. Hence, the sequential iterative linking approach probably provides further benefit given it may minimize the variability in identification of the *meaningful concept* due to the regular conversations that are facilitated across linking rounds. The high level of agreement between linkers should engender confidence that this approach is sound; particularly when the customized ICF linking rules

were refined to prompt linkers to consider the *likely meaningful concept* before linking each outcome.

Whilst this study pilots a method that may contribute to improved inter-rater reliability amongst novice linkers, caution must be taken in the interpretation of the final linking outcome. The sequential iterative linking method probably addresses the issue of linking reliability and efficiency by novices, but it does not ensure linking accuracy. We encourage future researchers to investigate whether this approach improves the accuracy of linking completed by novices in comparison with experts and consider further method enhancements that will contribute to improved accuracy, thereby minimizing the barriers to this research for novices.

Finally, care must be taken when comparing the agreement statistics reported in this investigation against other studies given the different methods of linking, as well as the likely variations in linking experience and content expertise. Given the sequential iterative linking method used in this investigation, 1,297 outcomes were independently linked over 5 rounds, each with a separate consensus meeting. We suggest that the round 1 linking results are likely reflective of novice ICF linkers with content expertise (e.g. in prosthetics) immediately post-training. The round 5 linking results better reflect that of experienced ICF linkers with content expertise who have had the opportunity to build a shared understanding over sequential iterative linking rounds.

CONCLUSION

There is little prior research exploring innovative methods of linking to the ICF that may improve reliability, particularly among novice linkers. The innovative sequential iterative linking method trialled in this pilot study provides for a learning curve where novice linkers have an opportunity to develop a shared understanding and refine the linking rules over time. This method allowed novice linkers to improve the level of agreement over time and achieve similar inter-rater agreement to that observed in other studies involving expert linkers. Further research is required to engender confidence that the sequential iterative linking method can lead to strong inter-rater reliability among other novice linkers across discipline areas.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the expertise provided by: Ms Aloka Seneviratne of La Trobe University who developed and helped refine the custom data extraction and linking spreadsheet. Furthermore, we acknowledge the expert guidance and supervisory support provided by Professor Alan Shiell of La Trobe University.

LC gratefully acknowledges the support provided by the American Orthotic and Prosthetic Association (AOPA) Centre for Orthotic and Prosthetic Learning and Outcomes/Evidence-Based Practice (COPL) grant (EBP-053119).

The authors have no conflicts of interest to declare.

REFERENCES

1. Fayed N, Cieza A, Edmond Bickenbach J. Linking health and health-related information to the ICF: a systematic review of the literature from 2001 to 2008. Disabil Rehabil 2011; 33: 1941–1951.
2. Cieza A, Brockow T, Ewert T, Amman E, Kollerits B, Chatterji S, et al. Linking health-status measurements to the international classification of functioning, disability and health. J Rehabil Med 2002; 34: 205–210.
3. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustün B, Stucki G. ICF linking rules: an update based on lessons learned. J Rehabil Med 2005; 37: 212–218.
4. Cieza A, Fayed N, Bickenbach J, Prodinger B. Refinements of the ICF Linking Rules to strengthen their potential for establishing comparability of health information. Disabil Rehabil 2019; 41: 574–583.
5. Cieza A, Stucki G. Content comparison of health-related quality of life (HRQOL) instruments based on the International Classification of Functioning, Disability and Health (ICF). Qual Life Res 2005; 14: 1225–1237.
6. Soberg HL, Sandvik L, Ostensjo S. Reliability and applicability of the ICF in coding problems, resources and goals of persons with multiple injuries. Disabil Rehabil 2008; 30: 98–106.
7. Sigl T, Cieza A, Brockow T, Chatterji S, Kostanjsek N, Stucki G. Content comparison of low back pain-specific measures based on the International Classification of Functioning, Disability and Health (ICF). Clin J Pain 2006; 22: 147–153.
8. Stamm T, Geyh S, Cieza A, Machold K, Kollerits B, Kloppenburg M, et al. Measuring functioning in patients with hand osteoarthritis – content comparison of questionnaires based on the International Classification of Functioning, Disability and Health (ICF). Rheumatology (Oxford) 2006; 45: 1534–1541.
9. Castro S, Ferreira T, Dababnah S, Pinto AI. Linking autism measures with the ICF-CY: Functionality beyond the borders of diagnosis and interrater agreement issues. Dev Neurorehabil 2013; 16: 321–331.
10. Chen S, Tao J, Tao Q, Fang Y, Zhou X, Chen H, et al. Rater experience influences reliability and validity of the Brief International Classification of Functioning, Disability, and Health Core Set for Stroke. J Rehabil Med 2016; 48: 265–272.
11. Starrost K, Geyh S, Trautwein A, Grunow J, Ceballos-Baumann A, Prosiegel M, et al. Interrater reliability of the extended ICF core set for stroke applied by physical therapists. Phys Ther 2008; 88: 841–851.
12. Wu SM, Whiteside U, Neighbors C. Differences in inter-rater reliability and accuracy for a treatment adherence scale. Cogn Behav Ther 2007; 36: 230–239.
13. Clarke L, Ridgewell E, Dillon M. Measuring the benefits of prosthetic interventions in health economic evaluations. Part 1: identifying and linking outcomes to the ICF framework. Under peer review 2021.
14. Clarke L, Dillon M, Shiell A. Health economic evaluation in orthotics and prosthetics: a systematic review protocol. Syst Rev 2019; 8: 152.
15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and

meta-analyses: the PRISMA statement. PLoS Med 2009; 6: e1000097.
16. International Organization for Standardization (ISO). Assistive products for persons with disability – Classification and terminology (ISO 9999), ISO/TC 173/SC 2. Geneva, Switzerland: ISO; 2016.
17. World Health Organization. ICF e-Learning Tool. Geneva: WHO; 2020 [accessed 2020 June 12]. Available from: https://www.icf-elearning.com/
18. World Health Organization. Towards a common language for Functioning, Disability and Health ICF. Geneva: WHO; 2002.
19. World Health Organization. How to use the ICF: a practical manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for comment. Geneva: WHO; 2013.
20. Deathe AB, Wolfe DL, Devlin M, Hebert JS, Miller WC, Pallaveshi L. Selection of outcome measures in lower extremity amputation rehabilitation: ICF activities. Disabil Rehabil 2009; 31: 1455–1473.
21. Hebert JS, Wolfe DL, Miller WC, Deathe AB, Devlin M, Pallaveshi L. Outcome measures in amputation rehabilitation: ICF body functions. Disabil Rehabil 2009; 31: 1541–1554.
22. Radhakrishnan S, Kohler F, Gutenbrunner C, Jayaraman A, Li J, Pieber K, et al. The use of the International Classification of Functioning, Disability and Health to classify the factors influencing mobility reported by persons with an amputation: an international study. Prosthet Orthot Int 2017; 41: 412–419.
23. Radhakrishnan S, Kohler F, Gutenbrunner C, Jayaraman A, Pieber K, Li J, et al. Mobility in persons with lower extremity amputations and influencing factors: using the International Classification of Functioning, Disability and Health to quantify expert views. Prosthet Orthot Int 2019; 43: 88–94.
24. Theeven PJ, Hemmen B, Brink PR, Smeets RJ, Seelen HA. Measures and procedures utilized to determine the added value of microprocessor-controlled prosthetic knee joints: a systematic review. BMC Musculoskelet Disord 2013; 14: 1–12.
25. Xu J, Kohler F, Dickson H. Systematic review of concepts measured in individuals with lower limb amputation using the International Classification of Functioning, Disability and Health as a reference. Prosthet Orthot Int 2011; 35: 262–268.
26. World Health Organization. ICF Browser. Geneva: WHO; 2017. Available from: https://apps.who.int/classifications/icfbrowser/
27. World Health Organization. ICF: International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
28. Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Gaithersburg, MD: Advanced Analytics, LLC; 2014. Available from: https://www.agreestat.com/book4/9780970806284_prelim_chapter1.pdf.
29. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990; 43: 551–558.
30. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990; 43: 543–549.
31. Zec S, Soriani N, Comoretto R, Baldi I. High agreement and high prevalence: the paradox of Cohen's Kappa. Open Nurs J 2017; 11: 211–218.
32. R Foundation. irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC) 2019. Available from: https://CRAN.R-project.org/package=irrCAC
33. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–174.
34. Ogonowski J, Kronk R, Rice C, Feldman H. Inter-rater reliability in assigning ICF codes to children with disabilities. Disabil Rehabil 2004; 26: 353–361.
35. Kohler F, Connolly C, Sakaria A, Stendara K, Buhagiar M, Mojaddidi M. Can the ICF be used as a rehabilitation outcome measure? A study looking at the inter- and intra-rater reliability of ICF categories derived from an ADL assessment tool. J Rehabil Med 2013; 45: 881–887.
36. Hagberg K, Brånemark R, Hägg O. Questionnaire for Persons with a Transfemoral Amputation (Q-TFA): initial validity and reliability of a new outcome measure. J Rehabil Res Dev 2004; 41: 695–706.
37. Coenen M, Stamm TA, Stucki G, Cieza A. Individual interviews and focus groups in patients with rheumatoid arthritis: a comparison of two qualitative methods. Qual Life Res 2012; 21: 359–370.
38. Cerniauskaite M, Quintas R, Boldt C, Raggi A, Cieza A, Bickenbach JE, et al. Systematic literature review on ICF from 2001 to 2009: its use, implementation and operationalisation. Disabil Rehabil 2011; 33: 281–309.