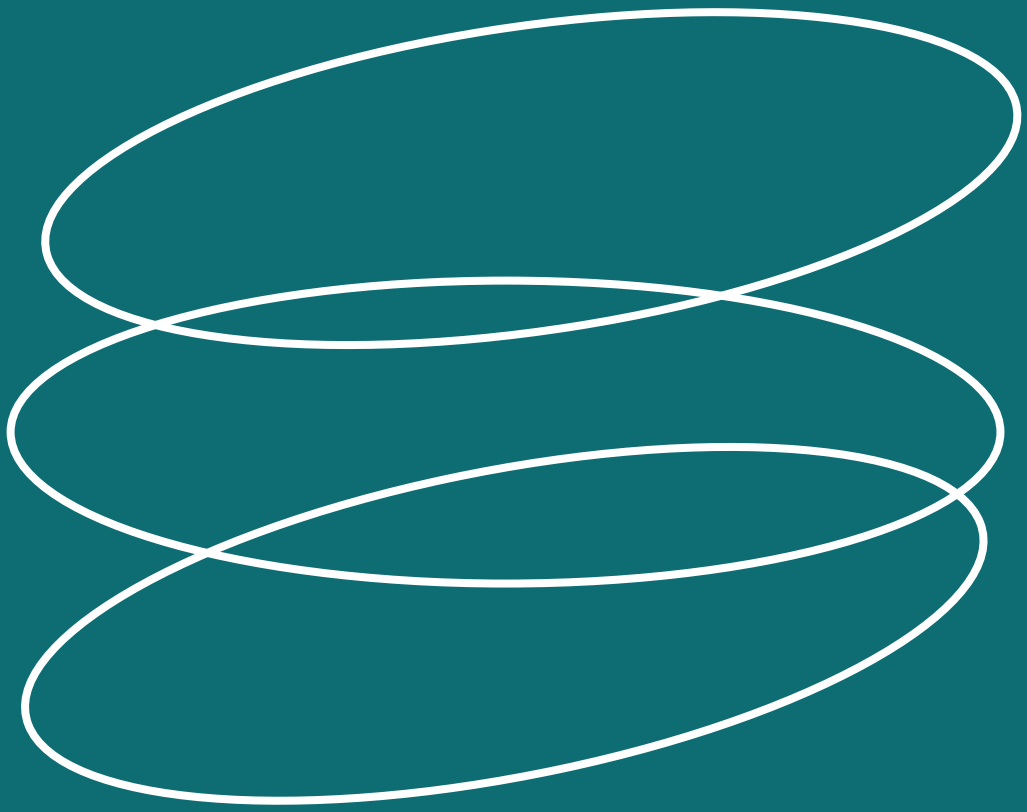
Strengthening Quality Improvement in Rehabilitation by Standardized Reporting of Functioning Outcomes



Cumulative PhD Thesis
Roxanne Maritz

University of Lucerne
2020

Strengthening Quality Improvement in Rehabilitation by Standardized Reporting of Functioning Outcomes

Cumulative Thesis
Doctor of Science, PhD in Health Sciences
Department of Health Sciences and Medicine
University of Lucerne

presented by
Roxanne Maritz

Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B, on behalf of the NRP74 StARS clinics. The Functional Independence Measure 18-item version can be reported as a unidimensional interval-scaled metric: Internal construct validity revisited. Journal of Rehabilitation Medicine. 2019; 51(3): 193-200.

Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B, on behalf of the NRP74 clinics. The Extended Barthel Index (EBI) can be reported as a unidimensional interval-scaled metric – A psychometric study. Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin. 2019; 29(4): 224-232.

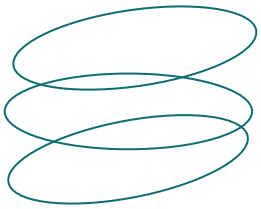
Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B, on behalf of the NRP74 clinics. Creating a common metric based on existing activities of daily living tools to enable standardized reporting of functioning outcomes achieved during rehabilitation. Journal of Rehabilitation Medicine. 2020; 52: jrm00085.

Maritz R, Ehrmann C, Prodinger B, Tennant A, Stucki G. The influence and added value of a Standardized Assessment and Reporting System for functioning outcomes upon national rehabilitation quality reports. International Journal for Quality in Health Care. 2020; 32(6): 379-387.

Accepted on 23.11.2020, on request by
Prof. Dr. Gerold Stucki, First Reviewer
Prof. Dr. Margaret Kelaher, Second Reviewer
Prof. Dr. Birgit Prodinger, Co-Promoter

Lucerne, Switzerland 2020

Table of contents



Acknowledgements

"If you want to go fast, go alone. If you want to go far, go together."
- African proverb

First, I would like to express my sincere gratitude to my first promoter Gerold Stucki and my co-promoter Birgit Prodinger for their continuous and valuable guidance, advice, encouragement and multifaceted expertise, which they generously shared with me. Working under their guidance has been a great privilege and I could not have imagined better mentors for my doctoral study.

I would also like to thank the rest of the NRP74 StARS project team including Jsabel Hodel, Cristina Ehrmann, Alan Tennant and Jerome Bickenbach for their much appreciated constructive feedback and support, for sharing their wealth of experience with me and for always having the bigger picture in mind.

Of course, my doctoral study and the NRP74 StARS project would not have been possible without countless others. Many thanks go to Swiss Paraplegic Research (SPF), particularly to Mirjam Brach, the University of Lucerne including the PhD Committee and the Swiss National Science Foundation's (SNSF) National Research Programme NRP74 Smarter Health Care for enabling me to conduct research related to my PhD. Furthermore, I would like to thank Melissa Selb and Carolina Fellinghauer from SPF, Anke Scheel-Sailer from the Swiss Paraplegic Centre and Luise Menzi from the National Association for Quality Development in Hospitals and Clinics (ANQ) for the much appreciated discussions, support on publications and their critical feedback and advice throughout the entire project. Many thanks also go to the NRP74 StARS advisory board and the 29 StARS clinics for the providing their ANQ datasets and for their insights from a political, practical and clinical perspective. Additionally I would like to thank Sarah Mantwill from the Swiss Learning Health System for her great support with the project's stakeholder involvement activities.

Moreover, I am sincerely grateful for having had the opportunity of a six-month PhD mobility research fellowship at the Evaluation and Implementation Science Group from the Centre for Health Policy at the University of Melbourne, Australia. Special thanks go to Margaret Kelaher, her team and again the SNSF, for making this fruitful scientific exchange possible.

Last but not least, I would like to thank my partner Pan, my parents, my sister and my dear friends for their precious support, encouragement and endless patience and my little daughter Wilma Malou for helping me put everything into perspective.

Summary

Background:
The primary goal of rehabilitation is to optimise functioning in everyday life. Therefore, functioning is a main quality indicator in the field of rehabilitation. Functioning is defined and classified by the World Health Organization's ICF the International Classification of Functioning, Disability and Health. In clinical practice, functioning outcomes are collected with a variety of ordinal-scaled assessment tools that are often only applicable within individual clinical encounters. Moreover, they cannot be aggregated in a valid manner or compared on a higher level. Therefore, there is a need for a Standardized Assessment and Reporting System (StARS) to ensure that functioning outcomes assessed with different tools can be aggregated, compared and taken as the basis for continuous learning and quality improvement at individual, institutional and national levels. This doctoral thesis takes Switzerland as an example, where musculoskeletal and neurological rehabilitation clinics can use different assessment tools, the Functional Independence Measure (FIM™) or the Extended Barthel Index (EBI), to report functioning outcomes for national rehabilitation quality reviews, what limits the comparability across clinics.

Objective:
To examine how an ICF-based and interval-scaled StARS for functioning outcomes can be created and implemented in Swiss rehabilitation quality reviews. Four specific aims were: 1) To examine whether the respective scores of the functioning assessment tools used in Swiss national quality reviews in neurological and musculoskeletal rehabilitation can be reported as unidimensional and interval-scaled metrics; 2) To create an ICF-based and interval-scaled common metric based on the functioning assessment tools used in Swiss national quality reviews in musculoskeletal and neurological rehabilitation; 3) To examine the influence and added value of an ICF-based and interval-scaled StARS on the current reporting of functioning outcomes in Swiss national rehabilitation quality reviews; 4) To develop strategies with relevant stakeholders for implementing the StARS in Swiss national quality reviews in rehabilitation.

Methods:
Four quantitative studies that addressed specific aims 1-3 and stakeholder involvement activities that addressed specific aim 4 were conducted. The basis of the four studies were datasets that included over 18'000 cases collected for the Swiss National Association for Quality Development (ANQ) in 2016 from 29 Swiss rehabilitation clinics providing musculoskeletal or neurological rehabilitation. In studies 1 and 2, the Rasch measurement model was applied in order to define whether FIM™ and EBI can be reported as unidimensional interval-scaled metrics. In study 3, an ICF-based interval-scaled common metric encompassing FIM™ and EBI was created as a core of a StARS by applying ICF Linking Rules to evaluate concept equivalence

and Rasch model requirements to establish score equivalence. In study 4, the impact and added value of the developed StARS for the reporting of functioning outcomes for Swiss national rehabilitation quality reviews was examined in comparison to the current ordinal-scaled system, including descriptive statistical methods and content exploration of further development areas of the current reviews based on relevant ICF Core Sets. The stakeholder involvement comprised a stakeholder brief and a stakeholder dialogue. The brief aimed to inform the stakeholders about the project, its findings and application areas and was developed with the involvement of stakeholders. The dialogue aimed to develop strategies for implementing a StARS for functioning outcomes in Switzerland.

Results:

Studies 1 and 2 showed that FIM™ and EBI can be reported as unidimensional and interval-scaled metrics for musculoskeletal and neurological rehabilitation, when Rasch-based transformation is applied. In study 3, concept and score equivalence of FIM™ and EBI could be established, resulting in an ICF-based interval-scaled common metric that encompasses the two assessment tools. In study 4, the comparison between the interval-scaled and the ordinal-scaled reporting showed that the achieved outcomes reported on an interval-scaled StARS tended to be smaller but more precise. Furthermore, study 4 demonstrated the added value of employing the ICF as the basis of a StARS, showing concrete functioning categories that can inform further development of national outcome quality reviews in rehabilitation. The output of the stakeholder brief and stakeholder dialogue was an implementation agenda in which the stakeholders decided on the next steps for implementing the developed StARS.

Conclusion:

The present thesis shows how a StARS for functioning outcomes can be created for national quality reviews in rehabilitation and that it can have a positive influence and added value in comparison to the currently applied ordinal-scaled reporting system. Moreover, the thesis illustrates the potential of involving stakeholders in developing implementation strategies to implement a StARS. An ICF-based and interval-scaled StARS enables valid aggregation and comparison of functioning outcomes assessed with different assessment tools resulting in richer data. This, in turn, enables learning from functioning outcomes and has the potential to inform and strengthen quality improvement in rehabilitation, in the sense of a learning health system.

List of abreviations

ADL	Activities of Daily Living
ANQ	Swiss National Association for Quality Development in Hospitals and Clinics
AROC	Australasian Rehabilitation Outcomes Centre
BI	Barthel Index
CIHI	Canadian Institute for Health Information
DIF	Differential Item Functioning
DRG	Diagnosis Related Groups
EBI	Extended Barthel Index
FIM™	Functional Independence Measure
ICF	International Classification for Functioning, Disability and Health
KVG	Swiss Health Insurance Act (Bundesgesetz über die Krankenversicherung)
LORY	Lucerne Open Repository
MSK	Musculoskeletal rehabilitation
NEUR	Neurological rehabilitation
NRP74	Swiss National Research Programme 74 Smarter Health Care
PhD	Doctor of Philosophy
SLHS	Swiss Learning Health System
SNSF	Swiss National Science Foundation
StARS	Standardized Assessment and Reporting System
SWOT	Strengths-Weaknesses-Opportunities-Threats
US	United States
WHO	World Health Organization

Chapter 1

Introduction

Introduction

This doctoral thesis demonstrates how quality improvement in rehabilitation can be strengthened by standardized reporting of functioning outcomes, based on the example of Switzerland. This first chapter introduces the rationale, context, key concepts and outline of the thesis that presents four studies and stakeholder involvement activities.

1.1 Rationale of the thesis

Functioning, the lived experience of health and its effect on daily life has been argued to be the third health indicator of the health system, complementing the established indicators of morbidity and mortality, and the key indicator for rehabilitation [1–3]. The concept of functioning is defined and classified in the World Health Organization's (WHO) International Classification of Functioning, Disability and Health (ICF) [3]. Functioning encompasses the outcomes of the interactions between a person's health condition, environment and personal factors, as well as body functions, body structures, activities and participation [3].

A problem currently faced in rehabilitation practice is that functioning information is collected with a wide variety of assessment tools, making it difficult to aggregate, compare and learn from functioning information [4]. There are two major challenges to comparing and aggregating functioning information assessed with different assessment tools:

1) The challenge of the content that differs between the tools: Even when two tools seem to cover similar concepts, such as Activities of Daily Living (ADL), the items of the assessment tools often differ in what they cover and their level of detail [5].
2) The challenge of the metric and scaling of the assessment tools: Different tools use different rating systems with differing total score ranges. Moreover, assessment tools assessing functioning information are often based on ordinal scale level [6]. In order to validly aggregate and compare functioning information, including the calculation of means or change scores, such as the difference between admission and discharge, interval scale level is needed [6–8].

Standardization of the assessment and reporting of functioning outcomes on the basis of the ICF would support the process of monitoring functioning information and integrating it in health information systems. Furthermore, it would enable the health care system to learn from functioning information and ultimately support the improvement of the quality of care [4, 9, 10]. A Standardized Assessment and Reporting System (StARS) for functioning information, including an ICF-based and interval-scaled common metric, can address both identified challenges and make functioning information assessed with different assessment tools comparable [4, 7, 11]. Using the ICF as a standard reference as the basis for the common metric allows for the comparison between the content of assessment tools, making it possible to evaluate whether

there is concept equivalence between the different tools [4, 11]. The common metric's interval-scale characteristic makes it possible to establish score equivalence of the encompassed assessment tools. Furthermore, it permits the meaningful aggregation and comparison of functioning information on a neutral common functioning reference metric, no matter with which tool the information was assessed [7].
Therefore, this doctoral thesis aims to examine how an ICF-based and interval-scaled StARS for functioning information can be created for the use in rehabilitation quality reports. Parallel to the development of a StARS, it was considered important to examine the influence and added value of a StARS in comparison to current ordinal-scaled reporting practice in rehabilitation quality reports and to involve relevant stakeholders in order to support future implementation [12].

1.2 Context of the thesis

Swiss public quality reviews in rehabilitation

This thesis takes the Swiss public quality reviews for musculoskeletal and neurological rehabilitation as a case in point. These reviews include the monitoring of patient functioning outcomes in inpatient rehabilitation in Switzerland through the coordination of measurement, analysis and public reporting of these outcomes on the level of the rehabilitation clinics, encompassing clinic comparisons [13]. The reviews and resulting reports are coordinated and published by the Swiss National Association for Quality Development in Hospitals and Clinics (ANQ). The ANQ was established in 2009, to develop and implement national quality and safety indicators in the field of acute care, mental health and rehabilitation [14]. The ANQ is the executive organ of the Swiss National Quality Agreement, which has been signed and is financed by all Swiss hospitals and clinics, insurance organizations and cantons, and in which the review of outcome quality is a key element. This form of nationwide partnership agreement is considered to be a pioneering method of providing national quality improvement across institutions in the health care sector [15]. The basis of the ANQ's activities is the Swiss Health Insurance Act (KVG), Article 48/9, which requires quality assurance across Swiss inpatient health care institutions, including periodic reviews of the outcome quality of health care services and clinic comparisons [16].
According to the ANQ review measurement plan, clinics from all areas of rehabilitation take measurements from all their patients at admission and discharge. To do so the clinics use different assessment tools depending on the rehabilitation area [13]. The annual ANQ reviews of musculoskeletal and neurological rehabilitation provide an interesting case in point for the development of a StARS for functioning outcomes, as the clinics can choose to report functioning outcomes with either the Functional Independence Measure (FIM™) [17] or the Extended Barthel Index (EBI) [18]. This practice reflects the reality that different clinics use different assessment tools for the measurement and monitoring of functioning outcomes. For the ANQ reviews, a clinic's outcomes need to be made comparable with other clinics across Switzerland, irrespective of which of the two tools are used to assess functioning outcomes. Both tools are considered to be very similar in content, both assessing the independence of a person in ADLs [19]. Furthermore, both tools are based on an ordinal scale, but are treated in ANQ reviews as if they are interval-scaled, i.e. means and change scores between admission and discharge are calculated, and the related scores are

treated as continuous variables in risk adjusted regression estimations [20–22].
The practice of not designating a single assessment tool to be used in the national quality reviews has the advantage that each clinic can choose whichever assessment tool that is more valuable for them. Both tools have their advantages and disadvantages, e.g. they differ in reliability and assessment time [19]. At the same time, this practice also impedes the comparison of functioning outcomes across clinics. To overcome the issue of comparability between clinics that use different assessment tools, the ANQ has commissioned the development of an expert-consensus-based transformation algorithm of the two assessment tools, referred to as the ANQ ADL score [23]. In comparison to a StARS for functioning outcomes that includes an ICF-based interval-scaled common metric, the ANQ ADL score has major shortcomings: its reliance on an ordinal-scale and the fact that it does not consider the different operational ranges of the assessment tools that it covers.
These special circumstances surrounding the Swiss ANQ outcome quality reviews in musculoskeletal and neurological rehabilitation together with the availability of rich and well-structured underlying data [20, 21] presented the opportunity to develop an example of an interval-scaled and ICF-based StARS for functioning outcomes to be used in national quality reviews.

Swiss National Research Programme Smarter Health Care NRP74
The thesis is a part of one of more than 30 research projects, included in the Swiss National Research Programme Smarter Health Care (NRP74) of the Swiss National Science Foundation (SNSF). The specific research project is entitled Enhancing continuous quality improvement and supported clinical decision making by standardized reporting of functioning, referred to as the NRP74 StARS project [24]. An overview of the NRP74 StARS research project organization can be found in Appendix 1.
A main goal of the NRP74 programme is to provide insights into ways of improving health outcomes with a particular focus on prevention and the treatment of chronic health conditions. Additionally, it aims at contributing to improved availability, accessibility, linkage and comparability of health data [25]. The NRP74 sets a special focus on knowledge translation and the collaboration with relevant stakeholders [26]. Given this, the NRP74 StARS project integrated stakeholder involvement activities throughout the span of the project, such as stakeholder consultation for grant submission, a kick-off meeting with project partner, the formation of an advisory board and two advisory board meetings, as well as two core activities related to the present thesis: a stakeholder brief and stakeholder dialogue.

1.3 Background and key concepts of the thesis

Health care quality and quality improvement
Quality and quality improvement have been and continue to be central issues in the field of health care [27]. Based on several definitions in the literature, WHO defined quality of care as “the extent to which health care services provided to individuals and patient populations to improve desired health outcomes” [28]. The work of pioneers such as Donabedian [29–31], Brook [32] and Rosenfeld [33], has been influential in advancing the topic, making major contributions to the definition, understanding and measurement of quality in health care [27]. Of special significance to this

doctoral thesis are the three attributes introduced by Donabedian for consideration when assessing quality in health care – namely structure, process and outcome [29]. This thesis focuses on the topic of outcome quality and related indicators. Outcome quality is defined as "change in a patient's current and future health status that can be attributed to antecedent health care" [29].

Measurement and monitoring are two vital aspects of quality improvement in health care. They provide the means for defining what health care institutions actually achieve and comparing this with targets in identify actions for improvement [34]. In order to be able to measure and monitor, quality has to be translated into measurable units, often referred to as quality indicators. Quality indicators are the tools for assessing health care structure, performance and outcomes attributes [35, 36]. To support health care facilities to measure and monitor quality, the standardization of such indicators is essential. The standardization of indicators also facilitates the comparison of quality between facilities [34]. Quality indicators also play a vital role in one method of quality monitoring that aims to connect measurement with subsequent quality improvement activities. This method compares health care providers through the quality indicators in performance or quality reports [37, 38]. The publicly reported ANQ reviews, highlighted in this thesis exemplify such reports on a national level [13, 15].

When quality indicators are publicly reported, the improvement of health care quality can be affected through two pathways according to Berwick et al.: the selection pathway that guides consumers to select high quality over low quality providers and the change pathway that aims to stimulate quality improvement among providers by identifying area of underperformance in comparison with others. Both pathways are interconnected by external motivation, that serves as the impetus for the commitment of organizations and providers to improve quality of care [37].

Rehabilitation

This thesis focuses on rehabilitation, the health strategy that comprises interventions to assist individuals experiencing disability to achieve and maintain optimal functioning in interaction with their environment [39]. Rehabilitation has been argued to be the health care strategy of the 21st century, as population ageing and higher incidence of chronic non-communicable diseases have contributed to rapid global increases in numbers of people experiencing problems in functioning [40, 41]. The importance of rehabilitation is emphasised by the WHO's Rehabilitation 2030 initiative, which highlights a substantial and increasing unmet need for rehabilitation worldwide. Many people with chronic health conditions need rehabilitation in order to stay or become as independent as possible, participate in education, be economically productive and fulfil meaningful roles in everyday life [42, 43].

Functioning and the standardized reporting of functioning outcomes

Functioning is a central topic of the present thesis. As functioning is one of the main indicators for a health system and the primary outcome of rehabilitation, it can be considered the key quality outcome indicator in rehabilitation [2, 44]. Functioning is the concept used by WHO to describe the lived experience of health in terms of body functions, body structures, activities and participation. The ICF provides the WHO

framework and standard reference to operationalize health as functioning in the form of a classification that includes chapters and codes [3]. The ICF comprises all domains of biological and lived health in terms of body structures (such as the structure of the brain), body functions (such as muscle power functions), activities and participation of an individual (such as walking or working) [3]. These functioning domains are the outcome of a dynamic interaction between health conditions (diseases, disorders, injuries, traumas, etc.) and contextual factors, including personal factors (such as age) and environmental factors (such as assistive devices), of which only the latter are classified in the ICF. All these domains and their interactions are reflected in the biopsychosocial model, also referred to as the ICF model (see Figure 1).

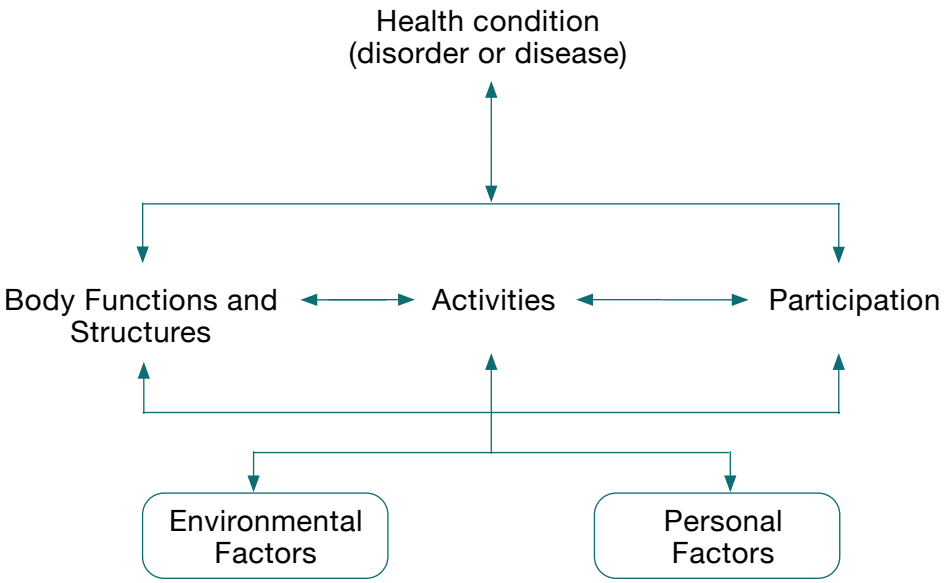


Figure 1: Interactions between the components of the ICF, World Health Organization 2001 [3]

According to the ICF, functioning denotes the positive aspects and disability the negative aspects of the outcome of the interactions between a person's health condition and contextual factors and encompasses body functions, body structures, activities and participation [3, 45]. Functioning information as indicated by the ICF is etiologically neutral. This means that it is associated with, but not casually linked to underlying health conditions. Individuals with the same health condition may experience different problems in their functioning, while individuals with different health conditions may experience the same problems in functioning [3, 45, 46]. In summary, functioning is a multidimensional and dynamic concept that addresses how health plays out in everyday life of a person, representing what matters to a person. It provides a picture of health that goes beyond morbidity and mortality [1, 2].

As a key quality outcome indicator in rehabilitation, the integration of functioning information in national health information systems, e.g. in outcome quality reports, is crucial for strengthening rehabilitation and quality of rehabilitation care [2, 47, 48]. In order to do so, as with quality indicators in general, it is important, that the reporting of functioning information is standardized [4, 7, 10, 34, 49], ideally through the use of

the internationally recognized ICF [48]. Functioning information can be assessed and reported in many ways, such as through observations documented as free text in electronic health records. In this thesis, the term "functioning outcomes" is used, when functioning information is assessed with established assessment tools [50]. These assessment tools are designed to assess functioning outcomes in a valid and reliable approach and are often referred to as standardized assessment tools. However, the wide variety of well-established assessment tools used to assess functioning outcomes [4] are not conductive to the standardization of the assessment and reporting of functioning outcomes in general. In order to achieve the general standardization of assessment and reporting of functioning outcomes, two options were identified. On one hand, specific assessment tools could be defined as a standard, for example at the national level. This option can be difficult to achieve, as there are many well-established assessment tools with clinical feasibility, carefully chosen by clinicians or providers to meet their specific assessment needs. On the other hand, different established assessment tools can be integrated in a StARS for functioning outcomes. This option has the advantage that providers and clinicians can continue to use the tools they are used to, while ensuring that the assessed outcomes are comparable [11].

Standardized reporting of functioning outcomes is also fundamental for a learning health system and quality improvement of the health system's response to people's functioning needs in the provision of rehabilitation [10]. Functioning outcomes can be used for learning, improvement and research on all three levels of a health system: micro, meso and macro [10, 51, 52]. The micro level is the clinical patient level; activates at this level include the individual rehabilitation planning, clinical decision making, goalsetting and the evaluation of the treatment [53]. The meso level is the health care institution level and involves the evaluation and optimisation of the service provision. The macro level is the level of rehabilitation policy that is guiding rehabilitation programming at the population level [10]. This thesis sets a focus at the meso level, with the StARS for national ANQ quality reviews that compares institutional functioning outcomes in rehabilitation. Nevertheless, it also touches upon the micro and macro levels, since functioning outcomes for the quality reports are assessed at the patient level and the StARS enables national data aggregation, which in turn, could be used for policy.

On the research side, much work has already been done highlighting the methods, the importance and need for standardized reporting of functioning through an ICF-based and interval-scaled StARS [4, 7, 11, 48, 49, 54]. The need for standardized reporting of functioning is also reflected in practice and measures have already been taken to make functioning outcomes comparable, such as the development of the ANQ ADL score in Switzerland that aims to compare functioning outcomes assessed with different ADL tools [23]. Unfortunately, the work of research and practice with regard to standardized reporting of functioning have yet to be brought together. For example, the ANQ ADL score is neither ICF-based nor interval-scaled. The present thesis aims to bridge the efforts of research and practice by developing a StARS for functioning outcomes based on existing research recommendations, and by developing respective implementation strategies for the practical context of Swiss national rehabilitation quality reviews.

Implementation and stakeholder involvement

As implementation is a key element of bridging research and practice, aspects of implementation with focus on the development of implementation strategies through stakeholder involvement, have been included in this thesis. Implementation is the process of putting to use or integrating evidence-based interventions in a setting [12]. One important part of the implementation process is knowledge translation, a dynamic process that includes synthesis, dissemination, exchange and application of knowledge, occurring within a complex system of interactions between researchers and knowledge users [55]. Implementation of research findings into practice is necessary in order to improve outcomes in health care [12]. Implementation and quality improvement in health care are therefore highly correlated.

Both implementation and quality improvement involve the united effort of stakeholders – health care professionals, patients and their families, researchers, payers, planers and educators – to bring health care towards better health [56]. Given its importance for implementation and quality improvement in health care, the interest in bridging the work of research and practice through stakeholder involvement has been growing [12].

As an extensive discussion of implementation would go beyond the scope of this thesis, it will draw attention to two implementation tools, whereby focusing on knowledge translation and stakeholder involvement, i.e. stakeholder dialogues and policy briefs (also referred to as stakeholder briefs) [57–59]. The application of both stakeholder involvement tools is recommended to help ensure the sustainable uptake of quality improvement across all three health system levels – micro, meso and macro- and in support of a learning health system [10].

1.4 Outline of the thesis

Given that there is a need for a StARS for functioning outcomes and that the Swiss context provides an interesting example for developing and implementing a StARS, the objective of this doctoral thesis was to examine how an ICF-based and interval-scaled StARS for functioning outcomes can be created and implemented in Swiss national quality reviews in rehabilitation. The following four specific aims were defined:

1) To examine whether the respective scores of the functioning assessment tools used in national quality reviews in neurological and musculoskeletal rehabilitation in Switzerland can be reported as unidimensional and interval-scaled metrics.
2) To create an ICF-based and interval-scaled common metric based on the functioning assessment tools used in Swiss national quality reviews in musculoskeletal and neurological rehabilitation.
3) To examine the influence and added value of the ICF-based and interval-scaled StARS on the current reporting of functioning outcomes in Swiss national rehabilitation quality reviews.
4) To develop strategies with relevant stakeholders for implementing the StARS in national quality reviews in Swiss rehabilitation.

In order to achieve the objective and specific aims 1-3, four related studies were conducted, resulting in peer-reviewed scientific publications (represented in Chapters 2-5). In order to achieve specific aim 4, a stakeholder dialogue was conducted, for which a stakeholder brief was developed as a preparatory document. These stakeholder involvement activities and their output were not prepared for scientific publication but described as part of this thesis in Chapter 6.

Data used in the thesis' studies

To ensure that the resulting StARS is based on actual data from the Swiss national quality reviews, the research team contacted all 64 clinics providing musculoskeletal or neurological rehabilitation in Switzerland in 2016 to acquire data for use in the studies. This is the official procedure when ANQ data is used for research [60]. Consequently, we received the ANQ data from 29 rehabilitation clinics representing for the German, French and Italian-speaking regions of Switzerland and that include more than 18'000 cases. Additionally, the data collected for the development of the ANQ ADL score (265 cases) was provided by the ANQ and the five involved clinics for use in the present thesis. For both, the overall doctoral project and the ANQ ADL score study an ethics approval was granted from the respective Swiss Ethics Committees.

Studies 1 and 2: Examination whether FIM™ and EBI can be reported as unidimensional and interval-scaled metrics

Studies 1 and 2 were related to specific aim 1, and had the objective to examine whether the functioning assessment tools FIM™ (study 1) and EBI (study 2) used for Swiss national quality reviews in neurological and musculoskeletal rehabilitation can be reported as unidimensional and interval-scaled metrics. There is evidence that the FIM™, an internationally well-established and widely researched assessment tool measures two different constructs i.e. a motor and a cognitive subscale, and should therefore not be reported in the total score [61]. No such research has yet been conducted about the EBI, a tool developed and used in the German-speaking context.
Studies 1 and 2 both used the same method: Rasch analysis [62–64] that employed so called testlet approaches to accommodate local response dependency between the items of the assessment tools [65]. Both studies were based on representative calibration samples of musculoskeletal and neurological FIM™ and EBI data from the 29 participating Swiss rehabilitation clinics. The two studies represent the foundation for the creation of the ICF-based interval-scaled StARS. Each study determined for each assessment tool independently whether it is valid to report a single total score for the respective tool, as it is currently reported by the ANQ [22]. Furthermore, each study determined if it is possible to transform the ordinal scales onto interval scale level.

Study 3: Creation of an ICF-based and interval-scaled common metric encompassing FIM™ and EBI

This study was related to specific aim 2, i.e. the development of an ICF-based and interval-scaled common metric, including FIM™ and EBI for the use in Swiss national rehabilitation quality reviews. The common metric represents the core of a StARS. The methodological approach of this study was based on the two key requirements

for standardized reporting of health information [4]: 1) content equivalence and 2) score equivalence. To determine whether the two assessment tools can be considered as equivalent in regards to their content ICF linking was applied [66]. To examine whether an interval-scaled reference metric including FIM™ and EBI can be established the Rasch model and its requirements for scale equating were applied [62, 67]. The study was based on a secondary analysis of the sample of the 265 neurological rehabilitation cases, used for the design ANQ ADL score. This sample of patients was assessed with both assessment tools, i.e. a common person design [67, 68]. This study provides a concrete example of how an ICF-based and interval-scaled common metric can be created with the ultimate goal of enabling the comparability and aggregation of the outcomes of different rehabilitation clinics, resulting from the use of different ADL assessment tools.

Study 4: Examination of the added value and the influence of the ICF-based and interval-scaled StARS upon the current reporting of functioning outcomes in Swiss national quality reviews

Study 4 was related to specific aim 3, i.e. to examine the added value and influence of a StARS upon the reporting of functioning outcomes in national rehabilitation quality reviews. In addition to the creation of the ICF-based interval-scaled StARS, we wanted to examine whether it makes a difference in the reported functioning outcomes if the StARS' common metric is applied compared to the current ordinal-scaled reporting approach of the Swiss national quality reviews, using the ANQ ADL score. Employing descriptive statistics, the methodological approach focused on the effect of the StARS' characteristics: the influence of the interval scale on the clinics' functioning outcomes [69, 70] and the added value of the ICF basis on the further development of the current ANQ reviews based on relevant ICF Core Sets. The study was based on 18047 complete musculoskeletal and neurological rehabilitation cases from the 29 participating Swiss rehabilitation clinics. It was conducted as a preliminary step to prepare and inform related decisions regarding to the implementation of the developed StARS.

Stakeholder involvement: Development of strategies with relevant stakeholders for implementing the StARS in Swiss national rehabilitation reviews

The approach of a stakeholder dialogue was chosen in order to reach specific aim 4, i.e. the development of implementation strategies for the developed StARS together with relevant stakeholders [57–59]. This also involved the development of a stakeholder brief, an adapted version of a policy brief, as a short preparatory document to provide all stakeholder dialogue participants information about the dialogue's content, background, goals, processes and related research findings in a user-friendly language The stakeholder dialogue aimed to inform relevant stakeholders about the research project and to develop strategies for implementing a StARS for functioning outcomes in Swiss national rehabilitation quality improvement. The 24 stakeholders who participated in the one-day stakeholder dialogue in November 2019 were identified and contacted with the help of the NRP74 StARS project's advisory board. The participating stakeholders represented the federal offices of public health and statistics, health care departments of cantons, patients, rehabilitation clinics and health professionals, quality management organizations, financing institutions, rehabilitation associations and research.

References Chapter 1

1. Stucki G, Bickenbach J. Health, Functioning, and Well-being: Individual and Societal. Archives of physical medicine and rehabilitation 2019; 100: 1788-92.
2. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. European journal of physical and rehabilitation medicine 2017; 53: 134-8.
3. World Health Organization. International Classification of Functioning, Disability and Health - ICF. Geneva, Switzerland: WHO, 2001.
4. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. European journal of physical and rehabilitation medicine 2018; 54: 110-7.
5. Maritz R, Baptiste S, Darzins SW, Magasi S, Weleschuk C, Prodinger B. Linking occupational therapy models and assessments to the ICF to enable standardized documentation of functioning. Canadian journal of occupational therapy 2018; 85: 330-41.
6. Andrich D. Rating scales and Rasch measurement. Expert review of pharmacoeconomics & outcomes research 2011; 11: 571-85.
7. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. European journal of physical and rehabilitation medicine 2017; 53: 144-9.
8. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? Journal of rehabilitation medicine 2012; 44: 97-8.
9. Prodinger B, Cieza A, Oberhauser C, Bickenbach J, Ustun TB, Chatterji S, Stucki G. Toward the International Classification of Functioning, Disability and Health (ICF) Rehabilitation Set: A Minimal Generic Set of Domains for Rehabilitation as a Health Strategy. Archives of physical medicine and rehabilitation 2016; 97: 875-84.
10. Stucki G, Bickenbach J. Functioning information in the learning health system. European journal of physical and rehabilitation medicine 2017; 53: 139-43.
11. Prodinger B, Tennant A, Stucki G, Cieza A, Ustun TB. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. Journal of health services research & policy 2016; 21: 223-8.
12. Brownson RC, Colditz GA, Proctor EK. Dissemination and Implementation Research in Health: Translating Science to Practice. New York NY, United States: Oxford University Press USA, 2012.
13. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Review Information: Rehabilitation [Internet], 2018. Bern, Switzerland: https://www.anq.ch/en/departments/rehabilitation/ Accessed [2020 July 31].
14. Vincent C, Staines A, Swiss Federal Office of Public Health FOPH. Enhancing the Quality and Safety of Swiss Healthcare - A national report commissioned by the Federal Office of Public Health on the quality and safety of healthcare in Switzerland [Internet], 2019. Bern, Switzerland:

    https://www.bag.admin.ch/bag/en/home/versicherungen/krankenversicherung/krankenversicherung-qualitaetssicherung.html Accessed [2020 July 31].
15. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Quality Agreement [Internet], 2020. Bern, Switzerland: https://www.anq.ch/en/quality-agreement/ Accessed [2020 July 31].
16. Swiss Federal Council, Swiss Federal Council. Bundesgesetz über die Krankenversicherung (KVG) [Internet] [German], 1994. Bern, Switzerland:

    https://www.admin.ch/opc/de/classified-compilation/19940073/index.html Accessed [2020 July 31].
17. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. Advances in clinical rehabilitation 1987; 1: 6-18.
18. Prosiegel M, Böttger S, Schenk T, König N, Marolf M, Vaney C, Garner C, Yassoridis A. The Extended Barthel Index - a new scale for the assessment of disability in neurological patients [German]. Neurologie & Rehabilitation 1996: 7-13.

19. Marolf M, Vaney C, König N, Schenk T, Prosiegel M. Evaluation of disability in multiple sclerosis patients: a comparative study of the Functional Independence Measure, the Extended Barthel Index and the Expanded Disability Status Scale. Clinical rehabilitation 1996; 10: 309-13.

20. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, ANQ. Nationaler Vergleichsbericht 2016 Muskuloskelettale Rehabilitation [Internet] [German], 2018. Bern, Switzerland: https://results.anq.ch/fileadmin/documents/anq/27_31/20180913_ANQ_Reha_Nationaler-Vergleichsbericht_Muskuloskelettale-Rehabilitation_2016.pdf Accessed [2020 July 31].

21. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, ANQ. Nationaler Vergleichsbericht 2016 Neurologische Rehabilitation [Internet] [German], 2018. Bern, Switzerland: https://results.anq.ch/fileadmin/documents/anq/27_31/20180913_ANQ_Reha_Nationaler-Vergleichsbericht_Neurologische-Rehabilitation_2016.pdf Accessed [2020 July 31].

22. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Auswertungskonzept Rehabilitation (Modul 2 & 3) [Internet] [German], 2018. Bern, Switzerland: https://www.anq.ch/wp-content/uploads/2017/12/ANQ_Module_23_Auswertungskonzept.pdf Accessed [2020 July 31].

23. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, ANQ. Entwicklung und Validierung eines ADL-Überführungsalgorithmus auf Basis von FIM® und EBI [Internet] [German], 2018. Bern, Switzerland:

    https://www.anq.ch/wp-content/uploads/2018/03/ANQ_Reha_Abschlussbericht_ADL_Algorithmen.pdf Accessed [2020 July 31].

24. Swiss National Science Foundation, NRP74 - SNSF. Project 21: Standardised Assessment and Reporting System of functioning of people with chronic diseases (StARS) [Internet], 2017. Bern, Switzerland: http://www.nfp74.ch/en/projects/healthcare-across-sectors/project-prodinger Accessed [2020 July 31].

25. Swiss National Science Foundation, NRP74 - SNSF. The National Research Prgramme Smarter Health Care - NRP74 [Internet], 2017. Bern, Switzerland: http://www.nfp74.ch/en/the-nrp Accessed [2020 July 31].

26. Swiss National Science Foundation, NRP74 - SNSF. NRP74 Organisation [Internet], 2017. Bern, Switzerland: http://www.nfp74.ch/en/the-nrp/organisation Accessed [2020 July 31].

27. Johnson JK, Sollecito WA. The Global Evolution of Continuous Quality Improvement: From Japanese Manufacturing to Global Health Services. In: Johnson JK, Sollecito WA, editors. McLaughlin & Kaluzny's Continuous Quality Improvement in Health Care. Burlington MA, United States: Jones & Bartlett Learning; 2018. p. 1-28.

28. World Health Organization, WHO. What is quality of care and why is it important? [Internet], 2016. Geneva, Switzerland: https://www.who.int/maternal_child_adolescent/topics/quality-of-care/definition/en/ Accessed [2020 July 31].

29. Donabedian A. Volume I The Definition of Quality and Approaches to its Assessment. Ann Arbor, United States: Health Administration Press, 1980.

30. Donabedian A. Volume II The Criteria and Standards of Quality. Ann Arbor, United States: Health Administration Press, 1982.

31. Donabedian A. Volume III The Methods and Findings of Quality Assessment and Monitoring: An Illustrated Analysis. Ann Arbor, United States: Health Administration Press, 1985.

32. Brook RH, McGlynn EA, Cleary PD. Measuring Quality of Care. New England journal of medicine 1996; 335: 966-70.

33. Rosenfeld LS. Quality of Medical Care in Hospitals. American journal of public health and the nations health 1957; 47: 856-65.

34. Shaw C, WHO Regional Office for Europe (Health Evidence Network). How can hospital performance be measured and monitored? [Internet], 2003. Copenhagen, Denmark:

    http://www.euro.who.int/document/e82975.pdf Accessed [2020 Jul 31].

35. Jesus TS, Hoenig H. Postacute rehabilitation quality of care: toward a shared conceptual framework. Archives of physical medicine and rehabilitation 2015; 96: 960-9.

36. Camp PG, Cheung W. Are We Delivering Optimal Pulmonary Rehabilitation? The Importance of Quality Indicators in Evaluating Clinical Practice. Physical therapy 2018; 98: 541-8.

37. Berwick DM, James B, Coye MJ. Connections between quality measurement and improvement. Medical care 2003; 41: I30-8.

38. Berwick DM, Kelley E, Kruk ME, Nishtar S, Pate MA. Three global health-care quality reports in 2018. The Lancet 2018; 392: 194-5.

39. European Physical and Rehabilitation Medicine Bodies Alliance. White Book on Physical and Rehabilitation Medicine (PRM) in Europe. Chapter 1. Definitions and concepts of PRM. European journal of physical and rehabilitation medicine 2018; 54: 156-65.

40. Cieza A. Rehabilitation the Health Strategy of the 21st Century, Really? Archives of physical medicine and rehabilitation 2019; 100: 2212-4.

41. Stucki G, Bickenbach J, Gutenbrunner C, Melvin J. Rehabilitation: The health strategy of the 21st century. Journal of rehabilitation medicine 2018; 50: 309-16.

42. World Health Organization, WHO. Health information systems and rehabilitation. Rehabilitation 2030: A Call for Action [Internet], 2017. Geneva, Switzerland: https://www.who.int/disabilities/care/rehab-2030/en/ Accessed [2020 July 31].

43. World Health Organization, WHO. Rehabilitation in health systems: guide for action [Internet], 2019. Geneva, Switzerland: https://www.who.int/rehabilitation/rehabilitation-guide-for-action/en/ Accessed [2020 July 31].

44. Stucki G, Cieza A, Melvin J. The International Classification of Functioning, Disability and Health (ICF): a unifying model for the conceptual description of the rehabilitation strategy. Journal of rehabilitation medicine 2007; 39: 279-85.

45. World Health Organization, WHO. How to use the ICF - A Practical Manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for commen [Internet], 2013. Geneva, Switzerland: http://www.who.int/classifications/drafticfpracticalmanual2.pdf Accessed [2020 July 31].

46. World Health Organization, WHO. Towards a Common Language for Functioning, Disability and Health – ICF [Internet], 2002. Geneva, Switzerland:

https://www.who.int/classifications/icf/icfbeginnersguide.pdf?ua=1 Accessed [2020 July 31].

47. Stucki G, Bickenbach J, Melvin J. Strengthening Rehabilitation in Health Systems Worldwide by Integrating Information on Functioning in National Health Information Systems. American journal of physical medicine & rehabilitation 2017; 96: 677-81.

48. Dorjbal D, Cieza A, Gmunder HP, Scheel-Sailer A, Stucki G, Ustun TB, Prodinger B. Strengthening quality of care through standardized reporting based on the World Health Organization's reference classifications. International journal for quality in health care 2016; 28: 626-33.

49. Stucki G, Pollock A, Engkasan JP, Selb M. How to use the International Classification of Functioning, Disability and Health as a reference system for comparative evaluation and standardized reporting of rehabilitation interventions. European journal of physical and rehabilitation medicine 2019; 55: 384-94.

50. Sivan M, Phillips M, Baguley I, Nott M. Concepts of Rehabilitation - Outcome Measurment. In: Sivan M, Phillips M, Baguley I, Nott M, editors. Oxford Handbook of Rehabilitation Medicine. Oxford, United Kingdom: Oxford University Press; 2019.

51. Solvang PK, Hanisch H, Reinhardt JD. The rehabilitation research matrix: producing knowledge at micro, meso, and macro levels. Disability and rehabilitation 2017; 39: 1983-9.

52. Stucki G, Zampolini M, Juocevicius A, Negrini S, Christodoulou N. Practice, science and governance in interaction: European effort for the system-wide implementation of the International Classification of Functioning, Disability and Health (ICF) in Physical and Rehabilitation Medicine. European journal of physical and rehabilitation medicine 2017; 53: 299-307.

53. Rauch A, Cieza A, Stucki G. How to apply the International Classification of Functioning, Disability and Health (ICF) for rehabilitation management in clinical practice. European journal of physical and rehabilitation medicine 2008; 44: 329-42.

54. Prodinger B, Stamm T, Peterson D, Stucki G, Tennant A. Toward a Standardized Reporting of Outcomes in Hand Osteoarthritis: Developing a Common Metric of Outcome Measures Commonly Used to Assess Functioning. Arthritis care & research 2016; 68: 1115-27.

55. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, Robinson N. Lost in knowledge translation: time for a map? The Journal of continuing education in the health professions 2006; 26: 13-24.

56. Batalden PB, Davidoff F. What is "quality improvement" and how can it transform healthcare? Quality & safety in health care 2007; 16: 2-3.

57. Lavis JN, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP). Health Research Policy and Systems 2009; 7: I1.

58. Lavis JN, Boyko JA, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 14: Organising and using policy dialogues to support evidence-informed policymaking. Health Research Policy and Systems 2009; 7: S14.

59. Rubinelli S, von Groote PM. Stakeholder Dialogue as Deliberation for Decision Making in Health Policy and Systems: The Approach from Argumentation Theory. American journal of physical medicine & rehabilitation 2017; 96: S17-S22.

60. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, ANQ. Empfehlungen Verwendung von ANQ-Daten zu Forschungszwecken [Internet] [German], 2019. Bern, Switzerland: https://www.anq.ch/wp-content/uploads/2017/12/ANQ_Konzept_Nutzung_Forschungsdaten.pdf Accessed [2020 July 31].

61. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. Journal of rehabilitation medicine 2011; 43: 884-91.

62. Ehlan A, Kucukdeveci A, Tennant A. The Rasch Measurement Model. In: Franchignoni F, editor. Research Issues in Physical & Rehabilitation Medicine. Pavia, Italy: Salvatore Maugeri Foundation; 2010. p. 89–102.

63. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47.

64. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis and rheumatism 2007; 57: 1358-62.

65. Wilson M. Detecting and interpreting local item dependence using a family of Rasch models. Applied psychological measurment 1988; 2: 353-64.

66. Cieza A, Fayed N, Bickenbach J, Prodinger B. Refinements of the ICF Linking Rules to strengthen their potential for establishing comparability of health information. Disability and rehabilitation 2019; 41: 574-83.

67. Andrich D. The Polytomous Rasch Model and the Equating of Two Instruments. In: Christensen K, Kreiner S, Mesbah M, editors. Rasch Models in Health. Hoboken, United States: ISTE Ltd and John Wiley and Sons, Icn.; 2013. p. 163-96.

68. Kolen MJ, Brennan RL. Introduction and Concepts. In: Kolen MJ, Brennan RL, editors. Test equating, scaling, and linking. New York, United States: Springer; 2004. p. 1-28.

69. Verburg IW, Holman R, Peek N, Abu-Hanna A, de Keizer NF. Guidelines on constructing funnel plots for quality indicators: A case study on mortality in intensive care unit patients. Stat Methods Med Res 2018; 27: 3350-66.

70. Kuhrij L, van Zwet E, van den Berg-Vos R, Nederkoorn P, Marang-van de Mheen PJ. Enhancing feedback on performance measures: the difference in outlier detection using a binary versus continuous outcome funnel plot and implications for quality improvement. BMJ quality & safety 2020; Epub ahead of print.

Chapter 2

The Functional Independence Measure 18-item version can be reported as a unidimensional interval-scaled metric: Internal construct validity revisited

Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B, on behalf of the NRP74 StARS clinics. The Functional Independence Measure 18-item version can be reported as a unidimensional interval-scaled metric: Internal construct validity revisited. Journal of Rehabilitation Medicine. 2019; 51(3): 193-200.

This is an article published by the Foundation of Rehabilitation Information in the Journal of Rehabilitation Medicine on the 5th of March 2019, available online: https://doi.org/10.2340/16501977-2525

J Rehabil Med 2019; 51: 193–200

ORIGINAL REPORT

THE FUNCTIONAL INDEPENDENCE MEASURE 18-ITEM VERSION CAN BE REPORTED AS A UNIDIMENSIONAL INTERVAL-SCALED METRIC: INTERNAL CONSTRUCT VALIDITY REVISITED

Roxanne MARITZ, MA[1,2], Alan TENNANT, PhD[1,2], Carolina FELLINGHAUER, PhD[1], Gerold STUCKI, MD[1,2] and Birgit PRODINGER, PhD[1,2,3], on behalf of the NRP74 StARS clinics[4]

From the [1]Swiss Paraplegic Research, Nottwil, Switzerland, [2]Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland, [3]Faculty of Applied Health and Social Sciences, Technical University of Applied Sciences Rosenheim, Rosenheim, Germany and [4]NRP74 StARS clinics (aarReha Schinznach – Stefan Bützberger, Berner Klinik Montana – Dieter Ernst, Berner Reha Zentrum Heiligenschwendi – Jürg Wägli, Clinica di Riabilitazione EOC Novaggio & Faido – Giovanni Rabito, Clinica Hildebrand – Sandro Foiada, Clinique La Lignière – Nicolas Walther, Felix Platter Spital – Stefan Engelter, Hôpital du Valais Spital Wallis Centre Martigny, Sierre, Brig & Saint Amé – Els De Waele, Kantonsspital Baselland Bruderholz & Laufen – Beat Ritter, Klinik Schloss Mammern – Ruth Fleisch-Silvestri, Kliniken Valens Rehazentrum Valens, Rehazentrum Walenstadtberg & Rheinburg-Klinik – Stefan Bachmann, Rehaklinik Dussnang – Javier Blanco, Rehab Basel – Margret Hund – Georgiadis, Zürcher RehaZentrum Wald – Matthias Hermann, Rehaklinik Zihlschlacht – Michèle Bongetta, Spitäler Schaffhausen – Jan Kuchynka, Universitätsklinik Balgrist – Serge Altmann)

Objective: **Since the 1990s the Functional Independence Measure (FIM™) was believed to measure 2 different constructs, represented by its motor and cognitive subscales. The practice of reporting FIM™ total scores, together with recent developments in the understanding of the influence of locally dependent items on fit to the Rasch model, raises the question of whether the FIM™ 18-item version can be reported as a unidimensional interval-scaled metric.**

Design: **Rasch analysis of the FIM™ using testlet approaches to accommodate local response dependency.**

Patients: **A calibration sample containing 946 cases of data from 11,103 patients undergoing neurological or musculoskeletal rehabilitation in Switzerland in 2016.**

Results: **Baseline analysis and the traditional testlet approach showed no fit with the Rasch model. When items were grouped into 2 testlets, fit to the Rasch model was achieved, indicating unidimensionality across all 18 items. A transformation table to convert FIM™ raw ordinal scores to the corresponding Rasch interval scaled values was created.**

Conclusion: **This study provides evidence that FIM™ total scores represent a unidimensional set of items, supporting their use in clinical practice and outcome reporting when applying the respective transformation table. This provides a basis for standardized reporting of functioning.**

Key words: outcome assessment (healthcare); psychometrics; rehabilitation; activities of daily living; Rasch measurement model; Functional Independence Measure.

Accepted Jan 24, 2019; Epub ahead of print Mar 5, 2019

J Rehabil Med 2019; 51: 193–200

Correspondence address: Roxanne Maritz, Rehabilitation Services & Care Unit, Swiss Paraplegic Research, 6207 Nottwil, Switzerland. E-mail: roxanne.maritz@paraplegie.ch

LAY ABSTRACT

The aim of this study was to look in detail at the FIM™, an assessment tool often used for patients undergoing rehabilitation. Some users report the FIM™ as 2 scores: one related to motor tasks, the other to cognitive tasks; others recommend reporting it as a single score including both motor and cognitive tasks. This study explored whether it is statistically meaningful to sum all the points into a single FIM™ total score. The results support the current practice of summing the points into a single total score for patients undergoing musculoskeletal and neurological rehabilitation. The results also allowed an interval scale to be derived from the FIM™, enabling a broad range of calculations to be made using the FIM™ score, such as calculating the change in FIM™ outcomes from the time a patient is admitted to a rehabilitation clinic until their discharge.

The primary outcome of rehabilitation is functioning (1). In order to document and monitor functioning, existing data collection tools can be used (2). The Functional Independence Measure (FIM™) is an assessment tool that is widely used in rehabilitation worldwide (3–6). The FIM™ is applied at the patient level to measure change throughout rehabilitation, at the institutional level to measure outcome quality, and at the national level for performance reporting or quality monitoring. Examples of use of the FIM™ are the reports of the Swiss National Association for Quality Development in Inpatient Care (ANQ) (7), the US model system for spinal cord injury (8) and traumatic brain injury (9), the Canadian Institute for Health Information (CIHI) (10), and the Australasian Rehabilitation Outcome Centre (AROC) (11). Furthermore, FIM™ measures can serve as a basis for inpatient rehabilitation payment (12). As with any assessment tool, in order to report valid total scores, certain psychometric standards must be met, including the assumption of unidimensionality. Furthermore, if an assessment tool is used to calculate change scores, it must be interval-scaled, rather than ordinal-scaled (6, 13). The Rasch measurement model can be used to examine assumptions such as unidimensionality or local item dependencies. Where satisfactory fit of data

doi: 10.2340/16501977-2525

to the model is achieved, an interval-scaled metric can be derived from ordinal scales (14, 15).

Earlier analysis of the FIM™ using Rasch analysis in the 1990s indicated that the FIM™ 18-item version incorporates 2 different constructs, represented by a motor scale and a cognitive scale, each of which should be scored separately (16). However, in clinical practice both the reporting of 2 separate motor and cognitive total scores and the reporting of a single total score of the FIM™, is evident (7, 9, 11). Since this first Rasch analysis of the FIM™, many others have been published, mostly on its motor subscale (17), but also on adaptations of the FIM™ (18, 19). More recently, the issue of so-called local item dependency has received attention (20). Local item dependency occurs when instrument items remain correlated when conditioned on the trait, what is functional independence in the case of the FIM™. Local dependency is indicated by significant correlation of the standardized analysis residuals. Fit of the FIM™ motor scale to the Rasch model has been shown to be seriously affected by local item dependency, which, once accommodated, resulted in adequate model fit (17).

Thus, given the recent methodological developments with regards to addressing the issue of local dependency in health scales, and inconsistency in reporting the FIM™ in practice, a review of the FIM™ 18-item version seemed appropriate, in order to address the following question: Is it possible to add all FIM™ items together to obtain a valid unidimensional total score, taking into account the local dependency in its item set? The objective of this study was therefore to revisit the question of whether the FIM™ can be reported as a unidimensional interval-scaled metric when local dependency is taken into account. Two specific aims in relation to the study's objective were: (*i*) to explore the metric properties of the FIM™; and (*ii*) to determine whether an interval-scale scoring system of the FIM™ 18-item version can be made available and, if so, to create an interval-scale transformation of the FIM™ raw scores when administered in the context of national quality monitoring in neurological and musculoskeletal rehabilitation.

METHODS

Subjects and setting

Data collected routinely for the Swiss national quality reporting, coordinated by the ANQ, was used for secondary analysis. All 64 Swiss rehabilitation clinics that provided data to the ANQ in 2016 for musculoskeletal or neurological rehabilitation were contacted, of which 30 voluntarily agreed to provide their ANQ datasets. Since the clinics can choose between different assessment tools in ANQ data collection, not all datasets contained FIM™ data. Thus, this study used datasets from 23 rehabilitation clinics, with 11,103 complete cases in total, representative of 3 different Swiss language regions (German, French, Italian). The FIM™ was administered at admission and discharge. Ethics approval for the study was requested from the Swiss Ethics Commissions, which stated in a declaration of no objection that the project fulfils the general ethical and scientific standards for research with humans and poses no health hazards.

Functional Independence Measure

The FIM™ is an assessment tool comprising 18 items. Thirteen items belong to the motor subscale and 5 items belong to the cognitive subscale. All items are scored from 1 (total assistance) to 7 (complete independence). The FIM™ item scores are summed up to a total score, ranging between 18 and 126, or total motor score ranging between 13 and 91 and between 5 and 35 for the cognitive total score (4). The ANQ used German, French and Italian translations of the FIM™ based on its official English version, on which a translation agreement was made with the Uniform Data System for Medical Rehabilitation (UDSMR). As this is common practice, the translations have not been authenticated by the UDSMR. In order to qualify to administer the FIM™ , the health professionals received training provided by the ANQ according to the respective UDSMR policy.

Sampling

A random stratified calibration sample was created using R (21), since type I errors, i.e. rejecting a hypothesis even if it was true, are likely to appear with a large sample size in Rasch analysis (22). The aim was to create a sample of approximately 1,000 cases, representing 4 equally sized subsamples, each with sufficient sample size for a stable item calibration and statistical interpretation (23, 24). Each subsample focused on one of the 2 different time-points of measurement, and one of the 2 different health condition groups of musculoskeletal and neurological rehabilitation: musculoskeletal cases at admission (MSKt1), musculoskeletal cases at discharge (MSKt2), neurological cases at admission (NEURt1) and neurological cases at discharge (NEURt2). To obtain precision across the whole range of scores (total score range 108; 18–126) and representation of language regions, a random sample was taken from each available total score per subsample and language region group. Cases that were selected from the admission subsamples were excluded and not selected for the discharge subsamples (25). Prior to the random selection all cases with missing values in a person's contextual factors of interest (described in more detail below) and all cases that scored an extreme score (18 or 126), were deleted, since they are excluded from the calculation of item difficulties by the Rasch measurement model. The sampling strategy is shown in Fig. 1.

Data analysis

To summarize basic sample characteristics and response distributions of the FIM™, descriptive statistics were conducted with Stata Version 14.2 (26). In order to achieve the study's first specific aim Rasch analysis was conducted using RUMM2030 (27). The analytical focus gave reference to local response dependency represented by residual correlations. High residual correlations indicate that items are measuring the same thing too closely (13). Furthermore, threshold disordering was examined, which indicates that the different response categories of an item are not in a successive order, i.e. do not represent an increasing level of functional independence. In addition, differential item

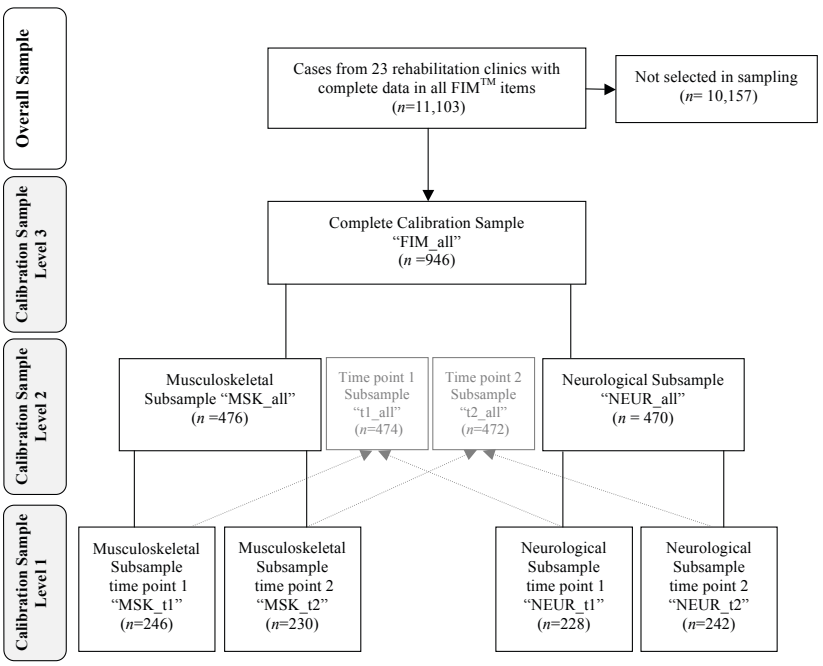


Fig. 1. Flow chart calibration sample with 3 different aggregation levels. FIM: Functional Independence Measure.

functioning (DIF) was evaluated, which indicates that, while accounting for the trait, an item works differently for certain groups defined by a contextual factor, such as gender or age. The partial credit model was applied, which has been shown previously to be the appropriate parametrization for the FIM™ (17, 28).

Baseline analyses

The baseline analysis tested how well the observed data from all 18 items fit the Rasch model (15). To do so, the individual and overall item-fit, the person-fit, the reliability indices α and person separation index (PSI), and the χ^2 *p*-value of the item-trait interaction standing for the fit of the data to the Rasch model were ascertained. The respective acceptable levels are represented in the bottom line of the corresponding results table. In addition, local response dependency among items was scrutinized, along with threshold disordering of item categories, and DIF for the following 7 factors: gender, age (4 age groups according to the interquartile ranges), nationality (Swiss or other), insurance (general, semi-private, private), rehabilitation group (neurological or musculoskeletal rehabilitation), clinic language (German, French or Italian) and time-point of measurement (admission t1, discharge t2). Both individual item-fit and DIF analyses *p*-values are Bonferroni adjusted in the RUMM2030 software.

Testlet approaches

Where the local independence assumption of the Rasch model was not met, testlet approaches were applied. A testlet is a simple sum score from a set of associated items, making the set into a single new "super"-item in order to absorb their dependencies (20, 29–31). The creation of testlets revealed positive results in earlier Rasch analyses of the FIM™ motor scale (17). Two different testlet approaches were used: one approach, referred to as traditional testlet approach, creating testlets oriented at conceptually associated items and based on their residual correlations (32). By grouping similar items into super-items, such as, for example, all the transfer items of the FIM™ , this traditional testlet approach highlights the potential differences, e.g. dimensionality between testlets unifies similar items, such as "self-care"' or "transfer". The other approach, referred to as alternative 2-testlet approach, divides conceptually similar items into 2 distinct testlets of equal size, taking alternative items in each testlet. This approach focuses on the total score of the FIM™ rather than the single items or groups of items by emphasizing the similarity of the items, as together they should measure the concept of functional independence. In delivering a bi-factor equivalent approach, the alternative 2-testlet approach has the advantage of creating testlets of equal size, as recommended by Andrich (29). Another advantage of the 2-testlet approach is that it allows for a conditional test of fit. Furthermore, all testlet-based approaches allow the calculation of the "explained common variance"' attributable to the general "first factor", indicating the proportion of variance retained to create a unidimensional latent estimate (29). Acceptable values of these additional statistics are indicated at the bottom of the respective testlet result table. The analysis of threshold disordering is not meaningful at the level of testlets, as a particular score can be derived in a number of ways, and is therefore not reported.

To ensure robustness of the results, the baseline analysis and the best-fitting testlet approach was conducted at 3 levels of aggregation of the calibration sample (see Fig. 1). In

Level 1 all 4 subsamples were analysed separately (MSKt1, MSKt2, NEURt1 and NEURt2). In Level 2 the rehabilitation group and time-point subsamples were aggregated respectively (MSKt1&t2, NEURt1&t2, t1MSK&NEUR, t2MSK&NEUR). Level 3 represents the aggregation of all 4 subsamples, i.e. the entire calibration sample (FIM_all). Together, these 3 aggregation levels resulted in 9 analysis steps.

For both testlet approaches, the emphasis is on making existing assessment tools work without the need to delete items or change the scoring structure.

Differential Item Functioning strategy

DIF was analysed in situations in which local dependencies could be accommodated satisfactorily with testlets. Where a lack of group invariance was observed, the testlets for the contextual factor were split on the basis of the strongest DIF, and continued until no further DIF was present (33). The split and unsplit solutions were then compared with each other on the basis of the Rasch person estimates, anchored to each other with an unsplit item free of DIF. An effect size calculation, based on the mean of the person estimates, their standard deviations, and the correlation of the split and unsplit version (34) was applied to determine whether DIF split was necessary for the final transformation table. If the effect size was below 0.2, DIF was considered small (35) and no action was taken to adjust for DIF.

Transformation table

The second specific aim of this study was to develop a transformation table in case fit to the Rasch model could be achieved. The solution with the best fit to the Rasch model was taken as a basis for this transformation, i.e. the solution with the most satisfactory core values for the entire calibration sample. The transformation table from FIM™ raw ordinal total scores to the corresponding interval-scaled values was based on the respective estimates according to the Rasch model.

RESULTS

Sample characteristics

The calibration sample included 946 cases. Of these, 476 were musculoskeletal cases and 470 neurological cases. A total of 474 cases were from time-point 1 admission, and 472 from time-point 2 discharge (see Fig. 1). FIM™ total scores had a mean of 81.7 (standard deviation (SD) = 27.5, median = 84). The mean age of subjects in the calibration sample was 71.6 years (SD = 14.5, 20–102 years). The calibration sample was 43% (n = 403) male and 57% (n = 543) female; 41% (n = 392) were from the German-speaking region of Switzerland, 25% (n = 238) from the French-speaking region and 34% (n = 316) from the Italian-speaking region; 84% (n = 798) of the sample were Swiss and 16% (n = 148) had another nationality. Insurance status was: 67% (n = 633) general, 18% (n = 172) semi-private, and 15% (n = 141) private.

Baseline Rasch analysis

In the 9 baseline analysis steps across the 3 aggregation levels of the calibration sample, no fit to the Rasch model was achieved (Table I). In all analyses the p-values of the item-trait χ^2 were significant. Furthermore, in all analysis steps there were items that showed local dependencies among each other, DIF and threshold disordering. Information on threshold disordering and local dependency of the baseline analyses are shown in Appendix S1[1].

[1]http://www.medicaljournals.se/jrm/content/?doi=10.2340/16501977-2525

Table I. Functional Independence Measure (FIM™) baseline analyses

Sample	n/CI	Item-fit residuals Mean (SD)	Person-fit residuals Mean (SD)	χ^2 p-value	PSI	α	DIF (items)	Paired t-test (Lower ci %), %
MSK_t1	246/4	0.193 (2.496)	−0.183 (1.304)	0.000	0.961	0.967	age (M), language (A, B, D, F, L, R)	9.8 (0.0)
MSK_t2	230/4	0.098 (2.191)	−0.165 (1.359)	0.000	0.966	0.968	language (B, D, F, L, N, P)	17.4 (0.0)
MSK_all	476/8	0.193 (3.255)	−0.155 (1.280)	0.000	0.963	0.967	gender (Q), age L, N), language (B, C, D, F, H, L, M, N, Q, R), time-point (L, M, N, O)	16.2 (0.0)
NEUR_t1	228/4	−0.046 (3.559)	−0.314 (1.745)	0.000	0.964	0.972	language (Q)	17.1 (14.3)
NEUR_t2	242/4	−0.461 (3.449)	−0.358 (1.595)	0.000	0.964	0.973	No DIF	15.3 (12.5)
NEUR_all	470/8	−0.369 (4.919)	−0.349 (1.678)	0.000	0.963	0.972	language (D, F, M, N, P, Q), time-point (L)	15.3 (13.3)
t1_all	474/8	0.101 (4.274)	−0.239 (1.609)	0.000	0.96	0.968	age (F, I, J, N, Q, R), language (B, D, F, L, N, Q), rehab-group (C, E, K, M, O, P, Q, R)	12.9 (10.9)
t2_all	472/8	−0.284 (3.957)	−0.293 (1.553)	0.000	0.964	0.971	language (B, D, M, N, Q), rehab-group (C, E, K, L, O, P, Q)	13.1 (11.2)
FIM_all	946/10	−0.077 (5.779)	−0.265 (1.609)	0.000	0.962	0.969	gender (L), age (N, O), language (B, D, F, H, L, M, N, Q, R), nationality (Q), insurance (O), time-point (L, M), rehab-group (C, E, K, L, M, O, P, Q, R)	11.1 (9.7)
Acceptable values		SD < 1.4	SD < 1.4	> 0.01	> 0.7	> 0.7	No DIF present	At least Lower ci < 5

MSK: musculoskeletal rehabilitation; NEUR: neurological rehabilitation; t1: admission; t2: discharge; all: combination of time-points and/or rehabilitation-groups; n: sample size; CI: class intervals; SD: standard deviation; PSI: Person separation index; α: Cronbach's alpha; DIF: differential item functioning; ci: confidence interval.

Testlet approaches

Within the traditional testlet approach 3 different versions of testlet combinations were applied, based on the underlying subscale structure of the FIM™. Two versions included 4 testlets for the motor scale, structured according to the FIM™ subtopics (self-care, sphincter control, transfers, locomotion) together with 2 combinations of the cognitive items. In one version all the cognitive FIM™ items were unified in one testlet, since they all showed local dependency among each other at the baseline analysis, resulting in a total of 5 testlets. In the other version, the cognitive items were split thematically according to the FIM™ subtopics into 2 testlets, communication and social cognition, resulting in a total of 6 testlets. The third version attempted to form similar sized testlets and was oriented at the residual correlations between the items and formerly reported clusters of the FIM™ (29, 36). In this version, 3 testlets were created: a self-care testlet incorporating items A–H, a mobility testlet incorporating items I–L, and a cognitive testlet incorporating items M–R. None of the 3 traditional testlet approaches, the 3-testlet, the 5-testlet and the 6-testlet version, resulted in fit to the Rasch model (see Table II).

In contrast, the alternative 2-testlet approach (with Testlet1 containing items A, C, E, G, I, K, M, O and Q, and Testlet2 containing items B, D, F, H, J, L, N, P and R) showed fit to the Rasch model across all 9 analyses steps. The *p*-values from the item-trait χ^2 were all non-significant at the 0.01 level, the reliability indexes all above 0.9, and the item- and person-fit estimates within the set acceptable values. The expected common variance values retained in the latent estimate were all just above 1, indicating some marginal remaining residual local dependency among the testlets. The fit of all testlet solutions is summarized in Table II, and the application of the 2-testlet approach to all aggregation levels of the calibration sample is shown in Appendix S2[1].

Differential Item Functioning strategy

Despite overall fit, some DIF remained in the 2-testlet solution for the whole calibration sample. For eliminating all DIF, the successful 2-testlet solution of the whole calibration sample had to be split twice. Testlet2 first had to be split by rehabilitation group. Secondly, the group of musculoskeletal rehabilitation from Testlet2 had to be split into the 2 time-points, i.e. admission and discharge. This resulted in the following superitems: Testlet1, Testlet2_NEUR, Testlet2_MSKt1, and Testlet2_MSKt2. Testlet1 was the anchor for the comparison of the person estimates of the split and the unsplit version. The effect size calculation resulted in 0.11 (see Appendix S3[1]), indicating that there was no need to split the final interval-scale transformation into different subgroups.

Transformation table

Based on the 2-testlet solution, an interval-based transformation table was created for all available FIM™ total scores, which can be used to transfer the

Table II. Testlet solutions on the level of the whole calibration sample (FIM_all)

n/CI	Testlets (items)	Item-fit residuals Mean (SD)	Person-fit residuals Mean (SD)	χ^2 *p*-value	PSI	α	DIF (Testlet)	A	Paired *t*-test, %	Cond. test of fit CI based
946/10	6 Testlets: Self-Care (A-F), Sphincter Control (G-H), Transfers (I-K), Locomotion (L-M), Communication (N-O), Social Cognition (P-R)	−0.156 (5.077)	−0.426 (1.200)	0.000	0.906	0.887	gender (T6), age (T2), language (T1, T2, T3, T4, T5, T6), insurance (T1, T5), time-point (T2, T4, T5), rehab-group (T1, T3, T4, T5, T6)	0.942	1.27	Only available for the 2-testlet approach
946/10	5 Testlets: Self-Care (A-F), Sphincter Control (G-H), Transfers (I-K), Locomotion (L-M), Cognition (N-R)	−0.010 (7.046)	−0.360 (1.138)	0.000	0.895	0.878	age (T2, T5), language (T1, T2, T3, T4, T5), nationality (T5), insurance (T1), time-point (T2, T4), rehab-group (T3, T4, T5)	0.930	1.16	
946/10	3 Testlets: Self-Care (A-H), Mobility (I-M), Cognition (N-R)	−1.419 (6.894)	−0.502 (1.049)	0.000	0.838	0.859	gender (T2, T3), age (T1), language (T1, T2, T3), nationality (T1, T3), insurance (T1), time-point (T2), rehab-group (T2, T3)	0.871	1.27	
946/10	2-testlets: Testlet1 (A, C, E, G, I, K, M, O, Q), Testlet2 (B, D, F, H, J, L, N, P, R)	−0.208 (0.317)	−0.614 (1.003)	0.408	0.980	0.981	rehab-group (T1, T2)	1.019	4.97	0.607
Acceptable values		SD < 1.4	SD < 1.4	> 0.01	> 0.7	> 0.7	No DIF	> 0.9	< 5.00	> 0.01

FIM all: Functional Independence Measure; all: combination of time-points and rehabilitation-groups; *n*: sample size; CI: class intervals; SD: standard deviation; PSI: person separation index; α: Cronbach's alpha; A: explained common variance; DIF: differential item functioning.

Table III. Functional Independence Measure (FIM™) total score transformation table: original scores to interval scores

Original FIM™ score	Rasch estimate	Transformed interval score	Original FIM™ score	Rasch estimate	Transformed interval score
18	−6.279	18.0	73	0.169	75.7
19	−5.686	23.3	74	0.232	76.3
20	−5.281	26.9	75	0.295	76.9
21	−5.005	29.4	76	0.358	77.4
22	−4.784	31.4	77	0.422	78.0
23	−4.594	33.1	78	0.485	78.6
24	−4.423	34.6	79	0.548	79.1
25	−4.263	36.0	80	0.612	79.7
26	−4.112	37.4	81	0.677	80.3
27	−3.966	38.7	82	0.741	80.9
28	−3.824	40.0	83	0.806	81.4
29	−3.686	41.2	84	0.871	82.0
30	−3.552	42.4	85	0.937	82.6
31	−3.420	43.6	86	1.004	83.2
32	−3.291	44.8	87	1.071	83.8
33	−3.166	45.9	88	1.139	84.4
34	−3.043	47.0	89	1.208	85.0
35	−2.923	48.0	90	1.277	85.6
36	−2.807	49.1	91	1.347	86.3
37	−2.693	50.1	92	1.418	86.9
38	−2.582	51.1	93	1.491	87.6
39	−2.473	52.1	94	1.564	88.2
40	−2.367	53.0	95	1.638	88.9
41	−2.264	53.9	96	1.714	89.6
42	−2.163	54.9	97	1.791	90.3
43	−2.064	55.7	98	1.869	90.9
44	−1.968	56.6	99	1.949	91.7
45	−1.874	57.4	100	2.030	92.4
46	−1.783	58.3	101	2.113	93.1
47	−1.693	59.1	102	2.197	93.9
48	−1.605	59.8	103	2.283	94.7
49	−1.519	60.6	104	2.371	95.4
50	−1.435	61.4	105	2.461	96.2
51	−1.353	62.1	106	2.553	97.1
52	−1.273	62.8	107	2.647	97.9
53	−1.194	63.5	108	2.742	98.8
54	−1.116	64.2	109	2.840	99.6
55	−1.041	64.9	110	2.941	100.5
56	−0.966	65.6	111	3.043	101.5
57	−0.893	66.2	112	3.148	102.4
58	−0.821	66.9	113	3.256	103.4
59	−0.750	67.5	114	3.365	104.3
60	−0.680	68.1	115	3.478	105.4
61	−0.611	68.7	116	3.593	106.4
62	−0.543	69.4	117	3.712	107.4
63	−0.476	70.0	118	3.834	108.5
64	−0.409	70.6	119	3.963	109.7
65	−0.343	71.1	120	4.098	110.9
66	−0.278	71.7	121	4.245	112.2
67	−0.213	72.3	122	4.410	113.7
68	−0.149	72.9	123	4.605	115.4
69	−0.085	73.5	124	4.853	117.7
70	−0.021	74.0	125	5.225	121.0
71	0.043	74.6	126	5.784	126.0
72	0.106	75.2			

ordinal-scaled FIM™ raw scores into interval-scaled FIM™ scores (see Table III).

DISCUSSION

This is the first study to provide evidence of the unidimensionality of the FIM™ 18-item version when administered to neurological and musculoskeletal patients in an inpatient rehabilitation setting. Although the baseline analyses and the traditional testlet approaches did not result in Rasch model fit, an alternative 2-testlet approach, emphasizing the sameness of the FIM™ items with 2 equally sized super-items, achieved model fit. The robustness of these results was confirmed in repeating the same strategy for all subsets of the calibration sample. These results provide evidence for the internal construct validity of the FIM™ total scores based on 18 items, and thus support its reporting as a total score in clinical practice. Based on the current results, an interval-scale transformation table of the FIM™ total scores for use in national quality monitoring for neurological and musculoskeletal patients could be provided.

The traditional testlet approach applied in this study builds on the successful Rasch analysis strategy for FIM™ motor items (17). Nevertheless, when adding the cognitive item set to the motor items, the 4 testlet solution in the original paper could not be confirmed in this study. Highlighting the sameness of all the items through the 2-testlet approach attained model fit. The assumption of multidimensionality is often pursued in FIM™-related Rasch analyses (36). The extent of local dependency among the 18 items, clustered into the underlying structures raises questions as to whether the FIM™ should *a priori* be divided into multidimensional concepts, as proposed by Linacre and colleagues (16). The successful summation of the 18 FIM™ items in the current study places emphasis on the higher order construct of functional independence, incorporating both motor and cognitive aspects. Likewise, the Rasch analysis performed supports the theory that, when activities of daily living are observed, motor activities reflect some cognitive aspects and vice versa (18, 37). From a clinical perspective, the FIM™ offers different levels of granularity for reporting. Scores can be reported at the level of the single items, the item headings, e.g. sphincter control, on the level of the motor and cognitive subscales, or the level of the overall 18-item summary of functional independence. Thus, different granular levels of reporting are available, depending upon the use required.

The study can be also be seen as initial evidence that the German, French and Italian translations of the FIM™ do not substantially differ from each other, given the absence of substantial DIF by language. Furthermore, this study provides first evidence for the internal construct validity of the FIM™ 18-item version for musculoskeletal patients, given that no substantial DIF was present between the musculoskeletal and the neurological rehabilitation group. Nevertheless, the use of cognitive items for a musculoskeletal patient population within a national outcome report, as in the ANQ, remains debatable, and care should be

taken to consider whether only motor items should be assessed. In the US model system for burn injury, for example, only the FIM™ motor subscale is assessed (38). However, with regards to the ageing population and related comorbidity (39), both subscales may be of interest in musculoskeletal patients.

This study has the limitations of secondary data analysis. For example, there is a lack of information on the accuracy and consistency of the data-entry process, the selection of DIF factors was limited to the variables of the dataset and the use of non-validated translations. However, this design enabled a well-tailored calibration sample to be obtained from a large sample size. Another limitation is in the 2-testlet approach, which provides the basis for the transformation table. On the one hand, this approach was successful in attaining model fit. On the other hand, the approach does not allow a statement to be made about the hierarchy and difficulty of single items or a conceptually related group of items, since it focuses on the whole construct being measured through the assessment tool. However, while, for the purpose of quality or outcome reports, the FIM™ is based on the total score or change scores, data collection is still conducted on an item level, which allows clinicians to gain insight into the development of a single patient in a certain item or group of items, or to conduct a quality check of scores at the item level if the FIM™ was, for example, applied within a payment system.

The analysis of threshold disordering is also not possible with the testlet approach. Indeed, there is some evidence that disordered thresholds can themselves be caused by local dependency (20, 40). For example, if items are analysed within their subscales, threshold ordering may appear correct, but become disordered when subscales are summated together. Thus, it becomes impossible to determine if disordered thresholds are a consequence of local dependency, as the solution for local dependency renders interpretation of traditional thresholds invalid (20). Since the 2-testlet approach is a relatively new one for health assessment tools, further investigations are needed to confirm the influence of local dependency on such matters. However, this approach has the advantage that the total scores of a well-established and widely used assessment tool, such as the FIM™, can be converted on an interval-scale level, without deleting or re-scoring items.

We recommend the use of the interval transformation table provided in this study for neurological and musculoskeletal patients for national rehabilitation quality monitoring, in order to be able to calculate interval-scaled patient change scores for the FIM™, compared with its original ordinal scoring system (6). If the total scores are available in a digital format, as in the ANQ datasets, transformation can be implemented easily in an electronic information system, by simply re-coding the total scores according to the table provided in the results. This interval scoring system has the advantage that it provides an important basis for the application of a standardized reporting system for functioning information (2, 41) in which the FIM™ could be integrated as a widely used instrument in rehabilitation. This is beneficial, as the standardized reporting of functioning information enables clinicians to continue using currently implemented assessment tools while also being able to compare and aggregate the information within and across tools, institutions or even countries. One caveat to this is that the interval-scale transformation is actually measured with error, as can be seen in its logit form in Appendix S4[1].

In conclusion, the results of this study support the internal construct validity of the FIM™ 18-item version and, consequently, the reporting of its total score, by applying the interval-scaled transformation table provided in this study. The fact that all the variance could be accommodated in the final estimate suggests that previous reports of multidimensionality may have been driven by a breach of the local independence assumption. This supports the intention of its developers and the way the FIM™ scores are used in clinical practice and in institutional and national monitoring. It is recommended to use the interval-scale transformation of the FIM™ total score for national quality monitoring for neurological and musculoskeletal patients, in order to adequately report change scores in patients' functioning. Furthermore, interval transformation provides a basis for integrating the FIM™ into a standardized reporting system for functioning information.

ACKNOWLEDGEMENTS

The authors thank Dr L. Menzi, Head of Rehabilitation ANQ, and K. Schmitt, Corporate Development Director of the Swiss Paraplegic Centre, for their good advice, provision of valuable information and fruitful discussions about the project.

This project is part of the cumulative Dissertation of Roxanne Maritz, which is funded by the Swiss National Science Foundation´s National Research Programme "Smarter Health Care" (NRP 74) within the NRP74 StARS Project "Enhancing continuous quality improvement and supported clinical decision making by standardized reporting of functioning.

The authors have no conflicts of interests to declare.

REFERENCES

1. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. Eur J Phys Rehabil Med 2017; 53: 134–138.
2. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. Eur J Phys Rehabil Med 2017; 53: 144–149.

3. Haigh R, Tennant A, Biering-Sorensen F, Grimby G, Marincek C, Phillips S, et al. The use of outcome measures in physical medicine and rehabilitation within Europe. J Rehabil Med 2001; 33: 273–278.
4. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The Functional Independence Measure: a new tool for rehabilitation. Adv Clin Rehabil 1987; 1: 6–18.
5. Fiedler RC, Granger CV. The Functional Independence Measure: a measurement of disability and medical rehabilitation. In: Chino N. MJL, editor. Functional evaluation of stroke patients. Tokyo: Springer-Verlag; 1996, p. 75–92.
6. Uniform Data System for Medical Rehabilitation. The FIM(R) instrument: its background structure and usefulness. Buffalo: UDSMR; 2012 [cited 2018 Jul 19]. Available from: https://www.udsmr.org/Documents/The_FIM_Instrument_Background_Structure_and_Usefulness.pdf.
7. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken. Review Information: Rehabilitation. 2018 [cited 2018 Jul 19]. Available from: https://www.anq.ch/en/departments/rehabilitation/.
8. Model System Knowledge Translation Center. The Spinal Cord Injury Model System (SCIMS)-SCIMS PowerPoint presentation. 2013 [cited 2018 May 25]. Available from: http://www.msktc.org/About-Model-Systems#SCI.
9. Model System Knowledge Translation Center. The Traumatic Brain Injury Model Systems-TBIMS PowerPoint Presentation. 2017 [cited 2018 May 25]. Available from: http://www.msktc.org/tbi.
10. Canadian Institute for Health Information. National Rehabilitation Reporting System-Privacy Impact Assessment. 2015 [cited 2018 Nov 05]. Available from: https://www.cihi.ca/en/nrs_pia_en.pdf.
11. Australasian Rehabilitation Outcomes Centre. The FIM™ instrument in Australia. [cited 2018 Nov 05]. Available from: https://ahsri.uow.edu.au/aroc/whatisfim/index.html.
12. Carter GM, Relies DA, Ridgeway GK, Rimes CM. Measuring functioning for Medicare inpatient rehabilitation payment. Health Care Financ Rev 2003; 24: 25–44.
13. Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011; 11: 571–585.
14. Ehlan A, Kucukdeveci A, Tennant A. The Rasch measurement model. In: F. F, editor. Research issues in physical & rehabilitation medicine. Pava: Maugeri Foundation; 2010, p. 89–102.
15. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007; 57: 1358–1362.
16. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994; 75: 127–132.
17. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the Functional Independence Measure (FIM) revisited. J Rehabil Med 2011; 43: 884–891.
18. Cantagallo A, Carli S, Simone A, Tesio L. MINDFIM: a measure of disability in high-functioning traumatic brain injury outpatients. Brain Inj 2006; 20: 913–925.
19. Medvedev ON, Turner-Stokes L, Ashford S, Siegert RJ. Rasch analysis of the UK Functional Assessment Measure in patients with complex disability after stroke. J Rehabil Med 2018; 50: 420–428.
20. Wilson M. Detecting and interpreting local item dependence using a family of Rasch models Appl Psychol Meas 1988; 12: 353–364.
21. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018 [cited 2018 Nov 05]. Available from: https://www.R-project.org/.
22. Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol 2008; 8: 33.
23. Linacre JM. Sample size and item calibration (or person measure) stability. Arch Rasch Meas SIG; 1994 [cited 2018 Nov 05]. Available from: https://www.rasch.org/rmt/rmt74m.htm.
24. Hagell P, Westergren A. Sample size and statistical conclusions from test of fit to the Rasch model accodring to the Rasch Unidimensional Measurement Model (RUMM) program in health outcome measurement. J Appl Meas 2016; 17: 416–431.
25. Mallinson T. Rasch analysis of repeated measures. Rasch Meas Trans 2011; 251: 1317.
26. StataCorp. Stata Statistical Software: Release 14. College Station, Texas: StataCor LP; 2015 [cited 2018 Nov 05]. Available from: http://www.stata.com.
27. Andrich D, Sheridan B, Luo G. RUMM2030 Version 5.4 for Windows. Rumm Laboratory Pty Ltd; 2012 [cited 2018 Nov 05]. Available from: http://www.rummlab.com.au/.
28. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174.
29. Andrich D. Components of variance of scales with a bifactor subscale structure from two calculations of alpha. Educ Meas 2016; 35: 25–30.
30. Wainer H, Kiely G. Item clusters and computer adaptive testing: a case for testlets. J Educ Meas 1987; 24: 185–202.
31. Tuerlinckx F, De Boeck P. The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychol Methods 2001; 6: 181–195.
32. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. Appl Psychol Meas 2017; 41: 178–194.
33. Andrich D, Hagquist C. Real and artificial differential item functioning in polytomous items. Educ Psychol Meas 2015; 75: 185–207.
34. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods 2002; 7: 105–125.
35. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale: Lawrence Erlbaum Associates; 1988.
36. Pretz CR, Kean J, Heinemann AW, Kozlowski AJ, Bode RK, Gebhardt E. A multidimensional Rasch analysis of the Functional Independence Measure based on the National Institute on Disability, Independent Living, and Rehabilitation Research Traumatic Brain Injury Model Systems National Database. J Neurotrauma 2016; 33: 1358–1362.
37. Tesio L, Valsecchi MR, Sala M, Guzzon P, Battaglia MA. Level of activity in profound/severe mental retardation (LAPMER): a Rasch-derived scale of disability. J Appl Meas 2002; 3: 50–84.
38. Model System Knowledge Translation Center. Burn injury resources. 2018 [cited 2018 Jul 16]. Available from: https://msktc.org/burn/burn-resources.
39. World Health Organization. World report on ageing and health Geneva, Switzerland: 2015.
40. Hammond A, Tennant A, Tyson SF, Nordenskiold U, Hawkins R, Prior Y. The reliability and validity of the English version of the Evaluation of Daily Activity Questionnaire for people with rheumatoid arthritis. Rheumatology 2015; 54: 1605–1615.
41. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. Eur J Phys Rehabil Med 2018; 54: 110–117.

Chapter 3

The Extended Barthel Index (EBI) can be reported as a unidimensional interval-scaled metric – A psychometric study

Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B, on behalf of the NRP74 clinics. The Extended Barthel Index (EBI) can be Reported as a Unidimensional Interval-Scaled Metric – A Psychometric Study. Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin. 2019; 29(4): 224-232.

This is an article published by Georg Thieme Verlag KG in the Journal Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin on the 13th of March 2019, available online: https://doi.org/10.1055/a-0820-4642

Published online: 2019-03-13

The Extended Barthel Index (EBI) can Be Reported as a Unidimensional Interval-Scaled Metric – A Psychometric Study

Der Erweiterte Barthel Index (EBI) kann als eindimensionale intervallskalierte Metrik berichtet werden – eine psychometrische Studie

OPEN ACCESS

Authors
Roxanne Maritz[1,2], Alan Tennant[1,2], Carolina Saskia Fellinghauer[1], Gerold Stucki[1,2], Birgit Prodinger[1,3], on behalf of the NRP74 StARS clinics*

Affiliations
1 Swiss Paraplegic Research, Nottwil, Switzerland
2 Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland
3 Faculty of Applied Health and Social Sciences, Rosenheim,Technische Hochschule Rosenheim, Germany

Key words
activities of daily living, assessment instruments, outcome assessment, rehabilitation, Rasch model, quality management

Schlüsselwörter
Alltagsaktivität, Ergebnismessung, Qualitätsmanagement, Rehabilitation, Rasch Modell, Assessmentinstrumente

received 07.09.2018
accepted 10.12.2018

Bibliography
DOI https://doi.org/10.1055/a-0820-4642
Published online: 13.3.2019
Phys Med Rehab Kuror 2019; 29: 224–232

ISSN 0940-6689

Correspondence
Roxanne Maritz, MA
Swiss Paraplegic Research
Guido A. Zäch Straße 4
6207 Nottwil
Switzerland
roxanne.maritz@paraplegie.ch

ABSTRACT

Background The Extended Barthel Index (EBI), consisting of the original Barthel Index plus 6 cognitive items, provides a tool to monitor patients' outcomes in rehabilitation. Whether the EBI provides a unidimensional metric, thus can be reported as a valid sum-score, remains to be examined.

Objective To examine whether the EBI can be reported as unidimensional interval-scaled metric for neurological and musculoskeletal rehabilitation.

Methods Rasch analysis of a calibration sample of 800 cases from neurological or musculoskeletal rehabilitation in 2016 in Switzerland.

Results In the baseline analysis no fit to the Rasch Model was achieved. When accommodating local dependencies with a testlet approach satisfactory fit to the Rasch Model was achieved, and an interval scale transformation table was created.

Conclusion The results support the reporting of adapted EBI total scores for both rehabilitation groups by applying the interval scaled transformation table presented in this study.

ZUSAMMENFASSUNG

Hintergrund Der Erweiterte Barthel Index (EBI), der den Barthel Index um 6 kognitive Items ergänzt, ist ein Assessmentinstrument für die Rehabilitation. Ob der EBI eine eindimensionale Metrik liefert und somit als valider Gesamtscore berichtet werden kann, ist unklar.

Ziel Untersuchung ob der EBI für die neurologische und muskuloskelettale Rehabilitation als eindimensionale intervallskalierte Metrik berichtet werden kann.

Methode Rasch-Analyse einer Stichprobe von 800 neurologischen und muskuloskelettalen Rehapatienten aus der Schweiz.

Ergebnisse In der Basisanalyse wurde keine Übereinstimmung mit den Annahmen des Rasch-Modells erreicht. Nachdem lokale Item-Abhängigkeiten mit 2 Testlets angepasst wurden, wurde die Übereinstimmung erreicht und eine intervallskalierte Transformationstabelle erstellt.

Konklusion Die Ergebnisse unterstützen die Verwendung eines angepassten EBI Gesamtscores für beide Rehabilitationsgruppen unter Anwendung der intervallskalierten Transformationstabelle.

* NRP74 StARS clinics: cereneo Schweiz – Robinson Kundert; Hôpital du Valais Spital Wallis, Centre Martigny, Sierre, Brig & Saint-Amé – Els De Waele; Klinik Schönberg – Philipp Banz; Kliniken Valens, Rehazentrum Valens, Rehazentrum Walenstadtberg & Rheinburg-Klinik – Stefan Bachmann, Luzerner Höhenklinik Montana – Jean-Marie Schnyder, Reha Rheinfelden – Thierry Ettlin

Introduction

Functioning is the primary outcome in rehabilitation [1]. Global Activities of Daily Living (ADL) assessment tools that aim to assess functioning are essential for the documentation of the rehabilitation progress and its outcome [2, 3]. Sum-scores of such ADL assessment tools are commonly created by simply summing up the scores of individual items, which often deliver only an ordinal scale of a person's dependency in ADL tasks. There is increasing evidence that treatment decisions based on ordinal level scores can be misinformed [4] as ordinal-level scores can lead to under- or overestimation of the treatment benefit of a person [5]. Therefore, it is essential to transform ordinal measures into interval scales [6]. For this purpose, valid assumptions such as unidimensionality and group invariance need to be established [7].

This issue can be addressed by applying assessment tool data to the Rasch Model. If fit to the Rasch Model can be achieved, and assumptions of local independence and group invariance are supported, an interval-based scoring system can be developed [8].

The Extended Barthel Index (EBI) is such a global ADL tool that is a well-established assessment tool in German speaking countries at the patient, the institutional and the national level [9]. In Germany the EBI is one of the assessment tools used within the ICD-10-GM System as a tool to code restrictions in functioning, that can be relevant for the DRG based payment system [10]. In Switzerland the EBI is one assessment tool used for the national quality monitoring in rehabilitation from the National Association for Quality Development in Hospitals and Clinics (ANQ) [11], part of the CHOP (Swiss classification of treatments for national medical statistics) [12] and will also be part of the DRG based payment system for rehabilitation called ST Reha, that is to be implemented in 2022 [13].

The Extended Barthel Index (EBI) was developed in order to widen the utility of the original Barthel index (BI) [9]. The original BI assesses 10 motor ADL items [14]. The extension of the EBI consists of 6 additional cognitive items, of which 5 are adapted from the FIM™ (Functional Independence Measure), and one – "Vision/Neglect" – is unique to the EBI [9]. Thus, the EBI is a combination of 2 of the most commonly used general outcome measures for rehabilitation, the BI and the FIM™ [15–18]. Due to its simpler rating system and the elimination of some redundant FIM™ items the EBI was recommended over the FIM™, as it increases user-friendliness and compliance [19]. While originally intended for patients with multiple sclerosis, the EBI was also validated and is often applied for other neurological patients, e. g., stroke, traumatic brain injury, or Parkinson's disease [9, 19–23]. Even though the EBI is used for high impact decisions at the patient, institutional and national levels in German speaking countries, no work has been undertaken to-date to explore whether the EBI allows for the calculation of valid sum scores, which would subsequently be eligible for a broad range of statistical analyses. As long as we do not know whether the EBI delivers an ordinal- or interval-scaled unidimensional metric [24] change scores that are based on the EBI can be misleading and have to be interpreted with caution.

Therefore, the objective of the current study was to examine whether the properties of the EBI support its reporting as a unidimensional interval-scaled metric, when administered for national quality monitoring of patients functioning outcomes in neurological and musculoskeletal rehabilitation. This objective resulted in two specific aims: I) To explore the internal construct validity of the EBI and II) to determine if an interval-scale scoring system of the EBI can be made available.

Methods

Subjects and Setting

We conducted a secondary analysis of data routinely collected for the ANQ for national quality monitoring of rehabilitation clinics in Switzerland. We contacted all 64 Swiss rehabilitation clinics which provided musculoskeletal or neurological rehabilitation data to the ANQ in 2016. Thirty clinics agreed to provide their datasets. As the ANQ data collection permits clinics to choose between different ADL assessment tools, not all datasets contained EBI data. For this study we could include datasets from 10 Swiss rehabilitation clinics containing EBI data with in total 5978 complete cases, representing the German and French Swiss language regions. The datasets included data of the EBI on item level, collected at 2 time points – admission and discharge. Ethical approval of the study was requested from all Swiss Ethic Commissions, which stated in a declaration of no objection that the project fulfils the general ethical and scientific standards for research with humans and opposes no health hazards.

Measure

The Extended Barthel Index (EBI) is a clinician-administered scale to assess a patient's need for help with activities of daily living. It consists of 16 items, 10 on physical functioning and 6 on cognitive functioning [9]. The physical functioning items are those from the original Barthel Index [14]: 1-Feeding, 2-Grooming, 3-Dressing, 4-Bathing, 5-Transfer, 6-Mobility, 7-Stairs 8-Toilet use 9-Bowel, and 10-Bladder. The 6 cognitive items are 11-Expression, 12-Comprehension, 13-Social interaction, 14-Problem solving, 15-Memory, and 16-Vision/Neglect. Items 11–15 are adapted from the FIM™. Only item 16 is unique in the EBI. Each item is scored from 0–4, resulting in a total score of 64 [20]. Similar to the BI, not all items represent all categories from 0–4, such as item 1-Feeding that can be scored 0, 2, 3 or 4 (category 1 is missing) or item 13-Social interaction with categories 0, 2, 4 (categories 1 and 3 are missing). The EBI was developed in German [9], the French translation of the EBI used by the participating French speaking clinics, is a non-validated version created by the ANQ.

Sampling

Since a Rasch analysis with a larger sample size is prone to type 1 errors [25], a random stratified calibration sample was obtained using R [26]. The calibration sample contained in total 800 cases, consisting of 4 subsamples containing each 200 cases, each large enough for statistical conclusions and stable item calibration [27, 28]. The 4 subsamples were chosen to equally represent the 2 rehabilitation groups and assessment time points: musculoskeletal cases at admission (MSKt1), musculoskeletal cases at discharge (MSKt2), neurological at admission (NEURt1) and neurological cases at discharge (NEURt2). Cases that were selected for the ad-

mission subsamples were excluded to be selected for the discharge subsamples [29]. Prior to the random selection we deleted all cases with missing values in a variable of interest and all cases with extreme scores (0 or 64) since they cannot be used to estimate item difficulties by the Rasch Measurement Model [30]. In order to be able to give a valuable statement about the whole range of possible total scores of the EBI and the 2 different language regions (German and French) we randomly selected one of each available total scores per subsample and language group. In order to reach 200 cases for each subsample, additional cases were selected by assigning a higher selection probability to rarer total scores in order to best represent the whole range of total scores of the scale. The sampling strategy, with its different subsamples is represented in ▶ **Online Appendix. 1**.

Data analysis

We used descriptive statistics to summarize basic sample characteristics and response distributions. In order to reach specific aim I, Rasch analysis was conducted with the RUMM2030 software [31]. The Partial Credit Model was used, as the EBI has polytomous items with varying lengths [32]. The non-continuous nature of the EBI items response categories required recoding into subsequent categories suitable for the Rasch analysis, resulting in a raw adapted total score ranging from 0–50. The conversion of the original scoring (0–64) to the adapted EBI scoring (0–50) on an item basis is presented in ▶ **Table 1**.

Baseline analysis

To test how well the observed EBI data fitted the Rasch Model, we conducted the baseline analysis on all levels of the calibration sample [33]. To do so we ascertained the person and item fit residuals, the reliability indices α and the Person Separation Index (PSI), and the chi^2 p-value of the item-trait interaction, with the respective acceptable levels represented in the bottom line of the corresponding result table. In addition we investigated local response dependency among items, threshold disordering, and differential item functioning (DIF) for 7 person factors: gender, age (four age groups according to the interquartile ranges), nationality (Swiss or other), insurance status (general, semi-private, private), rehabilitation group (neurological or musculoskeletal rehabilitation), clinic language (German or French) and time point of measurement (admission t1, discharge t2).

Testlet approaches

If the item local independency assumption was not met, testlet approaches combining items into super-items in order to absorb the dependencies in the data were adopted [34–37]. The application of testlets on a related assessment tool, more precisely the FIM™[38] has shown to be an appropriate strategy when dealing with the clustering of items in the underlying subscale structure. In this study we applied 2 different testlet approaches.

Initially, a traditional testlet approach was adopted. This approach emphasises the underlying structure of motor and cognitive items of the EBI. The creation of these testlets was furthermore oriented towards existing local dependencies among items, indicated as standardized residual correlations [39]. Subsequently another testlet approach, referred to as the alternative 2 testlet approach, was used to equally divide items from similar item groups in 2 equally sized testlets, in order to emphasise the 'sameness' of the total item set. This alternative testlet approach, which creates 2 super items, has the advantage of gaining access to additional fit and unidimensionality statistics in RUMM2030 such as the conditional test of fit comparing the observed data with the model expectations, while in the same time satisfying the prerequisite that testlets should be equal in length [34]. Both testlet approaches also allow to report the explained common variance associated with the unidimensional latent estimate, obtained within a bi-factor equivalent approach [34]. The acceptable ranges of these additional statistics are as well indicated at the bottom line in the respective result table [40]. We did not report threshold disordering for the testlet approaches, as it does not allow a meaningful interpretation.

To ensure robustness of the analyses, we conducted the baseline analyses and the testlet approach indicating the best fit to the Rasch Model at three aggregation levels of the calibration sample, represented in ▶ **Online Appendix. 1**. In Level 1 all four subsamples were analysed separately (MSKt1, MSKt2, NEUR t1 and NEURt2). In Level 2 the rehabilitation group and time point subsamples were aggregated separately (MSKt1&t2, NEURt1&t2, t1MSK&NEUR, t2MSK&NEUR). In Level 3 all data were combined, representing the entire calibration sample (EBIall). Likewise, the 3 aggregation levels resulted in nine analytical steps. Throughout, the emphasis of the analyses was upon making the existing EBI work, without the necessity of deleting items or changing its scoring structure other than just making items have consecutive values.

DIF strategy

We analysed DIF in situations in which local dependencies could be accommodated satisfactorily with testlets on the level of the whole calibration sample (EBIall). If lack of invariance between different DIF factors was observed, we split the testlets for the factor with the strongest DIF first and continued, stepwise, until no further DIF was present [41]. We conducted an effect size calculation in order to determine if the splitting makes a substantial difference and should be applied in the final transformation table. The effect size calculation based on the Rasch person estimates from the split and unsplit solutions with estimates from analyses anchored on a DIF free testlet. The effect size calculation was based on the mean of the person estimates, their standard deviations, and the correlation of the split and unsplit version [42]. If the effect size was below 0.2, considered as a small effect size [43], no action was taken to adjust the final transformation table for DIF.

Transformation table

In order to reach specific aim II we sought to create a transformation table in the case that fit to the Rasch Model could be achieved. Based on the solution with the best fit to the Rasch Model, represented by the most satisfactory core values for the whole calibration sample, we constructed an interval-based transformation table of the ordinal adapted EBI total scores (0–50), based on the respective estimates according to the Rasch Model.

▶ **Table 1** Item Conversion on item level original scores (0–64) to the adapted raw score (0–50).

No	Items	EBI 0–64 Categories	EBI 0–50 Categories
1	Feeding	0	0
		2	1
		3	2
		4	3
2	Grooming	0	0
		1	1
		2	2
		3	3
		4	4
3	Dressing	0	0
		1	1
		2	2
		4	3
4	Bathing	0	0
		1	1
		2	2
		3	3
		4	4
5	Transfer	0	0
		1	1
		2	2
		4	3
6	Mobility	0	0
		1	1
		2	2
		3	3
		4	4
7	Stairs	0	0
		1	1
		2	2
		4	3
8	Toilet use	0	0
		1	1
		2	2
		4	3
9	Bowels	0	0
		2	1
		3	2
		4	3
10	Bladder	0	0
		1	1
		3	2
		4	3
11	Expression	0	0
		1	1
		3	2
		4	3
12	Comprehension	0	0
		1	1
		3	2
		4	3
13	Social interaction	0	0
		2	1
		4	2
14	Problem solving	0	0
		2	1
		4	2
15	Memory	0	0
		1	1
		2	2
		3	3
		4	4
16	Vision/ Neglect	0	0
		1	1
		3	2
		4	3
Min		**0**	**0**
Max		**64**	**50**

Results

Sample characteristics

The calibration sample, containing 800 cases in total, contained 400 cases in each rehabilitation group (MSK, NEUR) and 400 in each time point of assessment (admission t1, discharge t2) as defined in the sampling criteria (▶ **Online Appendix. 1**). EBI sum scores (in the 0–64 scoring) had a mean of 43.7 (SD = 14.6, median = 46). The mean age of the selected cases of the calibration sample was 61 years (min = 18, max = 98). The calibration sample contained 53 % (n = 421) male and 47 % (n = 379) female cases, 54 % (n = 432) were in the German-speaking region of Switzerland and 46 % (n = 368) in the French-speaking region, 82 % (n = 659) of the sample were Swiss and 18 % (n = 141) had another nationality. Insurance status related to 80 % (n = 637) general, 11 % (n = 88) semi-private, and 9 % (n = 75) private.

Rasch analysis

Baseline analyses

In the 9 baseline analysis steps no fit to the Rasch Model was achieved (▶ **Table 2**). In all analyses the p-values of the item-trait chi² were significant. Furthermore, in all baseline analyses items showed DIF, threshold disordering and local dependency among diverse items. Threshold disordering and local dependency in the baseline analyses are represented in ▶ **Online Appendix 2**.

▶ **Table 2** EBI baseline analyses with different aggregation levels of calibration sample.

Sample	n / CI	Item fit residuals Mean (SD)	Person fit residuals Mean(SD)	chi² p-value	PSI	α	DIF (item No)	Paired t-test (Lower ci %)
MSKt1	200 / 3	−0.096 (1.814)	−0.438 (1.130)	0.000	0.861	0.856	gender (2), language (2, 5, 7, 8, 14, 15, 16), insurance (16)	7.5% (0.0%)
MSKt2	200 / 3	−0.675 (2.152)	−0.169 (0.900)	0.000	0.849	0.902	gender (2, 16), age (2, 3, 13), language (2, 4, 7, 14, 16), nationality (11, 12)	10.0% (0.0%)
MSKall	400 / 6	−0.583 (2.941)	−0.310 (1.063)	0.000	0.862	0.882	gender (2, 16), age (2, 3,), language (2,4, 5, 7, 8, 10, 14, 16), nationality (11); insurance (3), time-point (2, 3, 5, 8, 10, 13, 14, 15)	9.3% (0.0%)
NEURt1	200 / 3	−0.764 (2.174)	−0.310 (1.025)	0.000	0.911	0.941	age (4), language (3, 4, 5, 8, 11, 14)	8.5% (5.5%)
NEURt2	200 / 3	−0.533 (2.576)	−0.307 (1.175)	0.000	0.918	0.918	language (3, 4, 11, 14), nationality (11), insurance(3, 11)	10.0% (7.0%)
NEURall	400 / 6	−1.009 (3.371)	−0.328 (1.107)	0.000	0.913	0.941	age (3, 4), language (2, 3, 4, 5, 8, 11, 13, 14), nationality (11), insurance (3)	8.8% (6.6%)
t1all	400 / 6	−0.639 (3.109)	−0.333 (1.155)	0.000	0.895	0.914	age (4), language (2, 3, 4, 5, 8, 10, 14, 15, 16), rehab-group (1, 2, 3, 4, 5, 8, 11, 13, 14, 15, 16)	9.3% (7.1%)
t2all	400 / 6	−0.801 (3.101)	−0.259 (1.084)	0.000	0.896	0.933	gender (2, 16), age (2, 3), language (2, 3, 4, 7, 14, 16), insurance (11), nationality (3, 11,12), rehab-group (1, 2, 3, 4, 14, 15)	8.5% (6.4%)
EBIall	800 / 10	−1.083 (4.451)	−0.289 (1.108)	0.000	0.896	0.924	gender (2, 16), age (2, 3, 4, 13, 16), language (2, 3, 4, 5, 7, 8, 10, 11, 13, 14, 16), nationality (3, 11), insurance (3, 16), rehab-group (1, 2, 3, 4, 5, 11, 12, 13, 14, 15, 16), time-point (2, 3, 5, 7, 8, 14)	7.5% (6.0%)
Acceptable values		SD < 1.4*	SD < 1.4*	> 0.01	> 0.7	> 0.7	*No DIF*	*At least Lower ci < 5%*

EBI = Extended Barthel Index, MSK = Musculoskeletal rehabilitation, NEUR = Neurological rehabilitationt1 = admission, t2 = discharge, all = combination of time-points or/and rehabilitation-groups, n = sample size, CI = Class Intervals, SD = standard deviation, PSI = Person Separation Index, α = Cronbach's alpha, DIF = Differential Item Functioning, ci = Confidence Interval, * only applicable for analyses on the item level

Testlet approaches

The traditional testlet approach gave rise to 2 different options – a 4 and a 5 Testlets version of the EBI. For both options, the physical disability items were divided into 3 Testlets, with Testlet1 Self-care (including items 1-Feeding, 2-Grooming, 3-Dressing, 4-Washing), Testlet2 Locomotion (including items 5-Transfer, 6-Mobility, 7-Stairs) and Testlet3 Toileting (8-Toilet use, 9-Bowels, 10-Bladder). For the 4 Testlet version all 6 items of the cognitive scale were collapsed into one testlet. In the 5 Testlet version, the cognitive items were divided into the Testlet4 Communication (including items 11-Comprehension and 12-Expression) and Testlet5 (including 13-Social interaction, 14-Problem solving, 15-Memory and 16-Vision/Neglect). For both – the 4 Testlet and the 5 Testlet version, no fit to the Rasch Model was achieved (▶ **Table 3**).

In the 2-testlet approach, the items were identified as thematic subtopics and then divided equally into the respective 2 testlets: Testlet1 containing items 1-Eating, 3-Dressing, 5-Transfer, 7-Stairs, 9-Bowels, 11-Comprehension, 13-Social interaction, 15-Memory and Testlet2 containing items 2-Grooming, 4-Washing, 6-Mobility, 8-Toilet use, 10-Bladder, 12-Expression, 14-Problem solving, and 16-Vision/Neglect.

With the 2-testlet solution, fit to the Rasch Model was achieved across all nine analyses steps. The item-trait chi² statistics were non-significant, the reliability indexes all above 0.85, and the item and person fit estimates showed acceptable values. Furthermore, the conditional test of fit also indicated fit at eight of the nine analysis steps. Most A-values were marginally above 1, indicating some remaining local dependency among the testlets. The core values of the testlet approaches for the whole calibration sample (EBIall) are summarized in ▶ **Table 3**. The core values for the other 8 subsamples of the successful 2-testlet solution can be found in ▶ **Online Appendix 3**.

DIF strategy

The DIF Strategy is presented in more detail in (▶ **Online Appendix 4**). In order to solve the DIF in the fitting 2-testlet solution of the whole calibration sample, Testlet2 was split four times resulting in the following 6 super-items: Testlet1, Testlet2_NEURgerman, Testlet2_NEURfrench, Testlet2_MSKfrench, Testlet2_MSKgerman_female, Testlet2_MSKgerman_male. Testlet1 was the anchor for the comparison of the person estimates of the split and the unsplit version. The resulting effect size amounted 0.09, indicating that

▲ **Table 3** Testlet approaches with whole EBI calibration sample (EBIall).

Testlets (item No)	Item fit residuals Mean (SD)	Person fit residuals Mean (SD)	chi2 p-value	PSI	α	DIF (Testlet)	A (PSI)	t-test %	CI-based cond. test of fit
4 Testlets: Self-care (1–4), mobility (5–7), Toileting (8–10), Cognition (11–16)	−0.569 (4.952)	−0.338 (0.881)	0.000	0.780	0.845	age (1, 3), language (1, 2, 3, 4), time-point (1, 2), rehab-group (1, 2, 3, 4)	0.870	2.5%	only available for two-testlet approach
5 Testlets: Self-care (1–4), mobility (5–7), Toileting (8–10), Communication (11–12), Social cognition (13–16)	−0.891 (3.931)	−0.369 (0.854)	0.000	0.804	0.844	gender (5), language (1, 3, 5), nationality (4), time-point (1, 2, 5), - rehab-group (1, 2, 4, 5)	0.898	2.9%	only available for two-testlet approach
Two-Testlets: Testlet1 (1, 3, 5, 7, 9, 11, 13, 15), Testlet2 (2, 4, 6, 8, 10, 12, 14, 16)	0.031 (1.676)	−0.476 (0.898)	0.822	0.938	0.950	language (1,2) rehab-group (1,2)	1.047	3.1%	0.013
Acceptable values	*Not applicable for analyses on testlet level*	*Not applicable for analyses on testlet level*	*>0.01*	*>0.7*	*>0.7*	*No DIF*	*>0.9*	*<5%*	*>0.01*

there is no benefit in splitting the final interval scale transformation into different subgroups. (▶ **Online Appendix 5**)

Transformation table

Based on the 2-testlet solution an interval scale based transformation table was created for the EBI 0–50 total raw scores, that can be used to transfer the ordinal EBI score into interval EBI scores, when having data on the item level. This transformation is represented in ▶ **Table 4**.

Discussion

Summary of findings

This study examined the psychometric properties of the EBI, providing first evidence of its internal construct validity for neurological and musculoskeletal patients. Even though no fit to the Rasch Model was achieved at the baseline analyses and with the traditional testlet approaches, we could attain model fit by applying an alternative 2-testlet approach. The robustness of the fit was confirmed at all three aggregation levels and subsets of the calibration sample. The evidence of the EBI's unidimensionality, provides a statement for the internal construct validity of and therefore the reporting of EBI total scores. Furthermore, this study provides an interval scale transformation table of the EBI raw adapted total scores (from 0–50). To avoid bias in reporting change, it is necessary to use the EBI interval scores, as the transformation table shows that changes of a patient at the ends of the score range would be underestimated and changes happening in the middle of the score range would be overestimated if the ordinal EBI raw scores were applied. For example a patient with a EBI raw admission score of 25 and a raw discharge score of 30 would result in a change score of 5 on the raw ordinal basis but only in a change score of 2.9 on the interval level. The transformation table can also be applied for historical analyses when having data on an item level, by applying the conversion table (▶ **Table 1**). This study therefore further provides evidence for the use of the EBI as an ADL assessment tool, consistent with earlier findings [19].

The application of the 2-testlet approach, that divides similar items equally into 2 clusters, highlighting the sameness of all the items in an assessment tool, was successful in attaining model fit. Noteworthy, this approach puts emphasis on a higher order construct of the EBI, incorporating both motor and cognitive aspects, and is the closest that a 2-testlet approach can get to the actual total score. Still, the EBI can offer different levels of granularity: the level of single items out of which some relate conceptually to each other, e. g., item 6 Mobility and 7 Stairs, the level of sub-scales, e. g., the motor and cognitive subscales, and the level of the overall summary score, that is 16 items indicating the independence of a patient in ADL. Depending on the required use, all 3 levels of granularity are available for reporting. In this study the focus was at the level of the overall summary score – finally represented by 2 super-items – to achieve fit to the Rasch Model.

Furthermore, this study offers first evidence for the EBI's application for other patients than neurological patients and it is the first investigation of its French translation [44]. The results support that there is no substantial differential item functioning for the

▶ **Table 4** EBI Total Score Transformation Table – adapted 0–50 EBI raw Scores to EBI Interval Scores.

Adapted raw Score	Rasch Estimate	Transformed interval Score
0	−5.358	0.0
1	−4.621	3.4
2	−4.075	6.0
3	−3.670	7.8
4	−3.340	9.4
5	−3.063	10.7
6	−2.823	11.8
7	−2.611	12.8
8	−2.416	13.7
9	−2.235	14.5
10	−2.064	15.3
11	−1.901	16.0
12	−1.745	16.8
13	−1.594	17.5
14	−1.449	18.1
15	−1.309	18.8
16	−1.173	19.4
17	−1.041	20.0
18	−0.912	20.6
19	−0.786	21.2
20	−0.662	21.8
21	−0.539	22.4
22	−0.418	22.9
23	−0.298	23.5
24	−0.178	24.0
25	−0.057	24.6
26	0.064	25.2
27	0.187	25.7
28	0.312	26.3
29	0.439	26.9
30	0.569	27.5
31	0.703	28.1
32	0.841	28.8
33	0.983	29.4
34	1.131	30.1
35	1.284	30.8
36	1.443	31.6
37	1.609	32.3
38	1.783	33.2
39	1.965	34.0
40	2.155	34.9
41	2.354	35.8
42	2.562	36.8
43	2.782	37.8
44	3.013	38.9
45	3.261	40.0
46	3.531	41.3
47	3.836	42.7
48	4.204	44.4
49	4.712	46.7
50	5.412	50.0

musculoskeletal and the neurological group, and for the German and the French speaking region of Switzerland. The invariance of the two language versions supports the quality of the translations from its original in German into French. Notwithstanding this, there remain questions about the relevance of cognitive EBI items in the musculoskeletal population, particularly the meaningfulness of the item 16-Vision/Neglect remains debatable. For more extensive evidence on group invariance, e. g., regarding patient groups and the French language edition of the EBI, further investigation would be needed.

Limitations of the study

The study brings the limitations of secondary data analysis, for example the limited choice of the person factors for the DIF analyses or the lack of information on consistency and accuracy of data entry. The 2-testlet approach is new, and while it was successful in attaining model fit, it loses the granularity of the individual item approach, as no statement about the hierarchy and difficulty of single items or a conceptually related group of items can be made anymore. Of course this does not preclude the latter, but increasingly evidence is emerging that health assessment tools violate the local item independence assumption more often than not, and this has a damaging effect upon traditional scale interpretation [38, 45]. Thus it is difficult to see how some form of testlet solution could be avoided. The 2-testlet approach has the advantage that the total scores of a well-established assessment tool like the EBI can be converted on an interval scale level, without deleting or rescoring items.

In addition, on the levels of the testlet approaches, the analysis for threshold disordering is absent. There is some initial evidence that threshold disordering can be caused by local dependency [36]. If this is the case, it becomes impossible to conclude if the thresholds disordering is a consequence of local dependency or of it is due to item interpretation. As an example, while item thresholds are ordered within their subscales, thresholds can become disordered when subscales are summated together. Further investigations will be needed to confirm the influence of local dependency.

Another general limitation is the potential ceiling effect for the EBI. While the calibration sample used in the current study avoided that problem, by focusing on a broad representation of the EBI score range.

Application in practice

The clinical and practical relevance of this study is twofold: First, this study provides an empirical argument that the EBI items can be summed up to a single total score. This might not appear surprising since the single total score is widely used in practice. However, the unidimensionality of the EBI has not been proven empirically before. This evidence supports the use of the EBI as an assessment tool in practice. Second, the table to transform the raw score into an interval-based score provided in this study for neurological and musculoskeletal patients, allows for the monitoring of patient changes in EBI scores over time in an empirically sound way. Such monitoring is challenged when using the raw ordinal based EBI scores. The transformation table therefore enables a sound comparison of patient or clinic outcomes, which is a key characteristic for learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in which the EBI as frequently used assessment tool in the German

speaking area, could be included. This is beneficial as a standardized reporting of functioning information enables clinicians to continue using assessment tools while still being able to compare and aggregate the information within and across tools or institutions [47].

The 0–50 adapted raw scores proposed in this study can seem confusing when clinicians want to interpret single EBI scores and are used to the original 0–64 scoring system. However, as long as the data is available on the item level in a digital format, this score transformation can be implemented easily in the background of a dataset by a simple look-up table to convert individual item score back to the original, giving a 0–64 range.

Of note, for the EBI, there already exist different scoring systems. The one that is used in Germany in the ICD-10-GM system is different from the one of the original EBI scoring system that was used in this study, having different numbers of categories for certain items and having different item category values ranging from 0–15 [10]. In order to create an interval transformation table for other EBI scoring systems, the Rasch analysis would need to be repeated with data collected with the different scoring systems. The strategy applied in this study would give a good guidance to do so.

Conclusion

The results support the internal construct validity and therefore also the unidimensionality of the EBI for the neurological and the musculoskeletal rehabilitation groups and therefore the reporting of an adapted raw EBI total score. In order to do so the Rasch transformed and interval scaled EBI total scores ranging from 0–50 developed in this study should be used. This interval-based scoring system of the EBI provides the basis to integrate the EBI in a standardized reporting system of functioning information.

Acknowledgements

Thank goes to Dr. L. Menzi Head of Rehabilitation ANQ and K. Schmitt Corporate Development Director of the Swiss Paraplegic Centre for their good advice, the provision of valuable information and the fruitful discussions about the project. This project is part of the cumulative Dissertation of Roxanne Maritz, which is funded by the Swiss National Science Foundation's National Research Programme "Smarter Health Care" (NRP 74) within the project "Enhancing continuous quality improvement and supported clinical decision making by standardized reporting of functioning".

Conflict of interest

The authors have no other competing interests to declare.

References

[1] Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. European journal of physical and rehabilitation medicine 2017; 53: 134–138

[2] Nelles G. Neurologische Rehabilitation. Stuttgart: Georg Thieme Verlag, 2004

[3] Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. European journal of physical and rehabilitation medicine 2017; 53: 144–149

[4] Doganay Erdogan B, Leung YY, Pohl C et al. Minimal clinically important difference as applied in rheumatology: An OMERACT rasch working group systematic review and critique. The Journal of rheumatology 2016; 43: 194–202

[5] Gorter R, Fox JP, Twisk JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Med Res Methodol 2015; 15: 55

[6] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? Journal of rehabilitation medicine 2012; 44: 97–98

[7] Andrich D. Rating scales and Rasch measurement. Expert review of pharmacoeconomics & outcomes research 2011; 11: 571–585

[8] Christensen KB, Kreiner S, Mesbah M. Rasch Models in Health. London / Hoboken: ISTE Ltd / John Wiley & Sons, Inc.; 2013

[9] Prosiegel M, Böttger S, Schenk T et al. The Extended Barthel Index – a new scale for the assessment of diability in neurological patients [German]. Neurologie & Rehabilitation 1996; 7–13

[10] Deutsches Institut für Medizinische Dokumentation und Information DIMDI https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2018/zusatz-06-barthelindex.htm Accessed [2018 Nov 14]

[11] ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Bern https://www.anq.ch/wp-content/uploads/2017/12/ANQ_Module_23_Auswertungskonzept.pdf Accessed [2018 June 22]

[12] Federal Statistical Office. [https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/publikationen.assetdetail.1940914.html] Accessed [2018 June 22]

[13] Swiss DRG https://www.swissdrg.org/application/files/9315/0997/9991/ST_Reha_Format_und_Inhalt_Daten_2018.pdf Accessed [2018 June 22]

[14] Mahoney FI, Barthel DW. Functional Evaluation: The Barthel Index. Maryland state medical journal 1965; 14: 61–65

[15] Grill E, Stucki G, Scheuringer M et al. Validation of International Classification of Functioning, Disability, and Health (ICF) Core Sets for early postacute rehabilitation facilities: comparisons with three other functional measures. American journal of physical medicine & rehabilitation 2006; 85: 640–649

[16] Haigh R, Tennant A, Biering-Sorensen F et al. The use of outcome measures in physical medicine and rehabilitation within Europe. Journal of rehabilitation medicine 2001; 33: 273–278

[17] Houlden H, Edwards M, McNeil J et al. Use of the Barthel Index and the Functional Independence Measure during early inpatient rehabilitation after single incident brain injury. Clin Rehabil 2006; 20: 153–159

[18] Kwon S, Hartzema AG, Duncan PW et al. Disability measures in stroke: relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. Stroke 2004; 35: 918–923

[19] Marolf MV, Vaney C, Konig N et al. Evaluation of disability in multiple sclerosis patients: a comparative study of the Functional Independence Measure, the Extended Barthel Index and the Expanded Disability Status Scale. Clin Rehabil 1996; 10: 309–313

[20] Jansa J, Pogacnik T, Gompertz P. An evaluation of the Extended Barthel Index with acute ischemic stroke patients. Neurorehabilitation & Neural Repair 2004; 18: 37–41

[21] Bath PM, Iddenden R, Bath FJ et al. Tirilazad for acute ischaemic stroke. The Cochrane database of systematic reviews 2001; CD002087

[22] Jörger M, Beer S, Kesselring J. Impact of neurorehabilitation on disability in patients with acutely and chronically disabling diseases of the nervous system measured by the Extended Barthel Index. Neurorehabilitation & Neural Repair 2001; 15: 15–22

[23] Schuster-Amft C, Henneke A, Hartog-Keisker B et al. Intensive virtual reality-based training for upper limb motor function in chronic stroke: a feasibility study using a single case experimental design and fMRI. Disability & Rehabilitation: Assistive Technology 2015; 10: 385–392

[24] Adroher ND, Prodinger B, Fellinghauer CS et al. All metrics are equal, but some metrics are more equal than others: A systematic search and review on the use of the term 'metric'. PloS one 2018; 13: e0193861

[25] Smith AB, Rush R, Fallowfield LJ et al. Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology 2008; 8: 33

[26] R Core Team. R Foundation for Statistical Computing. Vienna: https://www.R-project.org/ Accessed [2018 Nov 14]

[27] Linacre JM. Archives of the Rasch Measurement SIG. https://www.rasch.org/rmt/rmt74m.htm Accessed [2018 Nov 14]

[28] Hagell P, Westergren A. Sample Size and Statistical Conclusions from Test of Fit to the Rasch Model Accodring to the Rasch Unidimensional Measurement Model (RUMM) Program in Health Outcome Measurement. Journal of Applied Measurement 2016; 17: 416–431

[29] Mallinson T. Rasch Analysis of Repeated Measures. Rasch Measurement Transactions 2011; 251: 1317

[30] Rasch G. Probabilistic models for some intelligence and attainment tests. The University of Chicago Press; Chicago: 1980

[31] Andrich D, Sheridan B, Luo G. Rumm Laboratory Pty Ltd. http://www.rummlab.com.au/ Accessed [2018 Nov 14]

[32] Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174

[33] Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis and rheumatism 2007; 57: 1358–1362

[34] Andrich D. Components of Variance of Scales With a Bifactor Subscale Structure From Two Calculations of alpha. Educational Measurement: Issues and Practice 2016; 35: 25–30

[35] Wainer H, Kiely G. Item clusters and computer adaptive testing: A case for testlets. Journal of Educational Measurement 1987; 185–202

[36] Wilson M. Detecting and interpreting local item dependence using a family of Rasch models. Applied psychological measurement 1988; 353–364

[37] Tuerlinckx F, De Boeck P. The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychological Methods 2001; 6: 181–195

[38] Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. Journal of rehabilitation medicine 2011; 43: 884–891

[39] Christensen KB, Makransky G, Horton M. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. Applied psychological measurement 2017; 41: 178–194

[40] Andrich D. The Polytomous Rasch Model and the Equating of Two Instruments. In: Christensen KB, Kreiner S, Mesbah M. (editors.) Rasch Models in Health. London, UK: ILSTE Ltd; 2013: 164–196

[41] Andrich D, Hagquist C. Real and Artificial Differential Item Functioning in Polytomous Items. Educ Psychol Meas 2015; 75: 185–207

[42] Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods 2002; 7: 105–125

[43] Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale: Lawrence Erlbaum Associates, 1988

[44] ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Bern, Switzerland https://www.anq.ch/wp-content/uploads/2018/02/ANQ_Module_2_EBI.pdf Accessed [2018 Jul 19]

[45] Hammond A, Tennant A, Tyson SF et al. The reliability and validity of the English version of the Evaluation of Daily Activity Questionnaire for people with rheumatoid arthritis. Rheumatology (Oxford, England) 2015; 54: 1605–1615

[46] Mainz J. Defining and classifying clinical indicators for quality improvement. International journal for quality in health care: journal of the International Society for Quality in Health Care 2003; 15: 523–530

[47] Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. European journal of physical and rehabilitation medicine 2018; 54: 110–117

Chapter 4

Creating a common metric based on existing activities of daily living tools to enable standardized reporting of functioning outcomes achieved during rehabilitation

Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B, on behalf of the NRP74 clinics. Creating a common metric based on existing activities of daily living tools to enable standardized reporting of functioning outcomes achieved during rehabilitation. Journal of Rehabilitation Medicine. 2020; 52: jrm00085.

This is an article published by the Foundation of Rehabilitation Information in the Journal of Rehabilitation Medicine on the 10th of July 2020, available online: https://doi.org/10.2340/16501977-2711

J Rehabil Med 2020; 52: jrm00085

ORIGINAL REPORT

CREATING A COMMON METRIC BASED ON EXISTING ACTIVITIES OF DAILY LIVING TOOLS TO ENABLE STANDARDIZED REPORTING OF FUNCTIONING OUTCOMES ACHIEVED DURING REHABILITATION

Roxanne MARITZ, MA[1,2], Alan TENNANT, PhD[1,2], Carolina FELLINGHAUER, PhD[1], Gerold STUCKI, MD, MS[1,2] and Birgit PRODINGER, PhD[1,2,3], on behalf of the NRP74 StARS clinics[4]

From the [1]Swiss Paraplegic Research, Nottwil, Switzerland [2]Department of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland, [3]Faculty of Applied Health and Social Sciences, Technical University of Applied Sciences Rosenheim, Rosenheim, Germany, [4]NRP74 StARS clinics involved in the underlying study (Clinica Hildebrand Centro di riabilitazione Brissago – Giovanni Rabito, Kliniken Valens – Stefan Bachmann, RehaClinic Bad Zurzach and RehaClinic Kilchberg – Serge Altmann and Peter S. Sandor, Klinik Bethesda Tschugg)

Objective: **Many different assessment tools are used to assess functioning in rehabilitation; this limits the comparability and aggregation of respective data. The aim of this study was to outline the development of an International Classification of Functioning, Disability and Health (ICF)-based interval-scaled common metric for 2 assessment tools assessing activities of daily living: the Functional Independence Measure (FIM™) and the Extended Barthel Index (EBI), used in Swiss national rehabilitation quality reports.**

Methods: **The conceptual equivalence of the 2 tools was assessed through their linking to the ICF. The Rasch measurement model was then applied to create a common metric including FIM™ and EBI.**

Subjects: **Secondary analysis of a sample of 265 neurological patients from 5 Swiss clinics.**

Results: **ICF linking found conceptual coherency of the tools. An interval-scaled common metric, including FIM™ and EBI, could be established, given fit to the Rasch model in the related analyses.**

Conclusion: **The ICF-based and interval-scaled common metric enables comparison of patients' and clinics' functioning outcomes when different activities of daily living tools are used. The common metric can be included in a Standardized Assessment and Reporting System for functioning information in order to enable data aggregation and comparability.**

Key words: outcome assessment (healthcare); psychometrics; rehabilitation; activities of daily living; Rasch Measurement Model; Functional Independence Measure; Barthel Index; quality in healthcare.

Accepted Jun 11, 2020; Epub ahead of print Jul 10, 2020

J Rehabil Med 2020; 52: jrm00085

Correspondence address: Roxanne Maritz, Swiss Paraplegic Research, Guido A. Zäch Strasse 4, 6207 Nottwil, Switzerland, E-Mail: roxanne.maritz@paraplegie.ch

LAY ABSTRACT

In our study we developed a common metric serving as a neutral comparator of two assessment tools which are used for assessing activities of daily living in rehabilitation patients in Switzerland. This common metric enables clinicians to use different established assessment tools assessing the same information, while being able to compare the respective information from those different scales on a larger level e.g. comparisons across clinics using different tools. This study is based on the example of Switzerland, where rehabilitation clinics can choose one of two measurement tools assessing activities of daily living, to report their outcome quality. With the common metric, the results from all the clinics can be compared with each other, no matter what tool was used for the assessment, enabling learning and improvement processes.

Functioning is the key indicator for rehabilitation as a health strategy (1). In order to strengthen rehabilitation it is essential to integrate functioning information, through the WHO's International Classification of Functioning, Disability and Health (ICF), into national health information systems, including reports on healthcare quality (2–4). Functioning information, including information on activities of daily living (ADL), is often collected using a variety of assessment tools, which can limit comparability across patients and clinics (5). Two prominent examples of such tools are the Functional Independence Measure (FIM™) and the Barthel Index (BI) (6, 7). Noteworthy, of these well-established tools there are many country and rehabilitation group adapted versions, such as the Extended Barthel Index (EBI) (8), the Modified Barthel Index (9) and the United Kingdom Functional Assessment Measure (UK FIM+FAM) (10). While all of these assessment tools focus on ADLs, their items and scoring structures differ.

There are 2 options for enhancing the comparability of functioning information: first, to define a single assessment tool as the standard, and, secondly, to establish a transformation system between existing established assessment tools. The first option would be difficult to achieve, as there are various reasons for the heterogeneity of the assessment tools in use, such as clinical utility and clinic-specific standards. The latter option, in which a common metric for functioning information is developed, to enable comparison and aggregation of information collected with different tools that measure the same concept, is more feasible (5, 11–15). For this purpose, a Standardized Assessment and Reporting

doi: 10.2340/16501977-2711

System (StARS) for functioning information, with an ICF-based interval-scaled neutral common metric as core element, would enable clinicians to continue using different assessment tools, while at the same time enabling the aggregation and comparison of conceptually equivalent information (5, 13). This common metric could serve as a reporting reference, e.g. for national rehabilitation quality reports, allowing for comparisons between institutions using different tools. A common metric also allows for the transformation of the score of one tool into the score of another conceptually equivalent tool. Furthermore, it provides a transformation of the ordinal-scaled ADL scores to an interval-scale level, which is required to calculate means and change scores between admission and discharge (16).

In order to demonstrate how a StARS of functioning information can be established, this paper takes Switzerland's national rehabilitation quality reports as an example, providing a simplified illustration for the heterogeneous landscape of assessment tools. For Swiss national quality reports in musculoskeletal and neurological rehabilitation, clinics can choose to report with the FIM™ or the EBI, but this impedes the comparison of clinics that use different assessment tools.

Earlier research by Prodinger et al. provided a score transformation of the FIM™ 13-items motor score and the BI on the basis of the Rasch model (17). While the FIM™ 13-item motor scale has been studied extensively (18), its 5 items of cognition have received less attention. Nevertheless, cognitive impairment is important in neurological disorders (19). The importance of cognition was the reason for development of the EBI, extending the BI with 6 cognitive items (8). The current paper therefore seeks to build on the evidence established by Prodinger et al. regarding the motor scales of these 2 assessment tools (17), using the same psychometric approach, but extending it to the 2 assessment tools' versions, including cognitive items. Furthermore, this paper provides a concrete example of how an ICF-based and interval-scaled common metric as a core of a StARS (5), can be created, so that the outcomes of different rehabilitation clinics, using different assessment tools assessing ADLs, can be compared and aggregated.

The objective of this study was to create an ICF-based interval-scaled common metric as a core element of a StARS for functioning information, based on the example of Swiss national quality reports. The approach included the assessment of the 2 key requirements for standardized reporting of health information: (*i*) to determine whether the 2 assessment tools can be considered conceptually equivalent, and (*ii*) to examine whether a reference metric including the FIM™ and EBI can be established by applying the Rasch model (20).

METHODS

Setting and subjects

In Switzerland, the National Association for Quality Development in Hospitals and Clinics (ANQ) coordinates the measurement and public reporting of outcome quality indicators for all rehabilitation clinics (21). For this purpose, functioning information from every patient in neurological and musculoskeletal rehabilitation is collected. Clinics can choose 1 of the 2 tools, either the FIM™ (18-item version) or the EBI for this part of the data collection. To overcome the issue of comparability, the ANQ has commissioned the development of an expert-based transformation algorithm of the 2 assessment tools, called the ANQ ADL Score. In order to create and validate the expert-based transformation algorithm, the Institute of Medical Sociology and Rehabilitation Science from the Charité – Universitätsmedizin Berlin, Germany conducted the respective study from 2015 to 2017 (22). The validation sample included 265 patients undergoing neurorehabilitation from 5 Swiss rehabilitation clinics, representative of the whole continuum of score ranges of the 2 assessment tools. All patients were assessed at admission with both assessment tools, effectively providing a reliable basis for scale-equating procedures, i.e. a common person design (23, 24). Both tools were assessed either in German (4 clinics) or Italian (1 clinic). The data collected in the ANQ ADL Score study, i.e. a prior research project, were used in secondary data analysis in the current study to provide the basis for the ICF-based and interval-scaled common metric. Both the ANQ ADL Score study and the current study were given ethics approval by the respective Swiss Ethics Commission.

ADL assessment tools

The FIM™ is an assessment tool administered by health professionals comprising 18 items. In order to qualify for FIM™ administration, the health professionals received training provided by the ANQ according to the respective FIM™ policy. FIM™ consists of 13 motor items and 5 cognitive items. All items are scored from 1 (total assistance) to 7 (complete independence), summing to a total score ranging between 18 and 126 (25).

The EBI is an assessment tool with 16 items, administered by health professionals. A user manual for the administration of EBI is available, but there is no specific training. Ten motor items are based on the original BI (26), 6 items cover cognitive functioning aspects, of which 5 are derived from the cognitive FIM™ items. One cognitive item is unique to EBI and refers to vision and neglect. All items are scored from 0 to 4, resulting in a total score of 0–64 (8). Not all EBI items contain all scoring categories (e.g. Item 1 Feeding can be scored 0, 2, 3 or 4), therefore an adapted 0–50 scoring version was proposed for Rasch analyses with EBI, which was taken as a basis for this study (27). A conversion from EBI 0–64 scores to EBI 0–50 scores, referred to as EBI50, on item basis can be found in Appendix S1[1].

Recent studies for both tools showed, that in the context of national quality reports they measure a unidimensional construct and can be reported as total scores on the interval-level, when Rasch-based transformation with bi-factor equivalent design is applied (27, 28). In these studies, neither tool showed differential item functioning for sex, age, nationality, healthcare insurance status of patients, time-point of measurement, rehabilitation group (neurological or musculoskeletal) or clinic language (German, French or Italian).

[1]http://www.medicaljournals.se/jrm/content/?doi=10.2340/16501977-2711

ICF linking

The first part of data analysis entailed linking each item of the respective tools to the ICF by using the ICF linking rules, an established method to enable comparability of health information (29). In order to satisfy the first requirement for standardized reporting and scale equating, i.e. to demonstrate the conceptual equivalence of the 2 assessment tools to be integrated into the common metric, the items from each assessment tool were linked to corresponding ICF categories. Furthermore, the perspectives from which information is collected and the categorization of response options were identified for both tools, in accordance with the ICF linking rules. Two researchers, with extensive linking experience (Maritz R., Selb M.), independently linked all items of the FIM™ and the EBI to the ICF, following step by step the refined version of the ICF linking rules (29) in Microsoft® Excel. The results of the independent linking were then compared and discussed. When no agreement on linking of an item could be reached, a third researcher (Prodinger B.) was involved to give advice and reach agreement.

Rasch analysis

The second part of the data analysis was based on Rasch analysis. In order to satisfy the second requirement for standardized reporting, i.e. score equivalence, the polytomous partial credit Rasch measurement model and associated requirements for equating of instruments were applied to derive an interval-scaled common metric from ordinal data (23, 30). These requirements include: unidimensionality; item invariance across sample subgroups, such as age or sex; and local independence, i.e. the demonstration that responses to any item should depend only on the trait (functional independence in the case of EBI and FIM™) and not on responses to other items (31). The analyses were conducted on the total score level of the 2 assessment tools. This was based on the reasoning that: (*i*) the total scores reflect the level of reporting in the national quality reports; (*ii*) previous findings support that the 2 tools can be reported as unidimensional metrics on the total score level (27, 28), representing the construct of functional independence, even though they incorporate both motor and cognitive items; and (*iii*) the recommendations for scale equating by Andrich are fullfilled (23). The data from the validation sample of the ANQ ADL Score study (described above under the subheading "Setting and subjects") were used for Rasch analyses. Basic sample characteristics and descriptive statistics were conducted using Stata Version 14.2, Rasch analysis was conducted with RUMM2030 professional version 5.4.

The analytical focus gave reference to the following 6 key criteria, helping to judge if fit to the Rasch model and its requirements for equating of 2 instruments was achieved: (*i*) the class-interval based conditional test of fit, assessing the observed and the expected scores under the model conditional on each total score through a Pearson χ^2 test; (*ii*) the item-trait interaction χ^2 test, reflecting fit of the data to the Rasch model, also referring to observed and expected scores on the level of class-intervals; (*iii*) the reliability indexes, reported as Cronbach's alpha and person separation index (PSI); (*iv*) differential item functioning (DIF), indicating if there is invariance for different subgroups; (*v*) unidimensionality, expressed as percentage of significant *t*-tests, using individual *t*-tests comparing person-ability estimates for each respondent derived from the subtest analysis; and (*vi*) threshold ordering, indicating whether the different scoring categories of an assessment scale are represented in a successive order (16, 23, 30). Acceptable levels of the key criteria are represented in the bottom row of the corresponding results Table I.

In order to fully examine the defined key criteria, a 2-tiered analysis was used to deal with some restrictions of the analysis software, as FIM™ has 109 scoring options (ranging from 18 to 126) and the RUMM2030 software allows for inclusion of only 101 scoring options. First, analysis of the assessment tools' total scores was performed, in which the FIM™ total scores were rescaled to 0 to 100 together with the EBI50 total scores. This first step is shown in the first row of the corresponding result in Table I, allowing for a conditional test of fit in which both total scores served as 2 items. The FIM™ score was then re-weighted by 1.09 to give the usual operational score of the FIM™. Secondly, the FIM™ items were divided into 2 testlets and the EBI50 items were combined in a third testlet. Which FIM™ item was contained in which testlet, was based on a previous research project about FIM™ total scores for use in national quality reports (28) and is indicated in the legend of Table I. This second analysis step enabled values of the variance in the latent estimate to be obtained, which imply the degree of local dependency remaining in the testlets (shown in the second row of the corresponding result in Table I). Furthermore, the 2 FIM™ testlets could then be taken together as a single super-testlet in a paired *t*-test analysis with the EBI total score, thus addressing the software limitation described, so that the FIM™ total scores can be reported on their original range.

Differential item functioning strategy

DIF was tested for sex, age (4 groups based on interquartile ranges), healthcare insurance status (general, semi-private, private), nationality (Swiss/other), duration of rehabilitation (4 groups based interquartile ranges) and clinic (5 rehabilitation clinics involved) using 4 class intervals. When uniform or non-uniform DIF *p*-values were <0.05, DIF was considered to be present, and the respective testlets were split, starting with the highest DIF, and continuing until no further DIF was present (32). The split and unsplit solutions were then contrasted on the basis of the Rasch person estimates, anchored to each other with an unsplit testlet free of DIF. To determine whether DIF split was necessary for the transformation table, an effect size calculation was performed, based on a Cohen's D calculation, including the mean of the person estimates, their standard deviations, and the sample size of the split and unsplit version (32). If the resulting

Table I. Results of the FIM™ EBI Rasch equating analyses

Data basis	Conditional test of fit *p*-value (DF)	PSI	α	Item- trait χ^2 *p*-value (DF)	Threshold disordering	DIF (testlets)	A Variance	paired *t*-test	Comment
FIM™0–100 – EBI50 total scores	0.504 (112)	0.958	0.848	0.861 (6)	No disordering	Clinic (FIM™, EBI50)	Not applicable	2.64 %	
FIM™ – EBI50 3 testlets FIM1, FIM2, EBI50	Not applicable	0.973	0.981	0.921 (9)	No disordering	Clinic (FIM1, FIM2, EI50)	0.999	6.04 % (lower CI= 3.4%)	FIM testlets taken together for *t*-test and transformation table
Acceptable values	>0.05	>0.70	>0.70	>0.05	No disordering	No substantial DIF	>0.90	<5.00 % (at least lower CI)	

PSI: person separation index; DIF: differential item functioning; A Variance: explained common variance (only available for testlets); DF: degrees of freedom; CI: 95% confidence interval. The FIM1 testlet included FIM™ items A, C, E, G, I, K, M, O & Q and the FIM2 testlet included FIM™ items B, D, F, H J, L, N, P & R. EBI: Extended Barthel Index; FIM™: Functional Independence Measure.

effect size was below 0.1, DIF was considered small and no action was taken, i.e. the unsplit solution was retained (33).

Common metric

If model fit was achieved with both Rasch analysis steps, an interval-scaled common metric was created based on the Rasch location estimates. The common metric is based on a paired t-test from the second Rasch analysis step, with the EBI total score testlet on the one hand and both FIM(TM) testlets together on the other hand. Likewise, the FIM(TM) total scores can be reported on their original range (18–126) and are therefore considered more accurate. The common metric was designed to range from 0 (complete dependence in ADL) to 100 (complete independence in ADL), oriented at similar research projects (13, 17), reflecting minimum and maximum logit estimates derived from the joint analysis.

RESULTS

International Classification of Functioning, Disability and Health linking

An overview of the ICF linking is shown in Table II, and detailed linking at the level of the items is shown in Appendix S2[1]. Both assessment tools' items represent a dependency perspective and all items responses were categorized in the form of an intensity of this dependency (1–7 for FIM™, 0–4 for EBI). The content of the FIM™ items was reflected in 24 ICF categories, and the content of the EBI in 26 ICF categories. Both assessment tools cover the same ICF categories, with the exception of EBI item 16 vision/neglect, linked to the ICF

Table II. Overview of the International Classification of Functioning, Disability and Health (ICF) linking table for the assessment tools: FIM™ and EBI

Perspective of items Categorization of item responses		FIM™ Dependency Intensity	EBI Dependency Intensity
ICF Code & Label			
b	BODY FUNCTIONS		
b1	Mental Functions		
b144	Memory functions	18) Memory	15) Memory
b156	Perceptual functions		16) Vision/Neglect
b2	Sensory functions and pain		
b210	Seeing functions		16) Vision/Neglect
b5	Functions of the digestive, metabolic and endocrine systems		
b525	Defecation functions	08) Bowel management	09) Bowel control
b6	Genitourinary and reproductive functions		
b620	Urination functions	07) Bladder management	10) Bladder control
b7	Neuromusculoskeletal and movement related functions		
d	ACTIVITIES AND PARTICIPATION		
d1	Learning and applying knowledge		
d175	Solving problems	17) Problem solving	14) Problem solving
d2	General tasks and demands		
d3	Communication		
d310	Communicating with - receiving - spoken messages	14) Comprehension	11) Comprehension
d315	Communicating with - receiving - nonverbal messages	14) Comprehension	11) Comprehension
d320	Communicating with - receiving - formal sign language messages	14) Comprehension	11) Comprehension
d325	Communicating with - receiving - written messages	14) Comprehension	11) Comprehension
d330	Speaking	15) Expression	12) Expression
d335	Producing nonverbal messages	15) Expression	12) Expression
d340	Producing messages in formal sign language	15) Expression	12) Expression
d345	Writing messages	15) Expression	12) Expression
d4	Mobility		
d410	Changing basic body position	09) Transfer bed-chair-wheelchair 10) Transfer toilet 11) Transfer tub/shower	05) Transfers
d420	Transferring oneself	09) Transfer bed-chair-wheelchair 10) Transfer toilet 11) Transfer tub/shower	05) Transfers
d450	Walking	12) Walking/using wheelchair	06) Mobility
d455	Moving around	13) Stairs	07) Stairs
d465	Moving around using equipment	12) Walking/using wheelchair	06) Mobility
d5	Self-care		
d510	Washing oneself	2) Grooming 3) Bathing	02) Grooming 04) Bathing
d520	Caring for body parts	2) Grooming	02) Grooming
d530	Toileting	6) Toileting 7) Bladder management 8) Bowel management	08) Toilet use 09) Bowel control 10) Bladder control
d540	Dressing	4) Dressing upper body 5) Dressing lower body	03) Dressing
d550	Eating	1) Eating	01) Feeding
d560	Drinking	1) Eating	01) Feeding
d7	Interpersonal interactions and relationships		
d710	Basic interpersonal relationships	16) Social interaction	13) Social interaction

EBI: Extended Barthel Index, FIM™: Functional Independence Measure.

categories *b156 Perceptual functions and b210 Seeing functions* that are not reflected in the FIM™. Both tools covered predominantly the activities and participation categories of the ICF, with a focus on the ICF chapters *d3 Communication, d4 Mobility and d5 Self-care*. Due to the high level of concordance, the tools were considered conceptually equivalent, i.e. measuring the same latent trait, both covering content related to the concept of functioning.

Sample characteristics

The sample incorporated 265 patients in neurological rehabilitation in Switzerland from 2016 and 2017. Four clinics from the German-speaking and one clinic from the Italian-speaking region of Switzerland each provided between 31 and 70 cases. The sex distribution was 50.8% male ($n = 123$) and 49.2% female ($n = 119$). The mean age of the sample was 67.2 years, ranging from 18 to 92 years, while the mean age of the cases of the different clinics ranged from 62.9 to 73.9 years. The majority (90.0%) were Swiss citizens, 10.0% had other nationalities. Almost three-quarters (74.0%) had general healthcare insurance, 18.6% semi-private insurance and 7.4% private insurance. The mean rehabilitation duration time was 37.7 days, ranging from 5 to 150 days (standard deviation 25.4). Minimal attained FIM™ score of 18 was attained by 7 cases, the maximal score of 126 was not achieved. The minimal EBI score of 0 was achieved by 2, the maximal score by 11 cases. There were missing values for the variables sex, age, healthcare insurance status, rehabilitation duration (each $n = 23$), and origin ($n = 24$).

Rasch analysis

Results of the Rasch equating analyses are shown in Table I. Fit of the data as total scores (first analysis step, first line Table I), and as 3 testlets (second analysis step with 1 EBI testlet, 2 FIM™ super-items, second line of Table I) showed good fit to the Rasch model. The conditional test of fit in the total score analysis and the item-trait χ^2 in both analyses were all non-significant. Reliability was consistent with high-stakes clinical decision-making. All the common variance was subsequently included in the interval-scaled transformation, and both analyses satisfied the monotonic relationship with the functioning trait, thus showing no disordering of thresholds, again a requirement for successful test equating at the scale score level. DIF was present at the clinic level for both tools. A comparison of person estimates between a non-split and split- solution resulted in an effect size of 0.02; thus, it was considered marginal and no action was taken. Detailed information on the DIF strategy and the effect size calculation are shown in Appendix S3[1].

Common metric

Based on the second analysis step, a common metric was created (see Table III), using the total scores of FIM™ and EBI as testlets. The related transformation table for FIM™ and EBI50 total raw scores can be retrieved from the metric and is shown in Appendix S4[1]. For example a FIM™ raw score of 114 can be translated to 72.9 in the common metric, or an EBI50 raw score of 47. The common metric also shows that the operational range of the FIM™ (common metric values from 8.4 to 96.6) is larger than that of the EBI (common metric values from 9.3 to 81.1), indicating that the FIM™ covers a wider range of patient abilities and that there might be ceiling effects with the EBI. The operational ranges of the 2 assessment tools, in contrast to the common metric, are shown in Fig. 1. The fact that the operational range of the common metric is wider than the range of both assessment tools on its own, can be explained by the fact that the calculation of the common metric is based on patients assessed with both tools together, i.e. the common person design of the sample.

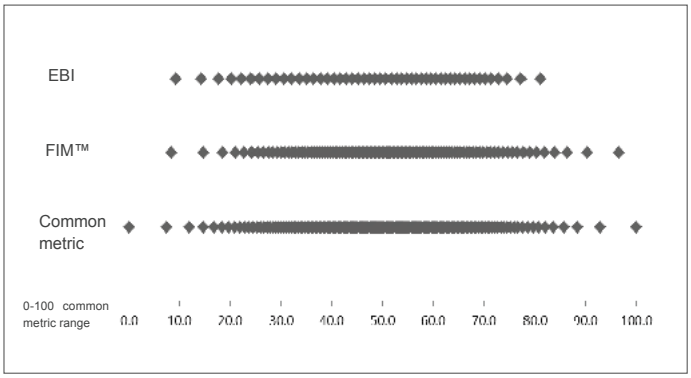


Fig. 1. The 0–100 ICF-based and interval-scaled common metric including the Extended Barthel Index (EBI) and the Functional Independence Measure (FIM™). ICF: International Classification of Functioning, Disability and Health.

Table III. International Classification of Functioning, Disability and Health (ICF)-based interval-scaled common metric 0–100, including Functional Independence Measure (FIM™) and Extended Barthel Index (EBI)50

ICF-based interval-scaled common metric 0-100 including FIM™ and EBI50

common metric raw	RE common metric	Common metric 0-100
0	-3,618	0,0
1	-3,089	7,4
2	-2,763	12,0
3	-2,563	14,8
4	-2,419	16,8
5	-2,304	18,4
6	-2,208	19,8
7	-2,125	20,9
8	-2,052	21,9
9	-1,985	22,9
10	-1,924	23,7
11	-1,868	24,5
12	-1,815	25,3
13	-1,765	26,0
14	-1,718	26,6
15	-1,673	27,2
16	-1,63	27,9
17	-1,588	28,4
18	-1,548	29,0
19	-1,509	29,5
20	-1,472	30,1
21	-1,435	30,6
22	-1,399	31,1
23	-1,364	31,6
24	-1,33	32,1
25	-1,296	32,5
26	-1,263	33,0
27	-1,23	33,5
28	-1,199	33,9
29	-1,167	34,3
30	-1,136	34,8
31	-1,106	35,2
32	-1,076	35,6
33	-1,046	36,0
34	-1,017	36,4
35	-0,988	36,8
36	-0,96	37,2
37	-0,932	37,6
38	-0,904	38,0
39	-0,877	38,4
40	-0,85	38,8
41	-0,824	39,1
42	-0,797	39,5
43	-0,771	39,9
44	-0,746	40,2
45	-0,72	40,6
46	-0,695	40,9
47	-0,67	41,3
48	-0,646	41,6
49	-0,622	42,0
50	-0,598	42,3
51	-0,574	42,6
52	-0,55	43,0
53	-0,527	43,3
54	-0,504	43,6
55	-0,481	43,9
56	-0,459	44,3
57	-0,436	44,6
58	-0,414	44,9
59	-0,392	45,2
60	-0,37	45,5
61	-0,348	45,8
62	-0,326	46,1
63	-0,305	46,4
64	-0,284	46,7
65	-0,262	47,0
66	-0,241	47,3
67	-0,22	47,6
68	-0,199	47,9
69	-0,178	48,2
70	-0,158	48,5
71	-0,137	48,8
72	-0,116	49,1
73	-0,096	49,3
74	-0,075	49,6
75	-0,055	49,9
76	-0,034	50,2
77	-0,014	50,5
78	0,006	50,8
79	0,027	51,1
80	0,047	51,3
81	0,067	51,6
82	0,088	51,9
83	0,108	52,2
84	0,128	52,5
85	0,149	52,8
86	0,169	53,1
87	0,19	53,3
88	0,21	53,6
89	0,231	53,9
90	0,251	54,2
91	0,272	54,5
92	0,293	54,8
93	0,313	55,1
94	0,334	55,4
95	0,355	55,7
96	0,376	56,0
97	0,397	56,2
98	0,419	56,6
99	0,44	56,9
100	0,462	57,2
101	0,483	57,5
102	0,505	57,8
103	0,527	58,1
104	0,549	58,4
105	0,571	58,7
106	0,593	59,0
107	0,616	59,3
108	0,638	59,6
109	0,661	59,9
110	0,684	60,3
111	0,707	60,6
112	0,731	60,9
113	0,754	61,2
114	0,778	61,6
115	0,802	61,9
116	0,826	62,3
117	0,85	62,6
118	0,875	62,9
119	0,9	63,3
120	0,925	63,6
121	0,95	64,0
122	0,975	64,3
123	1,001	64,7
124	1,027	65,1
125	1,053	65,4
126	1,08	65,8
127	1,107	66,2
128	1,134	66,6
129	1,162	67,0
130	1,19	67,4
131	1,218	67,8
132	1,247	68,2
133	1,276	68,6
134	1,306	69,0
135	1,337	69,4
136	1,369	69,9
137	1,401	70,3
138	1,434	70,8
139	1,468	71,3
140	1,504	71,8
141	1,541	72,3
142	1,58	72,8
143	1,62	73,4
144	1,663	74,0
145	1,709	74,6
146	1,757	75,3
147	1,809	76,0
148	1,864	76,8
149	1,925	77,7
150	1,991	78,6
151	2,064	79,6
152	2,146	80,8
153	2,241	82,1
154	2,354	83,7
155	2,497	85,7
156	2,693	88,4
157	3,009	92,8
158	3,52	100,0

FIM™ raw	RE FIM™	Common metric 0-100
18	-3,015	8,4
19	-2,568	14,7
20	-2,290	18,6
21	-2,116	21,0
22	-1,989	22,8
23	-1,886	24,3
24	-1,799	25,5
25	-1,723	26,5
26	-1,654	27,5
27	-1,590	28,4
28	-1,531	29,2
29	-1,475	30,0
30	-1,421	30,8
31	-1,370	31,5
32	-1,320	32,2
33	-1,272	32,9
34	-1,226	33,5
35	-1,181	34,1
36	-1,137	34,8
37	-1,094	35,4
38	-1,051	36,0
39	-1,010	36,5
40	-0,970	37,1
41	-0,931	37,6
42	-0,892	38,2
43	-0,854	38,7
44	-0,817	39,2
45	-0,781	39,7
46	-0,745	40,2
47	-0,710	40,7
48	-0,676	41,2
49	-0,642	41,7
50	-0,609	42,2
51	-0,576	42,6
52	-0,544	43,1
53	-0,513	43,5
54	-0,481	43,9
55	-0,450	44,4
56	-0,420	44,8
57	-0,390	45,2
58	-0,360	45,6
59	-0,331	46,0
60	-0,302	46,5
61	-0,273	46,9
62	-0,244	47,3
63	-0,215	47,7
64	-0,187	48,1
65	-0,159	48,5
66	-0,130	48,9
67	-0,102	49,3
68	-0,074	49,6
69	-0,046	50,0
70	-0,018	50,4
71	0,010	50,8
72	0,037	51,2

FIM™ raw cont'd	RE FIM™	Common metric 0-100
73	0,065	51,6
74	0,094	52,0
75	0,122	52,4
76	0,150	52,8
77	0,178	53,2
78	0,207	53,6
79	0,236	54,0
80	0,265	54,4
81	0,294	54,8
82	0,323	55,2
83	0,353	55,6
84	0,383	56,1
85	0,413	56,5
86	0,444	56,9
87	0,475	57,3
88	0,507	57,8
89	0,539	58,2
90	0,571	58,7
91	0,604	59,1
92	0,638	59,6
93	0,672	60,1
94	0,707	60,6
95	0,742	61,1
96	0,778	61,6
97	0,815	62,1
98	0,852	62,6
99	0,891	63,2
100	0,930	63,7
101	0,969	64,3
102	1,010	64,8
103	1,052	65,4
104	1,094	66,0
105	1,138	66,6
106	1,182	67,2
107	1,228	67,9
108	1,275	68,5
109	1,323	69,2
110	1,372	69,9
111	1,423	70,6
112	1,476	71,4
113	1,532	72,1
114	1,589	72,9
115	1,650	73,8
116	1,714	74,7
117	1,782	75,7
118	1,855	76,7
119	1,934	77,8
120	2,020	79,0
121	2,117	80,3
122	2,230	81,9
123	2,367	83,8
124	2,547	86,4
125	2,832	90,4
126	3,280	96,6

EBI50 raw	RE EBI50	Common metric 0-100
0	-2,951	9,3
1	-2,597	14,3
2	-2,349	17,8
3	-2,174	20,2
4	-2,029	22,3
5	-1,900	24,1
6	-1,779	25,8
7	-1,662	27,4
8	-1,548	29,0
9	-1,435	30,6
10	-1,326	32,1
11	-1,219	33,6
12	-1,116	35,1
13	-1,015	36,5
14	-0,917	37,8
15	-0,822	39,2
16	-0,730	40,5
17	-0,640	41,7
18	-0,553	42,9
19	-0,468	44,1
20	-0,385	45,3
21	-0,304	46,4
22	-0,224	47,5
23	-0,147	48,6
24	-0,071	49,7
25	0,004	50,7
26	0,077	51,8
27	0,149	52,8
28	0,220	53,8
29	0,290	54,7
30	0,360	55,7
31	0,429	56,7
32	0,497	57,6
33	0,566	58,6
34	0,634	59,6
35	0,702	60,5
36	0,770	61,5
37	0,838	62,4
38	0,906	63,4
39	0,974	64,3
40	1,042	65,3
41	1,111	66,3
42	1,180	67,2
43	1,250	68,2
44	1,324	69,2
45	1,401	70,3
46	1,487	71,5
47	1,588	72,9
48	1,713	74,7
49	1,898	77,3
50	2,173	81,1

RE: Rasch estimate (location); cont'd: continued.

DISCUSSION

This study provides an ICF-based and interval-scaled common metric, including the transformation between FIM™ and EBI for use in national rehabilitation quality reports, and thus facilitates a Standardized Assessment and Reporting System (StARS) for functioning information. In order to create the common metric, 2 steps were followed; first, ICF linking showed that the 2 assessment tools can be considered as conceptually equivalent. In the second step a common metric including the 2 tools could be established, as fit to the Rasch models' requirements for the equating of instruments was achieved. The current study provides an example of how ADL scales assessing the same information can be made comparable through the ICF on the conceptual level and an interval-scaled common reference metric on the scale level. The methodology applied can inform further research, in which conceptually similar functioning information is collected with different tools and needs to be made comparable or aggregated. Such a system could, for example, be used in areas other than national quality reports, such as systematic reviews and meta-analyses in the field of rehabilitation (34), or country comparisons of rehabilitation outcomes with different versions of the BI or the FIM.

The common metric established in this study has 4 major advantages over other transformation systems, such as, for example, an expert-based transformation system developed in the original ANQ ADL Score study (22) or scale equating without a reference metric, such as equating using the Leunbach model (35). First, the common metric approach is based on the ICF, the global standard for conceptualizing and describing functioning information, serving as neutral and conceptual reference to compare the content of the different tools included in the common metric (4, 36). Secondly, Rasch analysis allows an interval-based scoring and transformation table to be derived to support the calculation of valid change scores of functioning, e.g. between admission and discharge, which can inform clinical practice and research about functioning change in a quantifiable way (37). Thirdly, the common metric allows us to take the operational ranges of the 2 assessment tools into account, and can therefore operate at the level of the more detailed tool (17). Fourthly, through the ICF-based common metric, other assessment tools measuring functioning could be added in future (5). Given that this study was conducted in the context of Swiss national quality reports, only 2 assessments were included. In principle, any number of instruments can be integrated into a common metric as long as they are conceptually equivalent. Examples with more than 2 scales exist and have been published (5, 13, 20).

The comparison of the operational ranges of the 2 assessment tools showed that FIM™ covers a wider operational range than EBI. This finding can inform the choice of assessment tools and the interpretation of change scores. The basis of adding further assessment tools to the common metric would again be the linking of the respective assessment tool with the ICF and a person-equating design, in which the new tool to be added to the metric is assessed in parallel with either EBI or FIM™. As FIM™ is the assessment tool with the wider operational range, and is more widely used in rehabilitation in general, and was also previously used in other scale-equating projects, e.g. with the Barthel Index or the minimal dataset (11, 17, 38), it appears to be the choice as a linking scale.

Attention should also be paid to the equating design, as indirect transformation between 2 tools, i.e. the equating of instruments via one instrument that is already included in a transformation set, have been shown to be less precise than the direct transformation, i.e. the equating of different instruments in the same study via a common person design (35). However, indirect transformations are likely to be more feasible in terms of data collection. This reflects typical challenges in the practice of quality improvement work (39). A limitation of the study is that the dataset only covers data from 2 assessment tools and includes only neurological rehabilitation patients. The common metric would also be needed for other assessment tools or rehabilitation groups, such as that of musculoskeletal patients in the example of the Swiss quality reports. As previous studies of each assessment tool showed no substantial DIF between neurological and musculoskeletal rehabilitation patients (27, 28), the transformation table could preliminarily be applied to the musculoskeletal ANQ reports. Nevertheless, this should be validated in a future equating study including musculoskeletal data. Another limitation of the study was the restrictions of the RUMM2030 software not being able to cover total score or testlets with more than 100 thresholds (e.g. a score of 0–100), leading to some adaptions in the analyses. Likewise, the Rasch analyses had to be 2-tiered in order to enable judgement about the model fit. Furthermore, the current system does not facilitate the separation of cognitive from motor performance, as it is based on the total scores of the ADL tools, which reflect the level of reporting of the Swiss quality reports in rehabilitation. However, previous studies on EBI (27) and FIM™ (28) showed that the total scores, including motor and cognitive items, can be reported as a unidimensional construct, i.e. functional independence. Another methodological challenge was that the weight of ICF linking was not clearly defined, i.e. there is no clear cut-off when

conceptual equivalency is achieved. However, with the methodological approach chosen, the conclusion drawn, based at the conceptual level on the results of the ICF linking, was supported at an empirical level by the Rasch analysis.

A strength of the present study, in addition to the above-mentioned advantages of the common metric, includes the use of the Rasch model as a basis to create the common metric. In the Rasch model the raw score is a sufficient statistic, such that there is a one-to-one correspondence between the raw score and the latent estimate, which is not the case in other item response theory or classical test theory models (16). Furthermore, the resulting estimates of the Rasch model, which build the basis for the common metric, are distribution free, given the person and item parameter separation. The common metric is therefore not sample dependent and can be applied to any relevant sample (16).

Application in practice

The ICF-based and interval-scaled common metric including the FIM™ and the EBI, as provided in this study, can be applied in national quality reports for neurological rehabilitation and, based on previous evidence with regards to DIF, for musculoskeletal rehabilitation. Application of the transformation table enables the comparison of clinics that are using different assessment tools and, at the same time, the calculation of valid change scores for patients between admission and discharge, as the transformation is based on the interval-scale level. Extension of the common metric with other functioning-related measures is possible. Possible extensions could be informed on the basis of the conducted ICF linking in comparison with the ICF Rehabilitation Set, also referred to as the ICF Generic-30, which was designed to define ICF categories as a minimal standard for reporting and assessing functioning and disability in clinical populations along the continuum of care (40).

Conclusion

The ICF-based and interval-scaled common metric provided in this study supports the assessment of patient outcomes using different ADL assessment tools (FIM™ and EBI) in clinical practice, while at the same time being able to compare the related outcomes of different clinics. It lays the groundwork for a standardized reporting system of functioning information for use in national quality improvement reports.

ACKNOWLEDGEMENTS

The authors thank Melissa Selb from the ICF Research Branch for her support in the ICF linking, Martin Brünger, Stefanie Köhn, Anna Schlumbohm and Karla Spyra from the Institute of Medical Sociology and Rehabilitation Science of the Charité – Universitätsmedizin Berlin, Germany for providing and supporting us with the data from the ANQ ADL Score Study, Luise Menzi, Head of Rehabilitation ANQ, Klaus Schmitt, former Corporate Development Director and Anke Scheel-Sailer, Head of Clinical Quality Management Research Department from the Swiss Paraplegic Centre, for their provision of valuable advice for the project. This project is part of the cumulative Dissertation of Roxanne Maritz, which is funded by the Swiss National Science Foundation's National Research Programme "Smarter Health Care" (NRP 74) within the NRP74 StARS project.

The authors have no conflicts of interest to declare.

REFERENCES

1. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. Eur J Phys Rehabil Med 2017; 53: 134–138.
2. Stucki G, Bickenbach J, Melvin J. Strengthening rehabilitation in health systems worldwide by integrating information on functioning in national health information systems. Am J Phys Med Rehabil 2017; 96: 677–681.
3. World Health Organization (WHO). Rehabilitation in health systems: guide for action. 2019 [cited 2020 Mar 5). Available from: https://www.who.int/rehabilitation/rehabilitation-guide-for-action/en/.
4. World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva: WHO, 2001.
5. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. Eur J Phys Rehabil Med 2018; 54: 110–117.
6. Haigh R, Tennant A, Biering-Sorensen F, Grimby G, Marincek C, Phillips S, et al. The use of outcome measures in physical medicine and rehabilitation within Europe. J Rehabil Med 2001; 33: 273–278.
7. Grill E, Stucki G, Scheuringer M, Melvin J. Validation of International Classification of Functioning, Disability, and Health (ICF) Core Sets for early postacute rehabilitation facilities: comparisons with three other functional measures. Am J Phys Med Rehabil 2006; 85: 640–649.
8. Prosiegel M, Böttger S, Schenk T, König N, Marolf M, Vaney C, et al. Der erweiterte Barthel - Index (EBI) – eine neue Skala zur Erfassung von Fähigkeitsstörungen bei neurologishen Patienten. Neurol Rehabil 1996; 1: 7–13.
9. Ohura T, Hase K, Nakajima Y, Nakayama T. Validity and reliability of a performance evaluation tool based on the modified Barthel Index for stroke patients. BMC Med Res Methodol 2017; 17: 131.
10. Nayar M, Vanderstay R, Siegert RJ, Turner-Stokes L. The UK Functional Assessment Measure (UK FIM+FAM): psychometric evaluation in patients undergoing specialist rehabilitation following a stroke from the National UK Clinical Dataset. PloS One 2016; 11: e0147288.
11. Li CY, Romero S, Bonilha HS, Simpson KN, Simpson AN, Hong I, et al. Linking existing instruments to develop an activity of daily living item bank. Eval Health Prof 2018; 41: 25–43.
12. Smith RM, Taylor PA. Equating rehabilitation outcome scales: developing common metrics. J Appl Meas 2004; 5: 229–242.
13. Prodinger B, Stamm T, Peterson D, Stucki G, Tennant A. Toward a standardized reporting of outcomes in hand osteoarthritis: developing a common metric of outcome measures commonly used to assess functioning. Arthritis Care Res 2016; 68: 1115–1127.
14. Victorson D., Schalet BD, Kundu S, Helfand BT, Novakovic K, Penedo F, et al. Establishing a common metric for self-reported anxiety in patients with prostate cancer: Linking

the Memorial Anxiety Scale for Prostate Cancer with PROMIS Anxiety. Cancer 2019; 125: 3249–3258.
15. Oude Voshaar MAH, Vonkeman HE, Courvoisier D, Finckh A, Gossec L, Leung YY, et al. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Life Res 2019; 28: 187–197.
16. Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011; 11: 571–585.
17. Prodinger B, O'Connor RJ, Stucki G, Tennant A. Establishing score equivalence of the Functional Independence Measure motor scale and the Barthel Index, utilising the International Classification of Functioning, Disability and Health and Rasch measurement theory. J Rehabil Med 2017; 49: 416–422.
18. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. J Rehabil Med 2011; 43: 884–891.
19. Berlucchi G. Brain plasticity and cognitive neurorehabilitation. Neuropsychol Rehabil 2011; 21: 560–578.
20. Prodinger B, Tennant A, Stucki G, Cieza A, Ustun TB. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. J Health Serv Res Policy 2016; 21: 223–228.
21. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken. Review information: rehabilitation. 2020 [cited 2020 Mar 6]. Available from: https://www.anq.ch/en/departments/rehabilitation/.
22. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken. Abschlussbericht Entwicklung und Validierung eines ADL–Überführungsalgorithmus auf Basis von FIM® und EBI 2018. [cited 2020 Mar 6]. Available from: https://www.anq.ch/wp-content/uploads/2018/03/ANQ_Reha_Abschlussbericht_ADL_Algorithmen.pdf.
23. Andrich D. The Polytomous Rasch Model and the equating of two instruments. In: Christensen KB, Kreiner, S, Mesbah M, editors. Rasch models in health. Hoboken: ISTE Ltd and John Wiley and Sons Inc.; 2013, p. 163–196.
24. Kolen MJ, Brennan RL. Introduction and concepts. In: Kolen MJ, Brennan RL, editors. Test equating, scaling, and linking. New York: Springer; 2004, p. 1–28.
25. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The Functional Independence Measure: a new tool for rehabilitation. Adv Clin Rehabil 1987; 1: 6–18.
26. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. Md State Med J 1965; 14: 61–65.
27. Maritz R, Tennant A, Fellinghauer CS, Stucki G, Prodinger B. The Extended Barthel Index (EBI) can be reported as a unidimensional interval-scaled metric – a psychometric study. Phys Med Rehab Kuror 2019; 29: 224–232.
28. Maritz R, Tennant A, Fellinghauer C, Stucki G, Prodinger B. The Functional Independence Measure 18-item version can be reported as a unidimensional interval-scaled metric: internal construct validity revisited. J Rehabil Med 2019; 51: 193–200.
29. 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.
30. Ehlan AH, Kucukdeveci AA, Tennant A. The Rasch measurement model. In: Franco Franchignoni (editor). Research issues in physical & rehabilitation medicine. Pavia: Maugeri Foundation; 2010, p. 89–102.
31. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007; 57: 1358–1362.
32. Andrich D, Hagquist C. Real and Artificial differential item functioning in polytomous items. Educ Psychol Meas 2015; 75: 185–207.
33. Rouquette A, Hardouin JB, Vanhaesebrouck A, Sébille V, Coste J. Differential Item Functioning (DIF) in composite health measurement scale: recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. PloS One 2019; 14: e0215073.
34. Stucki G, Pollock A, Engkasan JP, Selb M. How to use the International Classification of Functioning, Disability and Health as a reference system for comparative evaluation and standardized reporting of rehabilitation interventions. Eur J Phys Rehabil Med 2019; 55: 384–394.
35. Adroher ND, Kreiner S, Young C, Mills R, Tennant A. Test equating sleep scales: applying the Leunbach's model. BMC Med Res Methodol 2019; 19: 141.
36. World Health Organization, WHO. How to use the ICF – a practical manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for comment. 2013 (cited 2020 Mar 6). Available from: http://www.who.int/classifications/drafticfpracticalmanual2.pdf.
37. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med 2012; 44: 97–98.
38. Velozo CA, Byers KL, Wang YC, Joseph BR. Translating measures across the continuum of care: using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. J Rehabil Res Dev 2007; 44: 467–478.
39. Needham DM, Sinopoli DJ, Dinglas VD, Berenholtz SM, Korupolu R, Watson SR, et al. Improving data quality control in quality improvement projects. Int J Qual Health Care 2009; 21: 145–150.
40. Prodinger B, Cieza A, Oberhauser C, Bickenbach J, Ustun TB, Chatterji S, et al. Toward the International Classification of Functioning, Disability and Health (ICF) Rehabilitation Set: a minimal generic set of domains for rehabilitation as a health strategy. Arch Phys Med Rehabil 2016; 97: 875–884.

Chapter 5

The influence and added value of a Standardized Assessment and Reporting System for functioning outcomes upon national rehabilitation quality reports

Maritz R, Ehrmann C, Prodinger B, Tennant A, Stucki G. The influence and added value of a Standardized Assessment and Reporting System for functioning outcomes upon national rehabilitation quality reports. International Journal for Quality in Health Care. 2020; 32(6), 379-387.

This is an article published by Oxford University Press in the International Journal for Quality in Health Care on the 1st of July 2020, available online:
https://doi.org/10.1093/intqhc/mzaa058

International Journal for Quality in Health Care, 2020, 32(6), 379–387
doi: 10.1093/intqhc/mzaa058
Advance Access Publication Date: 1 July 2020
Research Article

Research Article

The influence and added value of a Standardized Assessment and Reporting System for functioning outcomes upon national rehabilitation quality reports

ROXANNE MARITZ[1,2], CRISTINA EHRMANN[2], BIRGIT PRODINGER[1,2,3], ALAN TENNANT[2], and GEROLD STUCKI[1,2]

[1]Department of Health Sciences and Medicine, University of Lucerne, 6002 Lucerne, Switzerland, [2]Swiss Paraplegic Research, 6207 Nottwil, Switzerland, and [3]Faculty of Applied Health and Social Sciences, Technical University of Applied Sciences Rosenheim, 83024 Rosenheim, Germany

Address reprint requests to: Roxanne Maritz, Swiss Paraplegic Research, Guido A. Zächstrasse 4, 6207 Nottwil, Switzerland Fax: +41 41 939 66 40; E-mail: roxanne.maritz@paraplegie.ch

Received 28 February 2020; Revised 24 April 2020; Editorial Decision 11 May 2020; Accepted 11 May 2020

Abstract

Objective: To demonstrate the influence and added value of a Standardized Assessment and Reporting System (StARS) upon the reporting of functioning outcomes for national rehabilitation quality reports. A StARS builds upon an ICF-based (International Classification of Functioning, Disability and Health) and interval-scaled common metric.
Design: Comparison of current ordinal-scaled Swiss national rehabilitation outcome reports including an expert-consensus-based transformation scale with StARS-based reports through descriptive statistical methods and content exploration of further development areas of the reports with relevant ICF Core Sets.
Setting: Swiss national public rehabilitation outcome quality reports on the clinic level.
Participants: A total of 29 Swiss rehabilitation clinics provided their quality report datasets including 18 047 patients.
Interventions: Neurological or musculoskeletal rehabilitation.
Main outcome measures: Functional Independence Measure™ or Extended Barthel Index.
Results: Outcomes reported with a StARS tended to be smaller but more precise than in the current ordinal-scaled reports, indicating an overestimation of achieved outcomes in the latter. The comparison of the common metric's content with ICF Core Sets suggests to include 'energy and drive functions' or 'maintaining a basic body position' to enhance the content of functioning as an indicator.
Conclusions: A StARS supports the comparison of outcomes assessed with different measures on the same interval-scaled ICF-based common metric. Careful consideration is needed whether an ordinal-scaled or interval-scaled reporting system is applied as the magnitude and precision of reported outcomes is influenced. The StARS' ICF basis brings an added value by informing further development of functioning as a relevant indicator for national outcome quality reports in rehabilitation.

Key words: quality of health care, public reporting of healthcare data, outcome assessment (health care), rehabilitation, international classification of functioning, disability and health, psychometrics

Introduction

The measurement and monitoring of clinical performance are central for hospital quality improvement [1]. For the monitoring of institutional outcomes in national quality reports, the main health indicators of a health system need to be addressed, and for rehabilitation this indicator is functioning [2, 3]. Functioning is classified by the World Health Organization's International Classification of Functioning, Disability and Health (ICF), incorporating both biological health—the intrinsic health capacity described as body functions and structures, as well as lived health—the actual engagement of a person in activities and life situations in interaction with the environment [4, 5].

Currently, in rehabilitation, functioning outcomes are assessed with a variety of ordinal-scaled assessment tools, which makes it difficult to compare, aggregate [6, 7] and eventually learn from the related information for improvement processes. Therefore, standardization is essential for measurement of achieved outcomes within clinics, and critical for comparisons between clinics [1]. A concrete example comes from the Swiss public national rehabilitation outcome quality reports, in which musculoskeletal and neurological rehabilitation clinics can choose between two ordinal-scaled assessment tools assessing functioning outcomes in the domain of activities of daily living (ADL)—the Functional Independence Measure (FIM^{TM}) or the Extended Barthel Index (EBI) [8].

To standardize outcomes, two approaches exist: 1) define specific assessment tools, which have to be used by all stakeholders or 2) enable standardized reporting and thus comparability of routinely used assessment tools [6]. For the latter, different approaches, such as expert-consensus-based transformations [9] or Standardized Assessment and Reporting Systems (StARS) for functioning outcomes, based on a statistical, i.e. Rasch-based scale transformation approach [10] can be applied. Expert-consensus-based transformations have the advantage that experienced clinicians are involved in the process but the disadvantage that ordinal-scale properties of outcomes remain, thus restricting valid calculations of means or change scores [7, 11]. In contrast, a StARS for functioning outcomes includes a common metric as a core element, which has two main features: first, it is conceptually based on the ICF as the international standard for reporting functioning information and second, it is interval-scaled, thus, allowing for any parametric analyses in reporting and monitoring outcomes [6, 12].

National quality reports provide an excellent opportunity to examine the influence of different approaches toward standardized reporting of clinic outcomes. Therefore, the objective of the current study was to demonstrate the influence and added value of a StARS, with its interval-scaled ICF-based common metric, upon the reporting of functioning outcomes in national rehabilitation quality reports. Specific aims were related to the common metric's two main features:

1) To examine the influence of the common metric's interval-scaling feature in comparison to (i) functioning outcomes reported with ordinal-scaled assessment total scores and (ii) an ordinal-scaled expert-consensus-based transformation of these scores.
2) To outline the added value of the common metric's ICF basis for the identification of potential further functioning outcome indicators relevant for rehabilitation.

Switzerland was used as a case in point for this study, as both a currently applied ordinal-scaled expert-consensus-based system and a newly developed interval-scaled and ICF-based StARS exist for musculoskeletal and neurological rehabilitation national quality outcome reports.

Methods

Setting, participants and interventions

Secondary analysis of Swiss outcome quality reports in musculoskeletal and neurological rehabilitation, which are coordinated and published by the National Association for Quality Development in Hospitals and Clinics (ANQ) [8], was conducted. From 64 Swiss rehabilitation clinics providing musculoskeletal or neurological rehabilitation in 2016, 29 clinics agreed to provide their ANQ datasets for our study. Ethical approval was received from the Swiss Ethics Committees.

Outcome measures

Sociodemographic, treatment, health status and functioning-related data are routinely assessed for the ANQ quality reports. To assess functioning outcomes in musculoskeletal and neurological rehabilitation, clinics can choose between—FIM^{TM} and EBI [13, 14].

The FIM^{TM} includes 18 items: 13 items related to motor and five to cognitive skills. All items are scored from 1–7 resulting in an ordinal-scaled total score between 18 (total dependence) and 126 (complete independence) [13]. The EBI includes 16 items: 10 motor items based on the Barthel Index [15] and six cognitive items, of which five are derived from the cognitive FIM^{TM} items [14]. All items are scored 0–4, resulting in an ordinal-scaled total score between 0 (total dependence) and 64 (complete independence). While both EBI and FIM^{TM} are administered by health professionals, related training is only mandatory for FIM^{TM}. Recent studies in the context of quality reports showed that both tools can be reported as an interval-scaled total score when Rasch-based transformation is applied [16, 17].

Expert-consensus-based ANQ-ADL score and ICF-based interval-scaled common metric

To enable comparison of all rehabilitation clinics in national reports, irrespective of whether FIM^{TM} or EBI was assessed, two options exist:

(A) The ANQ-ADL score currently used in the ANQ reports, consists of an ordinal-scaled expert-consensus-based transformation algorithm between FIM^{TM} and EBI on item basis [9]. It ranges from 0 (complete dependence) to 60 (complete independence). It allows the comparison at item level but has its limitations: (1) its

Table 1 Main features of the ANQ-ADL score and the ICF-based interval-scaled common metric

	ANQ-ADL score	Common metric
Scale level	Ordinal-scaled	Interval-scaled
Included assessment tools	FIM™ and EBI	FIM™ and EBI
Scale range	0–60	0–100 (adaptable)
Development	Expert-consensus process, validation with a validation sample	Content equivalence assessed with ICF Linking Rules, score equivalence assessed with Rasch-based scale equating approach based on the same validation sample as the ANQ-ADL score
Strengths	– Involvement of experienced clinicians	– Based on the international standard for reporting functioning outcomes (ICF)
	– Item-based approach	– Interval-scale allows for calculations such as means and change scores
		– Considers the operational range of the integrated assessment tools
		– Includes all items of both tools
Weaknesses	– Ordinal-scale does not allow for calculations	– Total score-based approach
	– Does not consider the operational range of the included assessment tools	– More specialized statistical resources (Rasch analysis) required for development
	– Does not include EBI Item 16 'Vision/Neglect', as there is no corresponding item in FIM™	

ANQ-ADL score = Swiss National Association for Quality development in hospitals and clinics Activities of Daily Living Score, ICF = International Classification of Functioning, Disability and Health, FIM™ = Functional independence measure, EBI = Extended Barthel Index.

ordinal-scaling, (2) exclusion of EBI Item 16 'vision/neglect' and (3) automatic match of minimum and maximum scores of the two scales not considering their different operational ranges. The ANQ-ADL score was validated using a representative sample of 265 neurorehabilitation patients all being assessed with both FIM™ and EBI [9].

(B) The newly developed ICF-based interval-scaled common metric [12]. It includes FIM™ and EBI on total score level and was developed on the basis of the ANQ-ADL score validation sample, applying ICF Linking Rules [18] and Rasch methods for scale equating [10]. Its advantages include (1) its interval-scale, needed for calculations currently conducted in ANQ reports, (2) its consideration of the operational range of included tools [12] and (3) its ICF basis (see Appendix A1) allowing to compare the metric's content, e.g. with other tools. The common metric was designed to range from 0 to 100, which can be adjusted, as it is based on logit Rasch values (see Appendix A2).

The main features of these two options are summarized in Table 1.

Data analysis

The examination of the influence of the common metric's interval-scaling feature (specific aim 1) included three steps: (1) examination of the difference between reporting of functioning outcomes with ordinal-scaled FIM™, EBI and respectively ii) ANQ-ADL scores and the interval-scaled common metric; (2) examination of the difference between risk-adjusted funnel-plots of clinic performance based on the ordinal-scaled ANQ-ADL score and the interval-scaled common metric and (3) examination of floor and ceiling effects of ordinal-scaled FIM™, EBI and ANQ-ADL score as well as the interval-scaled common metric. Only those cases that could be clearly assigned to neurological (NEUR) or (MSK) rehabilitation and had complete data for admission and discharge of FIM™ or EBI, as well as the risk-adjustment variables, were included.

In the fourth step (4), we compared the ICF categories covered by the common metric with relevant ICF Core Sets to outline the added value of the common metric's ICF basis (specific aim 2).

The analyses were conducted with RStudio (steps 1–3) and Microsoft Excel (step 4).

1) *Difference between reporting of ordinal-scaled assessment tools respectively ANQ-ADL scores and the interval-scaled common metric*

We created a descriptive table for the comparison of admission, discharge and change scores, i.e. discharge score minus admission score, separately for MSK and NEUR rehabilitation on the clinic level, including respective standard deviations. In order to compare the respective values of ordinal-scaled FIM™, EBI and ANQ-ADL score to the interval-scaled common metric, we adapted the range of the common metric according to the scale it was compared to, i.e. 18–126 for FIM™, 0–64 for EBI and 0–60 for the ANQ-ADL score, on the basis of its Rasch logits.

2) *Difference of risk-adjusted funnel-plots for clinic performance between the ordinal-scaled ANQ-ADL score and the interval-scaled common metric*

We reproduced the funnel-plots of clinic performance from the ANQ reports, once based on the ANQ-ADL scores and once on the common metric for both rehabilitation groups. We used the same risk-adjustment method as ANQ in 2016, i.e. simple linear regression including the discharge ANQ-ADL respectively common metric scores as dependent variable and the following independent variables: gender, age, nationality, residence before admission, residence after discharge, health insurance status and type, diagnosis group, Modified Cumulative Illness Rating Scale (CIRS), duration of rehabilitation and admission ANQ-ADL respectively common metric scores [19]. We then compared the two funnel plots within one rehabilitation group and analyzed which clinics changed in regard

to the three funnel-plot categories (significant upward deviation, no significant deviation and significant downward deviation from regression estimate).

3) *Floor and ceiling effects of ordinal-scaled FIM™, EBI and ANQ-ADL score and the interval-scaled common metric*

As floor and ceiling effects are important quality criteria of outcome measures in health [20], we assessed the percentage of people from each rehabilitation group attaining minimum and maximum scores in FIM™, EBI, ANQ-ADL score and common metric separately for admission and discharge. We defined an indication for floor respectively ceiling effect if > 5% and a clear floor respectively ceiling effect if > 15% reached minimum respectively maximum scores [20].

4) *Added value of the common metric's ICF basis*

The original linking of the items contained in the common metric to the ICF using ICF Linking Rules [12, 18] resulted in 26 covered ICF categories (see Appendix A1). These categories were contrasted to categories of relevant ICF Core Sets in order to define gaps and further development opportunities for the StARS common metric for the ANQ outcome quality reports. ICF Core Sets are purpose-tailored shortlists of ICF categories developed in a standardized multimethod scientific process [21]. There exist two generic ICF Sets, and diagnosis and rehabilitation group-specific sets [22], each with brief and comprehensive versions. We contrasted the common metric's ICF categories with the two generic ICF Sets (Generic-7, Generic-30) [23, 24] and the eight rehabilitation group-specific ICF Core Sets for MSK and NEUR, each in its acute and postacute respectively brief and comprehensive version [25].

Results

Sample characteristics

The overall sample included 18047 complete cases in musculoskeletal (MSK, *n* = 12160) and neurological (NEUR, *n* = 5887) rehabilitation form 26 clinics, of which 18 were located in the German-speaking, five in the French-speaking and three in the Italian-speaking part of Switzerland. Twelve clinics provided both MSK and NEUR rehabilitation, 11 provided only MSK and three only NEUR rehabilitation. Nineteen clinics were assessing FIM™ (*n* = 11 636) and seven were assessing EBI (*n* = 6411). The gender distribution for MSK was 36.7% male (*n* = 4461) and 63.3% female (*n* = 7699), and for NEUR rehabilitation 52.5% male (*n* = 3091) and 47.5% female (*n* = 2796). The mean age of the MSK sample was 69.8 years ranging from 18 to 102. The mean age of the NEUR sample was 64.9 years ranging from 18 to 99. Average rehabilitation duration of MSK patients was 21 days (ranging from 7–182 days) and 37 days for NEUR patients (ranging from 7 to 351 days).

Difference between reporting of functioning outcomes with ordinal-scaled scores and the interval-scaled common metric

Table 2 shows the admission, discharge and change scores on clinic level, separately for MSK and NEUR rehabilitation. In 20 of the 23 MSK rehabilitation clinics, the change scores are higher when the ordinal scales of FIM™, EBI and ANQ-ADL score are used in comparison to the interval-scaled common metric. This was also the case for 14 of the 15 NEUR clinics, indicating a tendency to overestimate outcomes when reported with ordinal-scaled scores. For both rehabilitation groups, the total standard deviation of the different values is smaller for the interval-scaled metric, indicating a greater degree of precision when the common metric is used.

Difference between the risk-adjusted funnel-plots of clinic performance

Figure 1 shows the four funnel-plots comparing risk-adjusted clinic performances when using the ANQ-ADL score and the ICF-based interval-scaled common metric for the two rehabilitation groups. In MSK rehabilitation, five clinics (22%) changed the funnel-plot categories. Two clinics changed from 'no significant deviation' to 'significant upward deviation'. The deviation refers to the regression estimate, which is based on the case-mix related risk-adjustment. So, an upward deviation indicates that clinics performed better than their case-mix-related mean estimation of their performance. One clinic changed from 'significant upward deviation' to 'no significant deviation' and two clinics changed from 'no significant deviation' to 'significant downward deviation'. In NEUR rehabilitation, one clinic (7%) changed from 'significant upward deviation' to 'no significant deviation'.

Floor and ceiling effects

Table 3 shows the results of the analysis for floor and ceiling effects. There was no indication of floor effects for all four scales and also for FIM™ and the common metric no indication for ceiling effects. For EBI, there was a clear ceiling effect for MSK rehabilitation at admission (17.7%) and discharge (21.8%) and an indication of ceiling effect for NEUR rehabilitation at admission (5.1%) and discharge (12.2%). For the ANQ-ADL score, there was an indication for ceiling effect for MSK at admission (7.5%) and discharge for both MSK (12.8%) and NEUR rehabilitation (8.6%). This supports the results from the development of the common metric, which showed that FIM™ has a larger operational range for patients in comparison to EBI (see Appendix A4).

Added value of the common metric's ICF basis

Table 4 shows the overview of the comparison of the common metric's ICF categories with relevant ICF Core Sets. The extensive comparison table on the level of the ICF categories can be found in Appendix A3. The ICF Core Sets are only covered by the common metric with a maximum of 40.0% (Generic-30 Set) and a minimum of 17.2% (ICF NEUR postacute Core Set comprehensive version). The most relevant ICF categories covered by 8 of the 10 analyzed ICF Sets, not present in the common metric were b130 'Energy and drive functions' and d415 'Maintaining basic body position'. The following relevant categories covered by eight or more of the analyzed Core Sets, which were already represented in the common metric, were b620 'Urination functions', d410 'Changing basic body position', d420 'Transferring oneself', d450 'Walking', d510 'Washing oneself', d520 'Caring for body parts', d530 'Toileting' and d550 'Eating', stressing the importance of these aspects.

Discussion

This study demonstrates the influence and added value of an ICF-based interval-scaled StARS for national quality reports, on two levels: (1) the statistical level contrasting the influence of the common metric's interval-scale in comparison to the ordinal-scaled instrument's raw score and an ordinal-scaled expert-consensus-based transformation and (2) the added value on the content level contrasting the

Table 2 Difference between reporting of functioning outcomes ordinal scores and the interval-scaled common metric: musculoskeletal rehabilitation (M) and neurological rehabilitation (N)

Clinic Nr.	Outcome measure (score range)	Admission score ordinal (SD)	Discharge score ordinal (SD)	Change score ordinal (SD)	Admission score metric (SD)	Discharge score metric (SD)	Change score metric (SD)
1 M	ADL Score (0–60)	43.9 (8.6)	50.7 (7.5)	6.8 (5.1)	35.5 (4.8)	39.2 (4.3)	3.7 (2.4)
2M[a]	*ADL Score (0–60)*	*50.4 (7.2)*	*55.3 (6.2)*	*4.9 (4.6)*	*43.1 (7.4)*	*48.1 (7.1)*	*5 (4.6)*
3 M	ADL Score (0–60)	43.9 (8.4)	53 (6.9)	9 (5.8)	35.9 (5)	42.3 (5.1)	6.5 (3.6)
4 M	ADL Score (0–60)	46.8 (10.7)	50.6 (9.3)	3.8 (4.7)	38.6 (7)	41.8 (7)	3.3 (3.4)
5 M	ADL Score (0–60)	41.3 (10.9)	49.9 (10.8)	8.6 (5.9)	34.7 (6.1)	40.5 (7)	5.8 (4)
6 M	ADL Score (0–60)	55.3 (8.9)	58.5 (4.3)	3.2 (6.7)	45.9 (6.4)	47.6 (3.7)	1.7 (3.5)
7 M	ADL Score (0–60)	45.2 (10.3)	51 (9.8)	5.8 (5.8)	37.5 (6.9)	41.8 (7.2)	4.3 (3.9)
8 M	ADL Score (0–60)	41.8 (11.6)	51.3 (10.1)	9.5 (7.3)	35.5 (7.4)	41.8 (7.5)	6.3 (4.2)
9 M	ADL Score (0–60)	35.2 (10.9)	47.3 (10)	12.1 (8.7)	31.1 (5.1)	36.8 (4.7)	5.7 (3.9)
10M[a]	*ADL Score (0–60)*	*51.8 (8.2)*	*54.9 (7.9)*	*3.1 (7)*	*45.1 (9.1)*	*48.9 (9.1)*	*3.8 (6.5)*
11 M	ADL Score (0–60)	50.4 (8.2)	55.2 (5)	4.8 (5.7)	40.8 (54)	44.3 (3.8)	3.5 (3.5)
12 M	ADL Score (0–60)	44.7 (9.6)	53.1 (7.7)	8.4 (6.9)	37.6 (5.7)	43 (5)	5.4 (4.1)
13 M	ADL Score (0–60)	37.3 (13.2)	51 (9.2)	13.6 (9.6)	32.3 (6.2)	39.9 (6.1)	7.6 (4.4)
14 M	ADL Score (0–60)	51.2 (5.5)	57 (3.5)	5.8 (4.3)	40.9 (3.9)	45.5 (3.6)	4.6 (3)
15 M	ADL Score (0–60)	41.6 (9.8)	49.9 (7.8)	8.3 (7.5)	34.4 (5.4)	39.2 (5.3)	4.9 (4.6)
16 M[a]	*ADL Score (0–60)*	*49.9 (4.1)*	*54.9 (3.1)*	*5.1 (2.7)*	*40.3 (3.9)*	*47.1 (5.4)*	*6.8 (4.3)*
17 M	ADL Score (0–60)	50.7 (9.2)	54.6 (6.8)	4 (6.2)	41.2 (6)	43.9 (4.9)	2.6 (3.7)
18 M	ADL Score (0–60)	45.2 (10.5)	49.2 (9.3)	4 (5.1)	37.6 (6.3)	40.1 (5.8)	2.5 (3.1)
19 M	ADL Score (0–60)	41.9 (10)	52.4 (8.3)	10.5 (7.8)	34.7 (4.8)	40.6 (5.5)	5.9 (4.6)
20 M	ADL Score (0–60)	50.4 (7.9)	54.3 (5.8)	3.9 (4.7)	41 (5)	43.5 (4.1)	2.5 (2.7)
21 M	ADL Score (0–60)	45.1 (10.6)	53 (8.5)	8 (7)	36.7 (6.1)	42.4 (6.1)	5.8 (3.8)
22 M	ADL Score (0–60)	47.7 (7.2)	53.9 (5.9)	6.3 (5.3)	37.9 (4.3)	41.7 (3.9)	3.8 (3.1)
23 M	ADL Score (0–60)	42.3 (9.6)	50.5 (8.1)	8.2 (6.8)	35.2 (5.5)	40.4 (5.6)	5.2 (3.8)
Total	**ADL Score (0–60)**	**46.5 (10)**	**53.3 (7.6)**	**6.8 (6.5)**	**38.1 (6.4)**	**42.5 (5.7)**	**4.4 (3.8)**
1 M	FIM™ (18–126)	92.7 (14.1)	102.8 (11.5)	10.1 (7.4)	82 (8.6)	88.6 (7.8)	6.6 (4.3)
2M[a]	*FIM™ (18–126)*	*108.8 (12.6)*	*116.7 (10.8)*	*8 (7.7)*	*95.6 (13.3)*	*104.5 (12.9)*	*9 (8.3)*
3 M	FIM™ (18–126)	93.5 (14.9)	109.2 (11.6)	15.7 (9.7)	82.5 (9)	94.2 (9.2)	11.7 (6.4)
4 M	FIM™ (18–126)	99.3 (18.4)	106.7 (16.1)	7.3 (8.2)	87.4 (12.6)	93.3 (12.7)	5.9 (6.1)
5 M	FIM™ (18–126)	89.9 (18.8)	104.1 (177)	14.2 (9.4)	80.5 (11)	91 (12.5)	10.5 (7.1)
7 M	FIM™ (18–126)	96.7 (18.6)	106.5 (16.9)	9.8 (10.1)	85.4 (12.4)	93.2 (12.9)	7.8 (6.9)
8 M	FIM™ (18–126)	90.8 (20.1)	106.3 (16.8)	15.5 (11.4)	81.9 (13.3)	93.3 (13.5)	11.4 (7.6)
9 M	FIM™ (18–126)	78.7 (16.8)	96.5 (14)	17.8 (12.9)	74 (9.2)	84.2 (8.4)	10.2 (7.1)
10M[a]	*FIM™ (18–126)*	*110.3 (15.3)*	*115.7 (14.6)*	*5.3 (1 1.5)*	*99.1 (16.4)*	*106 (16.4)*	*6.9 (11.8)*
13 M	FIM™ (18–126)	82.2 (20)	103.3 (14.1)	21.1 (13.6)	76.1 (11.1)	89.8 (10.9)	13.7 (7.9)
14 M	FIM™ (18–126)	107 (9.3)	11 5.7 (6.3)	8.7 (6.6)	91.7 (7.1)	99.9 (6.5)	8.2 (5.4)
15 M	FIM™ (18–126)	88.8 (16.1)	102.1 (12.5)	13.4 (12.1)	79.8 (9.6)	88.6 (9.5)	8.7 (8.2)
16 M[a]	*FIM™ (18–126)*	*105.7 (8.3)*	*117.1 (6.8)*	*113 (6.7)*	*90.6 (7)*	*102.8 (97)*	*12.2 (7.7)*
18 M	FIM™ (18–126)	97.5 (17.5)	103.9 (14.6)	6.5 (8.4)	85.7 (11.3)	90.1 (10.4)	4.5 (5.5)
19 M	FIM™ (18–126)	90.3 (14.6)	105.4 (12.7)	15.1 (11)	80.5 (8.6)	91.2 (9.9)	10.6 (8.2)
21 M	FIM™ (18–126)	95.2 (17.7)	108.8 (14.2)	13.5 (10.6)	84 (10.9)	94.4 (11)	10.4 (6.8)
23 M	FIM™ (18–126)	91.6 (15.8)	104.9 (135)	13.3 (10.4)	813 (10)	90.7 (10.1)	9.3 (6.9)
Total	**FIM™ (18–126)**	**95.5 (17.2)**	**107.7 (13.7)**	**12.1 (10.2)**	**84.4 (1 1.3)**	**93.5 (11)**	**9.1 (7)**
6 M	EBI (0–64)	59.2 (9.1)	62.3 (4.5)	3 (6.3)	59.4 (8.4)	61.6 (4.9)	2.2 (4.6)
11 M	EBI (0–64)	53.4 (8.9)	58.8 (5.4)	5.4 (6.1)	52.7 (7)	57.3 (5)	4.6 (4.5)
12 M	EBI (0–64)	48.2 (9.7)	56.9 (7.9)	8.7 (7)	48.5 (7.4)	55.6 (6.6)	7.1 (5.4)
17 M	EBI (0–64)	54 (9.4)	58 (7.2)	4 (5.9)	53.2 (7.9)	56.7 (6.5)	3.5 (4.8)
20 M	EBI (0–64)	54.1 (7.8)	58 (5.8)	4 (4.6)	52.9 (6.6)	56.2 (5.4)	3.3 (3.5)
22 M	EBI (0–64)	49.6 (7.4)	55.8 (6.2)	6.2 (5.1)	48.9 (5.6)	53.8 (5.1)	4.9 (4)
Total	**EBI (0–64)**	**52.8 (9.1)**	**58 (6.4)**	**5.2 (5.8)**	**52.2 (7.8)**	**56.4 (5.9)**	**4.2 (4.4)**
1 N	ADL Score (0–60)	46.2 (13)	49.2 (11.7)	3 (5.9)	37.3 (7.7)	39.5 (7.4)	2.2 (3.3)
2 N	ADL Score (0–60)	41.4 (16.7)	45.6 (15.4)	4.3 (7.6)	35.4 (11.6)	38 (10.4)	2.6 (6.3)
3 N	ADL Score (0–60)	31.8 (16)	42.2 (15.9)	10.3 (9.6)	29.5 (9.8)	35.7 (9.9)	6.2 (5.5)
4 N	ADL Score (0–60)	45.7 (14.5)	51.4 (11.2)	5.7 (9.3)	37.7 (9.3)	41.3 (7.4)	3.5 (6)
5 N	ADL Score (0–60)	31.5 (17.4)	42.4 (16.9)	10.9 (9.2)	28.6 (1 0.8)	34.9 (9.9)	6.3 (5.1)
[a]*6N*	*ADL Score (0–60)*	*44.9 (15.5)*	*47.5 (16)*	*2.6 (6)*	*39.7 (11.8)*	*43.5 (14.6)*	*3.8 (6.6)*
7 N	ADL Score (0–60)	41.2 (14.9)	47.9 (13.1)	6.6 (8.6)	35.4 (9.2)	39.8 (8.3)	4.4 (5.3)

Continued

Table 2 Continued

Clinic Nr.	Outcome measure (score range)	Admission score ordinal (SD)	Discharge score ordinal (SD)	Change score ordinal (SD)	Admission score metric (SD)	Discharge score metric (SD)	Change score metric (SD)
8 N	ADL Score (0–60)	33.2 (16.1)	46.2 (15.1)	13 (11.6)	30.1 (8.9)	37.8 (9.6)	7.7 (6.2)
9 N	ADL Score (0–60)	30.7 (15.2)	40.6 (16.3)	9.9 (10.2)	28.6 (8.7)	34.3 (9.6)	5.7 (6.4)
10 N	ADL Score (0–60)	42 (14)	44.6 (14.1)	2.6 (5.6)	35.3 (8.5)	37.1 (8.5)	1.9 (3.2)
11 N	ADL Score (0–60)	24.1 (19.2)	33.3 (20.2)	9.2 (13.6)	23.1 (1 3.9)	29.8 (13.8)	6.7 (9.1)
12 N	ADL Score (0–60)	40.9 (15.6)	48.3 (13.4)	7.5 (9.7)	35.5 (9.2)	40.2 (8.2)	4.7 (5.5)
13 N	ADL Score (0–60)	30.4 (14.5)	40.7 (12.2)	10.2 (7.2)	28.8 (9)	34.2 (5.6)	5.4 (5)
14 N	ADL Score (0–60)	31.4 (14.7)	43.8 (15.2)	12.4 (1 0.4)	29.3 (8.4)	36.2 (8.9)	6.8 (5.5)
15 N	ADL Score (0–60)	37.3 (16.3)	44.6 (15.2)	7.2 (10.9)	32.3 (9)	36.9 (8.7)	4.6 (6.5)
Total	**ADL Score (0–60)**	**37.7 (1 6.4)**	**45.6 (14.9)**	**7.9 (10)**	**32.9 (9.9)**	**37.8 (9.3)**	**4.9 (5.9)**
1 N	FIM (18–126)	96.2 (21.6)	101.5 (19.4)	5.3 (9)	85.2 (13.8)	89.1 (13.3)	3.9 (6)
3 N	FIM (18–126)	73.9 (27.3)	90.9 (26.5)	17 (15.3)	71.1 (17.6)	82.2 (17.9)	11.2 (9.9)
5 N	FIM (18–126)	73 (28.5)	89.9 (26.2)	16.9 (14.4)	69.6 (19.5)	80.8 (17.8)	11.3 (9.2)
[a]*6N*	*FIM (18–126)*	*97.5 (28.8)*	*102.9 (29.1)*	*5.4 (9.6)*	*89.4 (21.2)*	*96.3 (26.2)*	*6.8 (11.9)*
8 N	FIM (18–126)	75.8 (25.4)	96.7 (24.2)	20.8 (17.4)	72.1 (16)	86 (17.3)	13.9 (11.2)
9 N	FIM (18–126)	71.4 (25)	87.3 (26.2)	15.8 (16.8)	69.5 (15.7)	79.7 (17.4)	10.2 (11.5)
10 N	FIM (18–126)	91 (23.6)	95.6 (23.1)	4.6 (8.9)	81.5 (15.3)	84.8 (15.4)	3.4 (5.8)
11 N	FIM (18–126)	59.7 (33.2)	75.9 (34.8)	16.2 (23.1)	59.6 (25.1)	71.7 (24.9)	12.1 (16.4)
13 N	FIM (18–126)	72.6 (24.8)	88.2 (17.9)	15.6 (123)	69.8 (16.2)	79.5 (10)	9.7 (9)
14 N	FIM (18–126)	73.5 (24.3)	92.7 (24.6)	19.2 (15.8)	708 (15.1)	83.1 (16)	12.3 (9.9)
15 N	FIM (18–126)	82.5 (25.5)	94.4 (23.3)	11.9 (17.3)	76.1 (16.2)	84.5 (15.7)	8.4 (11.7)
Total	**FIM (18–126)**	**76.4 (27.3)**	**92 (25.9)**	**15.5 (17.1)**	**72.3 (17.9)**	**82.8 (17.8)**	**10.5 (11.4)**
2 N	EBI (0–64)	44.4 (17.5)	48.6 (16.1)	4.3 (8.2)	45.6 (15.3)	49.1 (13.6)	3.4 (8.2)
4 N	EBI (0–64)	48 (15.4)	53.7 (12.1)	5.7 (10.2)	48.7 (12.2)	53.3 (9.7)	4.6 (7.9)
7 N	EBI (0–64)	44.2 (15.4)	51.4 (13.7)	7.2 (8.9)	45.5 (12.1)	51.3 (10.9)	5.8 (6.9)
12 N	EBI (0–64)	44.7 (15.3)	52.2 (13.2)	7.5 (9.4)	45.7 (12)	51.9 (10.7)	6.1 (7.2)
Total	**EBI (0–64)**	**44.6 (15.4)**	**51.8 (13.5)**	**7.2 (9.2)**	**45.8 (12.2)**	**51.6 (10.8)**	**5.8 (7.2)**

ADL Score = Activities of Daily Living Score, FIM™ = Functional Independence Measure, EBI = Extended Barthel Index, N = N.

[a]Clinics represented in italics show a higher change score when the interval-scaled metric is applied in comparison the majority of clinics, which show a smaller change score when the interval-scaled metric is applied.

Table 3 Examination of floor and ceiling effects

Scale and rehabilitation group	% of people reaching minimum score at admission (*N*)	% of people reaching maximum score at admission (*N*)	% of people reaching minimum score at discharge (*N*)	% of people reaching maximum score at discharge (*N*)
FIM™ MSK	0.0 (1)	0.4 (35)	0.0 (0)	1.3 (103)
FIM™ NEUR	2.1 (76)	0.2 (8)	0.8 (31)	1.1 (39)
EBI MSK	0.0 (1)	17[b] (741)	0.0 (0)	21.8[b] (911)
EBI NEUR	0.1 (2)	5.1[a] (114)	0.0 (1)	12.2[a] (273)
ADL Score MSK	0.0 (2)	7.5[a] (917)	0.0 (0)	12.8[a] (1553)
ADL Score NEUR	1.6 (92)	3.1 (182)	0.6 (35)	8.6[a] (508)
Common metric MSK	0.0 (1)	0.3 (35)	0.0 (0)	0.8 (103)
Common metric NEUR	1.3 (76)	0.1 (6)	0.5 (31)	0.6 (34)

MSK = Musculoskeletal rehabilitation, NEUR = Neurological rehabilitation, ADL Score = Activities of Daily Living Score, FIM™ = Functional Independence Measure, EBI = Extended Barthel Index.

[a]Indication of ceiling effect (>5%).

[b]Clear ceiling effect (>15%).

common metric's functioning categories with the content of relevant ICF Core Sets.

When the interval-scaled common metric is applied and contrasted to the currently used ordinal-scaled functioning outcomes, change scores on the clinic level tended to be smaller on the common metric but more precisely estimated. The main reason for this is that the units in the ordinal scale are not equal and tend to be smaller in the center of a scale than at the margins [7]. Consequently, when contrasted with the common metric, patients passing over the center of the ordinal scale pick up raw score points quickly, whereas the opposite is true for those moving across the margins. The metric removes this bias and provides a more accurate estimation of the actually achieved change. Even though it is known that ordinal-level scales lead to over- or underestimation of health-related outcomes [6, 7, 11, 26], many comparable outcome reports do use ordinal-scaled data without considering this fallacy [27, 28]. The results of

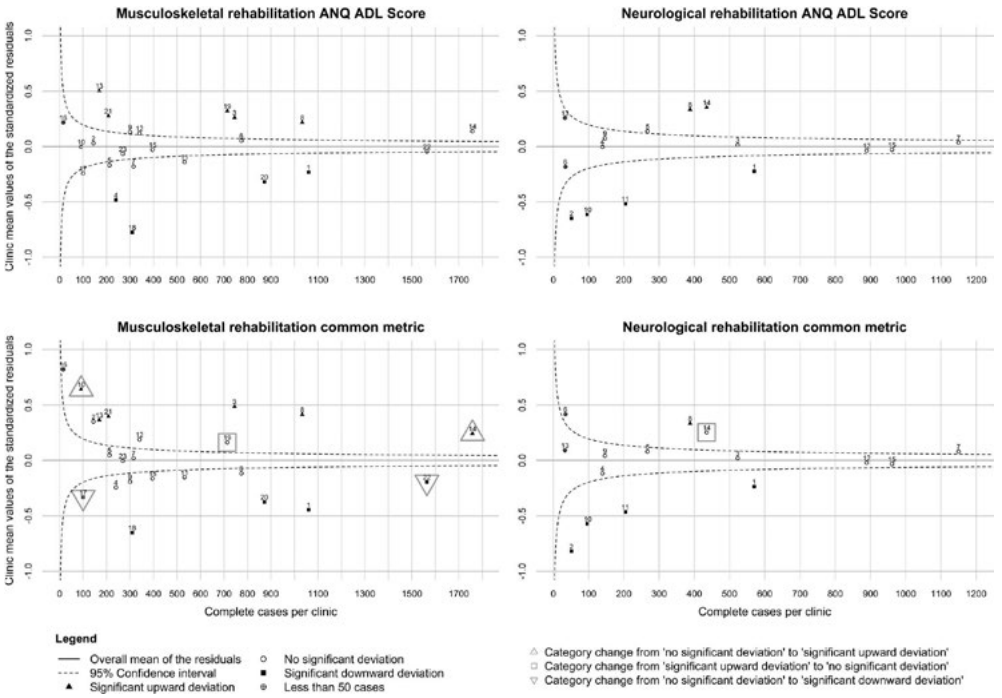


Figure 1 Funnel-plots comparing risk-adjusted clinic performance when using the ANQ-ADL score or the ICF-based interval-scaled common metric. ANQ = Swiss National Association for Quality Development in Hospitals and Clinics, ADL Score = Activities of Daily Living Score, ICF = International Classification of Functioning, Disability and Health.

the present study suggest that the biased ordinal-scaled reporting potentially leads to erroneous clinical decision-making and unfair benchmarking of clinic performance. No statistical hypothesis involving ordinal-scaled data should be tested before the ordinal-scaled data are transformed onto interval-scale level [29].

In the current study, MSK rehabilitation was affected more than NEUR rehabilitation by the difference between ordinal- and interval-scaled reporting approaches, no matter if risk-adjustment was conducted or not. The MSK clinics had baseline scores closer to the scales' upper limits in comparison to the NEUR clinics with scores located more around the center of the scales, indicating the importance that not only change scores on its own, but mean admission and discharge scores should be reported [11]. Furthermore, the MSK sample also showed stronger ceiling effects, especially with EBI, reflecting that FIM^{TM} and EBI are discriminating for the population they were developed for, i.e. the FIM^{TM} [13] for generic rehabilitation and EBI for NEUR rehabilitation [14]. The information of floor and ceiling effects can inform the clinics' decision for a tool most suitable for their specific patient population.

The common metric's ICF basis allowed the comparison with relevant ICF Core Sets, showing potential development opportunities in the functioning outcome indicator included in the current reports in NEUR and MSK rehabilitation such as 'energy and drive functions' but also confirmed relevant functioning outcome aspects that are already represented.

A StARS, with the common metric as the core element, can also be applied for other contexts outside of quality reports such as the comparison of outcome measures in meta-analyses [29]. In any case, a StARS has to be developed for its purpose [30], and it makes sense to consider its influence and added value before its actual implementation.

A limitation of the current study is that the analysis of the influence and added value is at the level of rehabilitation groups. As such, it would be interesting to consider the influence on a more detailed level such as diagnosis-related groups, for example stroke in neurological rehabilitation. A further limitation is that the study is based on a descriptive approach, which helps to describe the differences between the two reporting approaches but does not allow to make statements whether the discovered difference of the common metric is significant or not.

Conclusions

This study shows that it matters if functioning outcomes are reported on ordinal- or interval-scale level. A StARS can help to incorporate several conceptually similar assessment tools into one interval-scaled reporting system, thus enabling the comparison across clinics using different tools, as well as the calculations of means and change scores. Furthermore, the ICF basis of the common metric serves as an opportunity to inform further development of internationally

Table 4 Overview of the comparison of the common metric's ICF categories to relevant ICF Core Sets

ICF Core Set	Generic-7	Generic-30	MSK acute brief	MSK acute comp.	MSK post acute brief	MSK post acute comp.	NEUR acute brief	NEUR acute comp.	NEUR post acute brief	NEUR post acute comp.
Number of ICF categories in Set	7	30	27	48	31	70	33	85	38	116
Coverage of ICF categories with common metric	2 of 7	12 of 30	9 of 27	9 of 48	9 of 31	13 of 70	10 of 33	16 of 85	12 of 38	20 of 116
% of coverage with common metric	28.60%	40.0%	33.3%	18.8%	29.0%	18.6%	30.3%	18.8%	31.6%	17.2%

ICF = International Classification of Functioning, Disability and Health, MSK = Musculoskeletal, NEUR = Neurological, comp. = Comprehensive.

relevant functioning outcome indicators in rehabilitation quality reports.

Supplementary material

Supplementary material is available at *INTQHC Journal* online.

Acknowledgements

Special thanks go to the 29 NFP74 StARS clinics for providing us with their ANQ data aarReha Schinznach; Berner Klinik Montana; Berner Reha Zentrum Heiligenschwendi; Cereneo Schweiz AG; Clinica Hildebrand Centro di Riabilitazione Brissago; Clinique de Crans-Montana-HUG; Clinique La Lignière; Clinique romande de réadaptation Suva; EOC Clinica di Riabilitazione Faido & Novaggio; HVS sites Brig, Martigny, Sierre & Saint Amé; Kantonsspital Baselland Standorte Bruderholz & Laufen; Klinik Schloss Mammern; Klinik Schönberg; Kliniken Valens-Rehazentrum Valens, Rehazentrum Walenstadtberg und Rheinburg-Klinik Walzenhausen; Luzerner Höhenklinik Montana; Reha Rheinfelden; REHAB Basel; Rehaklinik Dussnang AG; Rehaklinik Zihlschlacht AG; Spitäler Schaffhausen; Universitäre Altersmedizin FELIX PLATTER; Universitätsklinik Balgrist; Zürcher RehaZentren Klinik Wald and to Dr. phil. Luise Menzi Head of Rehabilitation ANQ and Dr. med. Anke Scheel-Sailer, Head of Clinical Quality Management Research Department from the Swiss Paraplegic Centre for their provision of valuable advice for the project. This project is part of the cumulative Dissertation of Roxanne Maritz, which is funded by the Swiss National Science Foundation within the NRP74 StARS project.

Funding

This work was funded by the Swiss National Science Foundation's National Research Programme 'Smarter Health Care' (NRP 74) [grant number: 407440_167412/1].

Declaration of no objection

The authors have no conflicts of interests to declare.

References

1. Shaw C, *WHO Regional Office for Europe* (Health Evidence Network), Copenhagen, Denmark. http://www.euro.who.int/document/e82975.pdf (28 February 2020, date last accessed)
2. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. *Eur J Phys Rehabil Med* 2017;**53**:134–8.
3. National Academies of Sciences - Engineering - Medicine. *Crossing the global quality chasm: Improving health care worldwide*. Washington (DC): National Academies Press, 2018.
4. World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. Geneva: WHO, 2001.
5. Stucki G, Bickenbach J, Melvin J. Strengthening rehabilitation in health systems worldwide by integrating information on functioning in National Health Information Systems. *Am J Phys Med Rehabil* 2017;**96**:677–81.
6. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. *Eur J Phys Rehabil Med* 2018;**54**:110–7.
7. Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res* 2011;**11**:571–85.
8. *ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken*, Bern, Switzerland: https://www.anq.ch/en/departments/rehabilitation/ (28 February 2020, date last accessed)
9. *ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken*, Bern, Switzerland: https://www.anq.ch/de/rehabilitation-neuer-adl-score-macht-fim-und-ebi-resultate-vergleichbar/ (in German) (28 February 2020, date last accessed)

10. Andrich D. The polytomous rasch model and the equating of two instruments. In: KBK C, S. Mesbah M (eds.). *Rasch Models in Health*. Hoboken: ISTE Ltd and John Wiley and Sons, Icn, 2013, 163–96.
11. Stucki G, Daltroy L, Katz JN *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* 1996;**49**:711–7.
12. Maritz R, Tennant A, Fellinghauer C *et al.* Creating a common metric for activities of daily living scores to enable standardized reporting of functioning outcomes achieved during rehabilitation. *J Rehabil Med* 2020, **52**: Epub ahead of print.
13. Keith RA, Granger CV, Hamilton BB *et al.* The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil* 1987;**1**:6–18.
14. Prosiegel M, Böttger S, Schenk T *et al.* The extended Barthel index - a new scale for the assessment of disability in neurological patients [German]. *Neurol Rehabil* 1996;1:7–13.
15. Mahoney FI, Barthel DW. Functional evaluation: the Barthel index. *Maryland State Med J* 1965;**14**:61–5.
16. Maritz R, Tennant A, Fellinghauer C *et al.* The functional independence measure 18-item version can be reported as a unidimensional interval-scaled metric: internal construct validity revisited. *J Rehabil Med* 2019;**51**:193–200.
17. Maritz R, Tennant A, Fellinghauer CS *et al.* The extended Barthel index (EBI) can be reported as a unidimensional interval-scaled metric – a psychometric study. *Phys Med Rehab Kuror* 2019;**29**:224–32.
18. Cieza A, Fayed N, Bickenbach J *et al.* Refinements of the ICF linking rules to strengthen their potential for establishing comparability of health information. *Disabil Rehabil* 2019;**41**:574–83.
19. *ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken*, Bern, Switzerland: https://results.anq.ch/fileadmin/documents/anq/27_31/20180913_ANQ_Reha_Nationaler-Vergleichsbericht_Muskuloskelettale-Rehabilitation_2016.pdf https://results.anq.ch/fileadmin/documents/anq/27_31/20180913_ANQ_Reha_Nationaler-Vergleichsbericht_Neurologische-Rehabilitation_2016.pdf (in German) (28 February 2020, date last accessed)
20. Terwee CB, Bot SD, de Boer MR *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42.
21. Selb M, Escorpizo R, Kostanjsek N *et al.* A guide on how to develop an international classification of functioning, disability and health Core set. *Eur J Phys Rehabil Med* 2015;**51**:105–17.
22. ICF Research Branch, ICF Research Branch, Nottwil, Switzerland: https://www.icf-research-branch.org/icf-core-sets (28 February 2020, date last accessed)
23. Prodinger B, Cieza A, Oberhauser C *et al.* Toward the international classification of functioning, disability and health (ICF) rehabilitation set: a minimal generic set of domains for rehabilitation as a health strategy. *Arch Phys Med Rehabil* 2016;**97**:875–84.
24. Cieza A, Oberhauser C, Bickenbach J *et al.* Towards a minimal generic set of domains of functioning and health. *BMC Public Health* 2014;**14**:218.
25. Grill E, Strobl R, Muller M *et al.* ICF Core sets for early post-acute rehabilitation facilities. *J Rehabil Med* 2011;**43**:131–8.
26. Ørnbjerg LM, Christensen KB, Tennant A *et al.* Validation and assessment of minimally clinically important difference of the unadjusted health assessment questionnaire in a Danish cohort: uncovering ordinal bias. *Scand J Rheumatol* 2020;**49**:1–7.
27. AROC, Australasian Rehabilitation Outcomes Centre, Wollongong, Australia: https://ahsri.uow.edu.au/content/groups/public/@web/@chsd/@aroc/documents/doc/uow256897.pdf 28 February 2020, date last accessed.
28. *Canadian Institute for Health Information*, Ottawa, Canada: https://www.cihi.ca/en/nrs_pia_en.pdf 28 February 2020, date last accessed.
29. Stucki G, Pollock A, Engkasan JP *et al.* How to use the international classification of functioning, disability and health as a reference system for comparative evaluation and standardized reporting of rehabilitation interventions. *Eur J Phys Rehabil Med* 2019;**55**:384–94.
30. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the international classification of functioning, disability and health. *Eur J Phys Rehabil Med* 2017;**53**:144–9.

Chapter 6

Stakeholder involvement

Stakeholder involvement

This chapter gives an overview of the stakeholder involvement activities of the NRP74 StARS research project, with a focus on two core activities: the stakeholder brief and stakeholder dialogue. This chapter is directly related to specific aim 4 of the present thesis, i.e. to develop strategies with relevant stakeholders for implementing the StARS in Swiss national rehabilitation reviews.

6.1 Background

To help bridge research recommendations with regard to the development of a StARS for functioning outcomes and practice in Swiss rehabilitation quality reviews of using different assessment tools to assess functioning outcomes [1–7], the NRP74 StARS project highlighted the development of implementation strategies through stakeholder involvement and knowledge translation [8–10]. The NRP74 StARS research project integrated stakeholder involvement activities throughout its entire project span, such as stakeholder consultation for grant submission, a kick-off meeting with project partners, the formation of an advisory board, two advisory board meetings, a stakeholder brief and stakeholder dialogue [11]. The activities, their related goals and timelines are presented in detail in Table 1 at the end of this chapter. The focus of this chapter is on the two core activities related to the thesis' objective: the stakeholder brief and the stakeholder dialogue. Both of these are recommended tools for enabling the uptake of research evidence in support of quality improvement, in the sense of a learning health system [12–15]. After the completion of the NRP74 StARS project in spring 2021, an overarching project report describing all stakeholder involvement activities in detail will be published on the Lucerne Open Repository (LORY).

6.2 Stakeholder brief

Stakeholder brief methods

Policy or stakeholder briefs are documents that summarize research evidence, along with other forms of evidence and material, to provide context specific information on a priority issue and possible options to address this issue [15]. These briefs are increasingly used to inform stakeholder dialogues that involve stakeholders who would be affected by decisions regarding the respective priority issue [12–15]. In this thesis, we involved not only policy level stakeholders but also stakeholders from other levels of the health care system. Thus, we used the term "stakeholder brief" instead of "policy brief" for the brief developed for the stakeholder dialogue.

The objective of the stakeholder brief was to provide a short preparatory document that informed all stakeholder dialogue participants about the dialogue's content, background, goals and processes and about the related research project and its findings in a user-friendly language.

In response to the feedback given during discussions between the research team and the NRP74 StARS project advisory board, the research team decided to not only

present in the stakeholder brief one but rather in total four application areas for the results of the studies:

1) a StARS for functioning outcomes for national quality reviews in rehabilitation.
2) a StARS for general assessment and reporting of functioning outcomes in the clinical context of rehabilitation.
3) a StARS for functioning outcomes for the planning and performance mandates for Swiss cantons.
4) a StARS for functioning outcomes for national health statistics.

The reasoning for this decision was two-fold. First, these four application areas were all identified to be highly correlated. Secondly, addressing all four application areas made the best use of the time and interest expressed by the participating stakeholders, who came together for the one-day stakeholder dialogue.
Two methods were used to develop the draft of the stakeholder brief. First, Strengths-Weaknesses-Opportunities-Threats (SWOT) analyses, one developed for each of the four application areas, were used to systematically inform the stakeholder dialogue discussions regarding a StARS for functioning outcomes in each application area. SWOT analysis was chosen, as it consists of a simple framework for identifying and analysing factors that can have an impact on the implementation of a project or a product, and can provide a source of information for decision making. Furthermore, SWOT analyses enable health professionals and other stakeholders to participate more fully in facilitating the implementation of health care improvement activities [16].
Second, key informant interviews were used to ensure that the content of the stakeholder brief and specifically the description of the application areas and the SWOT analyses included in the brief contained all necessary information in a comprehensive way. Nine key informant interviews with stakeholders representing the four application areas were conducted by the research team. Each 1.5-hour interview followed a semi-structured interview guide, which was adapted according to the respective application area. The interviews were audio recorded and transcribed in the form of interview minutes, which were shared with and checked by the respective interviewees. The content of the interviews and the interviewee feedback were used to complete the final version of the stakeholder brief and to inform the planning stakeholder dialogue discussions, specifically relevant discussion issues to be considered.
The stakeholder brief was written in German, as this was the language of the stakeholder dialogue and was mailed four weeks in advance of the dialogue to all participants. In this e-mail, the stakeholders were asked to read the stakeholder brief in its entirety, to judge if they agree with the points of the SWOT analyses (yes/no, comments/additions) and to think about important next steps for the implementation of a StARS for functioning outcomes in the four application areas.

Results of the stakeholder brief

The following sections were included in the stakeholder brief:

- An introductory summary about the purpose of the stakeholder brief, the stakeholder dialogue and the related tasks of the stakeholders.
- An abstract that included a description of the underlying problem, the proposed solution i.e. a StARS for functioning outcomes, the research project, the four application areas and the goal of the stakeholder dialogue.
- A description of the NRP74 StARS research project including its background, its objectives, its methods and the results of the studies conducted.
- A description of the four application areas of a StARS for functioning outcomes in the Swiss health care context.
- SWOT analyses for a StARS for functioning outcomes in each of the four application areas.
- Information about the process and organisation of the stakeholder dialogue, including an agenda.
- A participant list for the stakeholder dialogue, acknowledgements, references, list of abbreviations and appendices, including an overview of additional sources of information considered for the stakeholder brief (such as the key informant interviews) and an informed consent form that allowed audio recording and pictures to be taken during the stakeholder dialogue.

The following parts of the stakeholder brief were translated into English, and can be found in the supplementary materials of this thesis: The abstract (see Appendix 2), the participant list (see Appendix 3), the agenda of the stakeholder dialogue (see Appendix 4) and the SWOT analysis for application area 1) a StARS for functioning outcomes for Swiss national quality reviews in rehabilitation (see Appendix 5).

6.3 Stakeholder dialogue

Stakeholder dialogue methods

A stakeholder dialogue is an evolving method that aims to support and engage stakeholders in implementing evidence-based solutions for a particular issue [13, 14, 17, 18]. Stakeholder dialogues are a promising mechanism for information sharing of research findings, networking, discussion, consensus development and planning of actions about related goals and/or processes, including policy, service delivery, financing and health information collection [13, 17]. Stakeholder dialogues should be designed to support interactions between researchers and stakeholders, timely identification and interpretation of the available research evidence and a timely identification of accord between research evidence and the beliefs, values, interests, goals and strategies of policymakers and other stakeholders [13]. In addition, stakeholder dialogues promote the identification of research issues that match actual needs of the health care system; this, in turn, fosters a culture of shared responsibility [17]. During a stakeholder dialogue, a group of stakeholders works collaboratively towards a common understanding of a problem and the best course of action to address this problem, taking a policy or stakeholder brief as a basis for their discussions [13, 14,

17]. Stakeholder dialogues are guided by a neutral moderator, who facilitates the discussions, stimulates participants to confront their views and manages confrontation over differences of opinion with fairness and assurance that all opinions are heard [17, 18].

The stakeholder dialogue conducted as part of the present thesis aimed to inform relevant stakeholders about the research project and to develop strategies for implementing a StARS for functioning outcomes in Swiss national rehabilitation quality improvement. Furthermore, strategies for developing and implementing a StARS for functioning outcomes in three other application areas, i.e. in the clinical context of rehabilitation, in planning and performance mandates of Swiss cantons and in national health statistics, were discussed but are not further reported in detail in the present thesis. The participants of the stakeholder dialogues were recruited through the NRP74 StARS project team and the projects' advisory board. The stakeholder dialogue was conducted according to the Chatham House Rules, meaning that the identities of the participants or the individual contributions are not reported in detail to promote an honest and productive dialogue [19]. The dialogue was moderated by a neutral moderator with expertise in health care and communication sciences, who was not a member of the NRP74 StARS research team [18].

The stakeholder dialogue included presentations and structured discussions that engaged all the stakeholders, partially organized in subgroups, about the implementation of a StARS for each of the four application areas. A panel discussion outlined the project's background, research results and the four application areas. The structured discussions were based on the information presented in the stakeholder brief, notably the SWOT analyses. A detailed agenda of the event is presented in Appendix 4.

A report of the stakeholder dialogue, based on the notes of the research team and audio recording of the stakeholder dialogue discussions was sent out to all of the stakeholder dialogue participants. The report also presented the output of the stakeholder dialogue, i.e. the implementation agenda. All participants had the opportunity to give feedback on the report. The feedback was then used to finalize the report.

Results of the stakeholder dialogue

The one-day stakeholder dialogue took place in November 2019 at the University of Lucerne in Switzerland. Out of 32 identified stakeholders, 24 stakeholders participated in the dialogue and represented: the federal offices of public health (n=1) and statistics (n=2), health care departments of cantons (n=1), patients (n=2), rehabilitation clinics and professionals (n=4), quality management organizations including the ANQ (n=3), institutions responsible for rehabilitation financing such as insurances and tariff commissions (n=5), rehabilitation associations (n=5) and research (n=1). See Appendix 3 for a detailed participant list.

The output of the discussions was an implementation agenda for the four application areas. This agenda included the following points in regard to application area 1) a StARS for national rehabilitation quality reviews:

- *Proposal of a StARS for functioning outcomes for all ANQ rehabilitation areas:* The NRP74 StARS project team will submit to the ANQ Quality Committee for Rehabilitation a proposal detailing the development and application of a StARS for functioning outcomes including all rehabilitation areas represented in ANQ reviews.
- *Processes to be followed within the ANQ for the potential implementation of a StARS for functioning outcomes:* The ANQ Quality Committee for Rehabilitation will subsequently discuss the proposal and make recommendations to the ANQ Board, the decision making body who will determine if the StARS can be implemented.
- *Clarification of opportunities related to the new Swiss national quality commission and agreements:* The NRP74 StARS project team will contact the responsible persons at the federal office of public health and the partners of the national quality agreements in order to clarify the possibilities and financing opportunities related to a StARS for functioning outcomes associated with the renewed national quality agreements and the newly established national quality commission (to be launched in 2021).
- *Creation of a StARS consortium for further collaboration and exchange:* All but one of the stakeholder dialogue participants expressed their commitment to establishing a consortium for further exchange and coordination between the stakeholders. The first task of this consortium should be the development of a roadmap that defines concrete responsibilities and clarifies the financing aspects of the next steps. The participants agreed that the Swiss Learning Health System (SLHS) initiative should be considered as coordinator of this consortium. The NRP74 StARS team will get in contact with the respective SLHS board regarding this.

Through the implementation agenda, next steps concerning the ANQ and important national quality initiatives were clarified and agreed upon. The stakeholders showed their commitment to further collaboration through the formation of a consortium.

6.4 Evaluation of the stakeholder brief and stakeholder dialogue

Evaluation method

The evaluation of stakeholder briefs and stakeholder dialogues is needed to improve our understanding of which particular design aspects and follow-up activities are well-received from the participating stakeholders, as well as our understanding of whether and how stakeholder dialogues support evidence-informed policy and decision making [13].

Both the stakeholder brief and stakeholder dialogue were evaluated in an online survey. The online survey was carefully designed and pre-tested by the research team according to specific principles [20]. The anonymized online survey was conducted in German using the survey tool SurveyMonkey® and included 16 questions including 5 point Likert scale answers and free text answers. The invitation to the survey was sent out via e-mail in the week following the stakeholder dialogue. A reminder to participate in the survey was sent out one week later.

Results of the evaluation

Of the 24 stakeholders who attended the dialogue, 19 (79%) replied to the survey. All survey participants rated their experiences with the stakeholder dialogue as good (53%) or very good (47%). The approach with the stakeholder brief as a preparatory document was rated by 53% as rather helpful and by 37% as very helpful and was largely read in its entirety (84%). According to the survey participants' free text comments, the stakeholder brief provided a good and understandable overview and introduction to the NRP74 StARS project and to its application areas. However, parts of the brief could have been more concise and the task regarding to the SWOT analyses could have been presented more clearly. The composition of the stakeholders was perceived as good and considered as appropriate. While three potentially missing stakeholders (the Association of Swiss Rehabilitation Clinics, medical officers from health insurance companies and representatives from large rehabilitation clinics) were identified, it was also reported that the group size was rather large. From the point of view of the participants, the objective of the dialogue was mainly fully (22%) or rather (56%) achieved. The participants found the dialogue a valuable opportunity to gain insight in the NRP74 StARS project and its application areas and to actively contribute to the discussions and in determining further steps. The participants rated the organization before and during the stakeholder dialogue as good (28%) or very good (72%). Moreover, the duration of the event was rated as appropriate by 89% of the participants, while 11% felt it that it was rather long. Lastly, the statements in the general comment section indicated that the opportunity for exchange and discussion with the other stakeholders was highly valued and the day was described as meaningful, stimulating and interesting.

Table 1 Overview of the stakeholder involvement activities within the NRP74 StARS research project

Stakeholder group	Stakeholder involvement activities and related goals
PROJECT PARTNERS 12 Swiss rehabilitation clinics, the ANQ and the ICF Research Branch	Acquisition and consultation of initial project partners for grant submission (2016) Goals: To gain initial project partners who support the research project. To get feedback for the research project grant submission. Kick-off meeting (April 2017) Goal: To gain preliminary feedback of on the first steps of the research project.
ADVISORY BOARD 19 members representing Government, cantons, financing, patients, rehabilitation service provider, rehabilitation associations including relevant health care professionals and health care quality management experts	Creation of the NRP74 StARS advisory board (2017) Goal: To identify suitable members and create an advisory board for feedback and exchange regarding the overall NRP74 StARS research project. NRP74 StARS advisory board meeting 1 (June 2019) Goals: To present and get feedback on preliminary project results. To ensure that the project is responding to and aligned with local, cantonal and federal initiatives and practices. To identify stakeholders for the stakeholder dialogue. NRP74 StARS advisory board meeting 2 (November 2020) Goals: To present final project results. To get feedback on the dissemination of the results. To define further steps and involvement regarding follow-up research and implementation activities.
NRP74 StARS KEY STAKEHOLDERS 24 stakeholders representing: federal offices of public health and statistics; cantonal health care departments; rehabilitation patients, clinics, professionals, associations, quality management, financing and research	Stakeholder dialogue including a stakeholder brief (November 2019) Goal of the stakeholder brief: To inform all stakeholder dialogue participants about its content, background, goals, processes and related research findings. Goals of stakeholder dialogue: To inform relevant stakeholders about the research project. To develop strategies for implementing a StARS for functioning outcomes in Swiss national rehabilitation quality improvement.
FURTHER STAKEHOLDERS AND THE PUBLIC	NFP74 StARS Symposium (September 2020)* Goals: To publicly present the final research project results. To discuss possible follow-up research activities and potential collaboration.

*Cancelled due to COVID-19, alternative measures were taken to meaningfully replace this stakeholder involvement activity

References Chapter 6

1. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. European journal of physical and rehabilitation medicine 2018; 54: 110-7.
2. Stucki G, Pollock A, Engkasan JP, Selb M. How to use the International Classification of Functioning, Disability and Health as a reference system for comparative evaluation and standardized reporting of rehabilitation interventions. European journal of physical and rehabilitation medicine 2019; 55: 384-94.
3. Prodinger B, Stamm T, Peterson D, Stucki G, Tennant A. Toward a Standardized Reporting of Outcomes in Hand Osteoarthritis: Developing a Common Metric of Outcome Measures Commonly Used to Assess Functioning. Arthritis care & research 2016; 68: 1115-27.
4. Prodinger B, Tennant A, Stucki G, Cieza A, Ustun TB. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. Journal of health services research & policy 2016; 21: 223-8.
5. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. European journal of physical and rehabilitation medicine 2017; 53: 144-9.
6. Dorjbal D, Cieza A, Gmunder HP, Scheel-Sailer A, Stucki G, Ustun TB, Prodinger B. Strengthening quality of care through standardized reporting based on the World Health Organization's reference classifications. International journal for quality in health care 2016; 28: 626-33.
7. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Review Information: Rehabilitation [Internet], 2018. Bern, Switzerland: https://www.anq.ch/en/departments/rehabilitation/ Accessed [2020 July 31].
8. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W, Robinson N. Lost in knowledge translation: time for a map? The Journal of continuing education in the health professions 2006; 26: 13-24.
9. Batalden PB, Davidoff F. What is "quality improvement" and how can it transform healthcare? Quality & safety in health care 2007; 16: 2-3.
10. Brownson RC, Colditz GA, Proctor EK. Dissemination and Implementation Research in Health: Translating Science to Practice. New York NY, United States: Oxford University Press USA, 2012.
11. Swiss National Science Foundation, NRP74 - SNSF. Project 21: Standardised Assessment and Reporting System of functioning of people with chronic diseases (StARS) [Internet], 2017. Bern, Switzerland: http://www.nfp74.ch/en/projects/healthcare-across-sectors/project-prodinger Accessed [2020 July 31].
12. Stucki G, Bickenbach J. Functioning information in the learning health system. European journal of physical and rehabilitation medicine 2017; 53: 139-43.
13. Lavis JN, Boyko JA, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 14: Organising and using policy dialogues to support evidence-informed policymaking. Health Research Policy and Systems 2009; 7: S14.
14. Lavis JN, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP). Health Research Policy and Systems 2009; 7: I1.
15. Lavis JN, Permanand G, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 13: Preparing and using policy briefs to support evidence-informed policymaking. Health Res Policy Syst 2009; 7 Suppl 1: S13.
16. Casebeer A. Application of SWOT analysis. British journal of hospital medicine 1993; 49: 430-1.
17. Swiss Learning Health System, SLHS. Stakeholder Dialogue [Internet], 2018. Luzern, Switzerland:

    https://www.slhs.ch/en/learning-cycles/stakeholder-dialogues Accessed [2020 July 31].
18. Rubinelli S, von Groote PM. Stakeholder Dialogue as Deliberation for Decision Making in Health Policy and Systems: The Approach from Argumentation Theory. American journal of physical medicine & rehabilitation 2017; 96: S17-S22.
19. Gordon D, McKay S, Marchildon G, Bhatia RS, Shaw J. Collaborative Governance for Integrated Care: Insights from a Policy Stakeholder Dialogue. Int J Integr Care 2020; 20: 3.
20. Best SJ, Krueger BS. Internet Survey Design. In: Fielding NG, Lee RM, Blank G, editors. The SAGE handbook of online research methods. Los Angeles, United States: Sage; 2008. p. 217-36.

Chapter 7

Discussion and conclusion

Discussion and conclusion

This thesis covers an important topic with a clear need for research – the standardization in the field of quality improvement [1–4], arguing that standardization can be achieved by the means of a StARS for functioning outcomes in national quality reviews in rehabilitation. The thesis demonstrates how it can be examined if certain assessment tools can be integrated in a StARS for functioning information, how a common metric that encompasses different functioning assessment scores can be created as the core of a StARS, how the influence and added value of a StARS can be examined and how corresponding implementation strategies can be mutually developed with relevant stakeholders. This final chapter summarizes the main findings of the four studies and stakeholder involvement activities and discusses the findings with reference to the research objective, specific aims and related literature. Furthermore, this chapter addresses strengths and limitations and examines the implications of the findings for practice, research and policy.

7.1 Summary of main findings

The main findings of this thesis can be illustrated through the specific aims of this thesis.

Specific aim 1, to examine whether the respective scores of the functioning assessment tools used in Swiss national quality reviews in neurological and musculoskeletal rehabilitation can be reported as unidimensional and interval-scaled metric, was addressed in two studies. Both studies, one for the assessment tool FIM™ and one for EBI, showed no fit with the Rasch model in their baseline Rasch analyses. Fit to the Rasch model was achieved only when the items of the assessment tools were grouped into two alternative testlets. The alternative testlets divided conceptually similar items into two super items of equal size and accommodated for local dependencies across the assessment tool items. Based on these two testlet solutions, a transformation table was created for each assessment tool to convert the raw ordinal scores of each tool to the corresponding interval-scaled Rasch estimate. Both studies provide evidence that the total scores of the FIM™ and of the EBI assessed for musculoskeletal and neurological rehabilitation can be reported as a unidimensional and interval-scaled metric, when this Rasch-based transformation table is applied. The approach taken in both studies focuses on the respective total scores rather than the single items or groups of items by emphasizing the similarity of the items measuring the construct of functional independence, which also represents the level reported in the Swiss national quality reviews. The findings of the two studies show that it is possible to integrate both tools in an interval-scaled StARS. Furthermore, the findings reflect the importance of addressing the issue of local dependency among items, a problem that often arises when employing tools developed according to classical test theory. This problem has recently received attention in the field of Rasch analysis for health related assessment tools [5–7]. The two studies provide a novel Rasch-based solution for dealing with local dependency and for enabling the use of well-es-

tablished assessment tools in clinical practice. The studies show that outcomes assessed with these tools can be transformed on interval scale level and therefore be validly aggregated for multiple purposes, allowing for a resource-friendly use of the same data.

Specific aim 2, to create an ICF-based and interval-scaled common metric including the functioning assessment tools used in Swiss national quality reviews in musculoskeletal and neurological rehabilitation was addressed in a third study. The linking to the ICF according to the ICF Linking Rules showed that FIM™ and EBI are conceptually congruent assessment tools, fulfilling the first requirement for scale equating, i.e. content equivalence. The subsequent analysis showed fit to the Rasch model, fulfilling the second requirement for scale equating, i.e. score equivalence. Consequently, an interval-scaled common metric, which also allows for a transformation of FIM™ and EBI scores, could be established based on the corresponding Rasch estimates. The ICF-based and interval-scaled common metric enables the comparison of the functioning outcomes of patients and clinics irrespective of the different tools used. The common metric builds the core of a StARS for functioning outcomes, which in turn, supports standardization, data aggregation and comparability, while also allowing for the use of different well-established assessment tools in practice [3, 8].

Specific aim 3, to examine the added value and influence of the interval-scaled and ICF-based StARS upon the current reporting of functioning outcomes in Swiss national quality reviews, was addressed in a fourth study. The comparison of interval-scaled reporting with the currently conducted ordinal-scaled reporting showed that the achieved outcomes reported on an interval-scale level through a StARS tended to be smaller but more precisely estimated. This finding indicates an overestimation of achieved outcomes in the current ordinal-scaled quality reviews. The added value of grounding the StARS in the ICF was examined by comparing the content of the common metric with relevant ICF Core Sets. This comparison revealed additional ICF categories, such as energy and drive functions or maintaining a basic body position, which should be included in the StARS in order to enhance the content and international comparability of the Swiss national rehabilitation quality reviews. This study suggests the decision to apply an ordinal-scaled versus an interval-scaled reporting system should be carefully considered, as the magnitude and precision of reported outcomes was demonstrated to be influenced by the type of scale employed in the reporting system. Furthermore, the study examined the added value of the StARS' ICF basis, showing concrete functioning categories that can inform the further development of national outcome quality reviews in rehabilitation. The standardization of functioning outcomes through the ICF is of value at different levels of the health system, such as the use in clinical decision making, comparison of institutions or rehabilitation programming on a policy level [9].

Specific aim 4, to develop strategies with relevant stakeholders for implementing the StARS in Swiss national rehabilitation reviews, was addressed in the form of stakeholder involvement activities, namely through a stakeholder brief and a stakeholder dialogue. The stakeholder brief was designed to introduce the need for and present

a proposal of a StARS for functioning outcomes in the Swiss health care system and to inform all stakeholder dialogue participants about the stakeholder dialogue and the corresponding research project. The one-day stakeholder dialogue was conducted to inform the 24 participating stakeholders about the research project and to develop implementation strategies for a StARS in Swiss rehabilitation. The output of the stakeholder dialogue comprised an implementation agenda, in which the stakeholders decided the next steps regarding the implementation. This agenda included four main points: i) the preparation of a proposal for a StARS for functioning outcomes for all ANQ rehabilitation areas, ii) the process to follow within the ANQ for the potential implementation of a StARS in Swiss rehabilitation quality reports, iii) the clarification of opportunities related to the new Swiss national quality commission and agreements, iv) the creation of a StARS consortium for further collaboration and exchange. The subsequent online evaluation of the stakeholder brief and stakeholder dialogue revealed that the stakeholders were satisfied with the brief and their involvement in the dialogue and that the stakeholder dialogue objective was reached. The participating stakeholders valued the opportunity for exchange and discussion with the other stakeholders and the research team highly and described the dialogue as meaningful, stimulating and interesting.

7.2 General discussion

This doctoral thesis provides a concrete example of the development of a StARS for functioning outcomes that can be implemented in national quality reviews in rehabilitation. The logic next question is: how exactly can a StARS for functioning outcomes contribute to strengthening quality improvement in rehabilitation? As introduced earlier, measurement and monitoring are vital to health care quality improvement, as they provide the means for defining the achievements of health care institutions for comparing these achievements with set targets in with the ultimate goal of identifying actions for improvement [1]. A StARS for functioning outcomes based on an interval-scaled and ICF-based common metric supports both measurement and monitoring and also enables the integration of functioning outcomes in health information systems. Given this, a StARS promotes learning from functioning information, building an important basis for quality improvement activities [3, 9, 10].

It is important to highlight that although a StARS for functioning outcomes makes data aggregation possible and fosters data comparison, learning and evidence-based decision making, a StARS only provides a basis for quality improvement. It does not automatically or directly lead to quality improvement. Quality improvement necessitates a culture in which this data is actively used, purposefully interpreted and translated into feasible quality improvement actions or processes [11–13]. A StARS is a means for supporting this culture. However, a StARS has to be actively used and continuously refined, in the sense of a continuous learning process. An example for further development of the StARS developed in this thesis, is the addition of other functioning assessment tools that better cover the aforementioned missing functioning aspects identified though ICF linking and the comparison with relevant ICF Core Sets [14, 15].

The characteristics of a StARS for functioning outcomes, i.e. its ICF basis and its in-

terval-scaled common metric, contribute to laying a foundation for learning and potential quality improvement. On one hand, the StARS' ICF basis enhances the quality review content and its comparison with other sources of functioning information. Using the ICF helps to identify what functioning content is already covered in the reviews and can inform further the development of the reporting system [16–19], so that a meaningful and more comprehensive picture of functioning is represented in the reviews. On the other hand, the interval-scaling characteristic of the StARS allows the valid aggregation and calculation of change scores of the relevant data [3, 8, 20–22]. Moreover, a StARS can strengthen quality improvement by providing comprehensive and consistent reporting of functioning outcomes, no matter if collected with different assessment tools. What matters is the information collected, not the tool. Thus, a StARS can potentially inform evidence-informed decision making [20]. The studies conducted within the scope of this doctoral thesis show how different assessment tools measuring the same aspects of functioning can be integrated in a StARS on an interval scale basis. This means that any subsequent data aggregation is valid and make the comparability of the data collected with different tools in different rehabilitation clinics possible. The ability to continuously compare data across clinics enriches the data, which can in turn inform continuous improvement. This continuous cycle of improvement constitutes a learning health system [9]. It is noteworthy that the current practice in the Swiss system of invalidly aggregating ordinal-scaled functioning data and using this data for calculating means and change scores is not uncommon nor isolated in Switzerland. The practice is seen in other comparable national rehabilitation quality reports or databases, such as the ones from the Australasian Rehabilitation Outcomes Centre (AROC), the Canadian Institute for Health Information (CIHI) or the US Model System [23–25].

Learning, evidence-informed decision making and potential improvement of care resulting from the reporting of functioning outcomes through a StARS are all possible at all three levels of the health system [9]. At the micro level, clinicians can compare individual patients with each other, or a clinician can compare her or his achievements in comparison to other clinicians treating patients groups with similar profiles. Consequently, this comparison can foster clinical decision making, goal setting or the evaluation of the treatment [26]. Furthermore, clinicians can continue to use well-established assessment tools that fit their clinical needs. At the meso level, the level reflected by the ANQ reviews, the service provision of a rehabilitation clinic can be compared with that of other clinics, irrespective of the assessment tool the clinic uses. Likewise, at the macro level, population level achievements of a region or a country can be contrasted to the another region or country. Such comparisons can inform and guide rehabilitation programming [9]. There is dynamic influence across these three levels. In other words, if quality improvement activities at the meso level are informed by functioning outcomes assessed and reported through a StARS, the quality improvement activities at the other two levels can potentially profit as well. In fact, the meso level is considered the connector of all the levels. The data that has to be collected at the patient level flows through the meso level service provider to the macro level for use in rehabilitation policy or national health care quality initiatives, among other things [4, 27]. Moreover, if a standard reference, such as the ICF as in the case of the present doctoral thesis, is used, consistency across all three levels is enhanced [9].

In order to support measurement and monitoring of useful quality indicators at all three levels, these indicators need to be carefully designed in order to measure whether the set objective is achieved. In rehabilitation, the overarching objective is the optimization of patient functioning [1, 28, 29]. Unfortunately, quality indicators are often selected on the basis of whatever data are routinely available on the micro level [1]. As the basis of the StARS, the ICF, specifically the ICF Core Sets (i.e. purpose tailored shortlists of ICF categories developed in a standardized multimethod scientific process) [30], can help the choice of meaningful functioning categories to be reported as functioning quality indicators [8, 15, 30].
In any case, if comparisons are made, irrespective of the health care level, it is important to consider proper risk adjustment methods in addition to a StARS, to further ensure that the comparison is done on equivalent data [31–33].
In addition to the design of the indicators, it is essential that users of the collected data at each health care level carefully consider how to use the data to optimally support quality improvement activities. For example, researchers encounter a challenge when they want to use data collected in the context of quality improvement for research projects. Many countries make a distinction between data used for quality improvement and data used for research, even though these two application areas are not reliably distinguishable [34, 35]. This was also the case in the research conducted as part of this thesis, as clear guidelines for the use of quality improvement data in research projects are not yet available [36].
Berwick et al. assert that functioning outcomes-informed quality improvement activities at all three levels supposedly act on the change pathway [37]. Although information used for the change pathway also automatically contains relevant information for informing patients or referring physicians and the related selection pathway, information on health care quality is currently still scarcely used for the selection pathway [37, 38]. In order to influence not only the side of the provider through the change pathway, but also the selection behaviour of the one choosing the provider, the granularity of reporting would most likely need to be higher and better fitted to the consumer's needs. Currently, the level of granularity of reporting often only stays at the level of the clinic or the rehabilitation department, e.g. neurological rehabilitation. More detailed reporting, such as on the level of diagnosis or the level of the treating clinician could support informed decisions on the selection pathway [39].
A StARS for functioning outcomes has additional application areas beyond its use in quality improvement. Other than the application areas already mentioned in the stakeholder brief and dialogue, such as the planning and performance mandates of cantonal or state authorities, and in national health statistics, two application areas warrant attention. One application area is the use of a StARS for functioning outcomes in clinical decision making, i.e. as the starting point for developing a clinical decision support tool [40–42]. A StARS could promote a more comprehensive understanding of functioning in patients with certain diagnoses or patient profiles, which in turn, can be fundamental for rehabilitation planning and decision making among health professionals and patients [43]. Another application area of a StARS is its use for data aggregation in research projects. A StARS can be used to learn from functioning information assessed with different tools applied in different research projects or databases. A StARS could foster data aggregation in meta-analyses such as in Cochrane reviews [44], or in other research projects that need to aggregate function-

ing outcomes derived from different sources or collected with different tools, e.g. making functioning outcomes of patients with hand osteoarthritis assessed with different commonly used assessment tools comparable [45].

7.3 Limitations and strengths

This section highlights internal and external limitations and strengths of the doctoral project.

One project internal limitation is that the StARS developed for this project encompasses only two functioning assessment tools. Moreover, these two tools only cover certain functioning aspects (according to ICF linking presented in Chapter 4) and do not reflect a complete picture of patient functioning outcomes. Nevertheless, these specific functioning aspects are the ones that are reflected in the current Swiss system for quality reviews in rehabilitation. In order to represent a more comprehensive and more meaningful picture of patient functioning, an ICF-based StARS would need to be developed according to the four steps recommended to be followed when documenting functioning with the ICF. These steps are: 1) defining ICF domains to document, 2) choosing the perspective to take, 3) selecting data collection tools and 4) deciding on a reporting approach, such as the use of an interval-scaled common metric [8].

Other project internal limitations exist in regard to the methodology employed. First, in order to include all scoring options of the FIM™ that exceeded the Rasch analysis RUMM2030 software limit of 101 scoring options, the creation of the common metric had to be adjusted following a novel two-tiered approach [46]. Furthermore, the creation of the common metric included only neurological rehabilitation cases, even though musculoskeletal rehabilitation cases would also be needed for the common metric. Having only neurological cases for the development of the common metric is attributed to the ANQ ADL score data, as this data only contained neurological cases. Nevertheless, based on the results of the first two studies (see Chapter 2 &3) it can be concluded that the common metric can also be applied for musculoskeletal rehabilitation, as there was no indication of different item functioning (DIF) for the two rehabilitation groups. Second, the examination of added value and impact of the common metric was also conducted in a novel self-designed format, as there are no established standards yet on how to examine the added value and impact of an ICF-based and interval-scaled StARS. This may have resulted in potential bias in study 4. Third, the StARS developed in this thesis is oriented towards the assessment tools' total scores and does not allow the reporting at the level of single items or subgroup of items. Reporting at the single item or subgroup level may be of interest to clinicians or other researchers. In the field of scale equating, different approaches exist; the research team decided for the total score-based approach suggested by Andrich [47]. The other scale equating approaches were not chosen due to certain disadvantages. For example, the expert-consensus-based approach applied in the development of the ANQ ADL score has the disadvantages that it is ordinal-scaled, item-based and does not consider the operational ranges of the included scales [48]. Another approach not chosen is the Leunbach's model, which only allows for direct equating between two scores and does not result in a neutral common metric. The

latter impedes the comparison of more than two tools [49].
A project external limitation is the fact that a StARS itself has its limits. There are certain risks in public quality reports that cannot be influenced by a StARS, like the quality of data reported by the clinicians or clinics and potential manipulations. Data quality issues may be due to unrealistic external targets or disincentives, such as quality-related payment schemes [1, 50, 51].
Furthermore, a StARS should be used with thought and prudence. Even though with the StARS developed in the scope of this doctoral thesis has enabled the aggregation and comparison of outcomes of the different rehabilitation groups of neurological and musculoskeletal rehabilitation, attention must be paid that "like is compared with like", as already mention in relation to the risk adjustment.
Moreover, in order to optimally learn from functioning outcomes used for improving the quality of care, data on structure and processes would be needed according to the Donabedian model [52] to understand what aspects influence desirable or non-desirable outcomes. Such information would also need to be assessed and reported in a standardized way. However, as data on structure and processes do not incorporate aspects of functioning, they cannot be included in a StARS for functioning outcomes as presented in this thesis.

A project internal strength is the project's novelty; it is the first concrete StARS for functioning outcomes that has been developed for use in national rehabilitation quality improvement. It is a concrete response to the expressed need for comparing the functioning outcomes of various clinics that have been assessed with different tools and to aggregate data in national rehabilitation quality reviews in musculoskeletal and neurological rehabilitation. With this concrete StARS standardization is possible without having to define a single assessment tool as a standard and more importantly, clinicians can continue to use the tools they have been using and which best fit their specific needs.
Additional project internal strengths lie in the methodology applied in this project. The developed StARS and the corresponding studies build upon a relevant and large patient sample with national coverage, representing all of Switzerland's language regions. The creation of the common metric employed the same data as the currently applied ANQ ADL score. This allowed a meaningful direct comparison of the two reporting systems. Another project internal strength is the active involvement of stakeholders in the development of implementation strategies in order to bridge research and practice. The doctoral thesis studies focused not only on the development of the StARS but also on its influence and added value in contrast to the existing national quality reporting system. This helped to inform corresponding implementation efforts. Moreover, concrete implementation strategies were worked out together with relevant stakeholders based on a stakeholder brief and through the conduct of a stakeholder dialogue [53–55].
There are also several project external strengths. First, the research presented in this thesis covers a topic in which research is needed [2, 4]. The research project is in line with WHO's Rehabilitation 2030 initiative that calls for the optimization of functioning through rehabilitation and for the integration of functioning data in health information systems. In this initiative, functioning data is seen as underpinning decisions in health

policy, management and clinical care and have the potential to facilitate the allocation of rehabilitation resources and strategic decision making [56]. The present research is also aligned with WHO's mission to assist countries to develop a comprehensive rehabilitation-specific strategic plan that should include monitoring, evaluation and review processes of rehabilitation services [57]. In addition, according to the stakeholders involved in the NRP74 StARS advisory board and stakeholder dialogue, the project covers a concrete need in the Swiss system for quality development in rehabilitation. They appreciated the potential application of a StARS for functioning outcomes in diverse application areas, such as performance mandates of the Swiss cantons or its use in Swiss national health statistics. Furthermore, the stakeholder dialogue led to the commitment of the participating stakeholders to follow-up activities and the plan to establish a StARS consortium supporting the implementation of the research presented in this thesis and other potential research projects regarding the standardization of functioning outcomes. The concrete planning of follow-up activities is in line with the guiding principles of how stakeholder dialogues should be conducted to support evidence-informed implementation and policymaking [53].

7.4 Implications

Practical implications

A major practical implication of a StARS for functioning outcomes is that it allows clinicians and clinics to continue using the assessment tools that they have been using, while also enabling the comparison and aggregation of functioning outcome data across clinics and clinicians who use different assessment tools. The tools used can be chosen based on the clinic's and clinicians practical needs, such as less time required for the assessment or a simpler rating system [58, 59]. Likewise, clinically meaningful assessment tools can continue to be used; a StARS does not require that a single tool is defined [60]. Furthermore, different tools are differently suited for specific patient populations. For example, the result of the last study presented in this thesis (see Chapter 5) showed that the EBI has ceiling effects in the musculoskeletal population, meaning that the EBI scale cannot adequately differentiate between patients with high functional independence. Also to be considered on this behalf, is that floor and ceiling effects may increase bias and uncertainty in statistical evaluations [61], what is relevant when statistics are considered for making decisions in clinical practice as well as other decisions at all levels of health care.

The studies presented in this thesis show that existing tools can be further used, but in order to do so they need to be critically evaluated regarding their unidimensionality and transformability onto the interval-scale level. The Rasch model, as applied in the underlying studies, can provide solutions for reporting the total scores of ordinal scaled assessment tools on interval scale level and the ICF can help to make the content of these tools comparable [62, 63]. The transformation of the assessed data to the interval scale level or the integration into a common metric can take place in the background, e.g. through the programming of automatic conversions when the data is entered into the electronic health record. The way the data is collected does not change, but the way in which the related functioning outcome data is aggregated and interpreted does.

As stated above in the discussion section, being able to learn from the valid aggregation and comparison of functioning outcomes and improvement in care are expected consequences of a StARS. The mechanisms of how this data can be best used to bring improvement in clinical practice requires further investigation, especially in consideration of different treatment modalities and patient profiles hat lead to better outcomes [64].

Research implications

The knowledge gained and the methodology applied in the research presented in this thesis can inform future research that focuses on standardizing the reporting of functioning outcomes. Further research is also needed on the mechanism of how the information provided through a StARS can influence and best support quality improvement activities to ultimately lead to quality improvement of care. Moreover, other examples of a StARS for functioning outcomes should be created. For example, a newly created StARS could target other functioning assessment tools, such as tools focusing on participation [17], other rehabilitation groups, such as cardiovascular rehabilitation, or other similar national rehabilitation outcome quality reviews such as the one from AROC [23].

Researchers need to consider that data resulting from functioning assessment tools might be ordinal-scaled and has to be transformed onto interval scale level before the data can be validly aggregated and used for calculations such as change scores, as shown in the present thesis. Otherwise an over- or underestimation of the achieved functioning outcomes can occur, as already found in previous research [22, 65, 66].

Another implication for research addresses implementation. Since the implementation strategies for the Swiss context have been identified in the research related to this thesis, carrying the process of implementation to the next and concrete level is warranted. For this purpose, conducting further implementation research would be beneficial. Specifically, research on what leads to a successful implementation of StARS for functioning outcomes and the ideal approach for informing quality improvement initiatives in rehabilitation [67]. For this implementation research, it would be valuable to again incorporate the perspective of relevant stakeholders, as the approach of actively seeking stakeholder involvement as part of the present thesis was much appreciated by the participating stakeholders, the research funder and the research team.

Policy implications

This thesis describes the development of a StARS for functioning as a health indicator and main indicator for rehabilitation. Functioning was operationalized using the ICF, the internationally recognized standard for documenting functioning [62, 68, 69]. The ICF basis of the StARS illustrated in this thesis makes it possible to conceptually decide whether existing assessment tools can be included in a StARS. Furthermore, the ICF basis supports the standardization and the international comparability of the tools encompassed in a StARS, as WHO Europe calls for in national quality monitoring programmes [1]. Likewise, the comparison of different clinics, regions or nations on a larger scale supports optimal planning and overarching quality improvement initiatives, whereby thinking outside the box of institution related quality improvement [70].

In terms of Switzerland, there is support for implementing the developed StARS for the national ANQ quality reviews in musculoskeletal and neurological rehabilitation and for developing and implementing it in other rehabilitation areas as recommended in the stakeholder dialogue. Implementing the developed StARS would ensure that the calculations and aggregations already conducted in the current quality reviews would be done on a valid interval-scaled basis [22, 71]. Furthermore, the StARS' ICF basis showed how the current Swiss system can be meaningfully expanded to complete the picture of patients' functioning outcomes in national quality reviews.
Moreover, the research in the present thesis can also been seen as an acknowledgement of the importance of functioning information for the field of rehabilitation and for the health system as a whole [72]. This was also confirmed by the participants of the stakeholder dialogue.
Last but not least, an additional implication on the policy level is related to research ethics and quality improvement data. The procedure for getting ethical clearance for conducting research in the field of quality improvement is challenging [34–36]. Policies should be developed for improving protections of participants in quality improvement projects while minimizing burden on researchers seeking to collect or use quality improvement data. These policies should include clear guidelines and reasonable ethical oversight that aligns risks with the needs of a learning health system [35].

7.5 Conclusion

This thesis demonstrates how a StARS for functioning outcomes for use in national quality reviews can be created, how its influence and added value can be examined and how related implementation strategies can be developed thorough stakeholder involvement. A StARS enables aggregation and comparison of functioning outcomes assessed with diverse established assessment tools through an ICF-based and interval-scaled common metric. A StARS also promotes learning from functioning information and can inform continuous improvement activities as part of a learning health system. The results of this thesis call for carefully considering whether a reporting system applies an interval or ordinal scale as the interpretation of reported functioning outcomes with the latter might lead to misinterpretation of the outcomes. The StARS' ICF basis brings added value by informing the further development of functioning as a relevant indicator for national outcome quality reviews in rehabilitation, while also fostering international comparison. Lastly, a StARS for functioning outcomes can be applied at the different levels of a health system- at the micro level to support clinical decision making, at the meso level to support institutional evaluation and quality improvement and at the macro level by guiding the national rehabilitation programming.

References Chapter 7

1. Shaw C, WHO Regional Office for Europe (Health Evidence Network). How can hospital performance be measured and monitored? [Internet], 2003. Copenhagen, Denmark:

   http://www.euro.who.int/document/e82975.pdf Accessed [2020 Jul 31].

2. Jesus TS, Hoenig H. Crossing the Global Quality Chasm in Health Care: Where Does Rehabilitation Stand? Archives of physical medicine and rehabilitation 2019; 100: 2215-7.

3. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. European journal of physical and rehabilitation medicine 2018; 54: 110-7.

4. Vincent C, Staines A, Swiss Federal Office of Public Health FOPH. Enhancing the Quality and Safety of Swiss Healthcare - A national report commissioned by the Federal Office of Public Health on the quality and safety of healthcare in Switzerland [Internet], 2019. Bern, Switzerland:

   https://www.bag.admin.ch/bag/en/home/versicherungen/krankenversicherung/krankenversicherung-qualitaetssicherung.html Accessed [2020 July 31].

5. Wilson M. Detecting and interpreting local item dependence using a family of Rasch models. Applied psychological measurement 1988; 2: 353-64.

6. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. Journal of rehabilitation medicine 2011; 43: 884-91.

7. Prodinger B, Hammond A, Tennant A, Prior Y, Tyson S. Revisiting the disabilities of the arm, shoulder and hand (DASH) and QuickDASH in rheumatoid arthritis. BMC musculoskeletal disorders 2019; 20: 41.

8. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. European journal of physical and rehabilitation medicine 2017; 53: 144-9.

9. Stucki G, Bickenbach J. Functioning information in the learning health system. European journal of physical and rehabilitation medicine 2017; 53: 139-43.

10. Prodinger B, Cieza A, Oberhauser C, Bickenbach J, Ustun TB, Chatterji S, Stucki G. Toward the International Classification of Functioning, Disability and Health (ICF) Rehabilitation Set: A Minimal Generic Set of Domains for Rehabilitation as a Health Strategy. Archives of physical medicine and rehabilitation 2016; 97: 875-84.

11. Lindstrom Egholm C, Helmark C, Christensen J, Eldh AC, Winblad U, Bunkenborg G, Zwisler AD, Nilsen P. Facilitators for using data from a quality registry in local quality improvement work: a cross-sectional survey of the Danish Cardiac Rehabilitation Database. BMJ open 2019; 9: e028291.

12. Canaway R, Bismark M, Dunt D, Kelaher M. Medical directors' perspectives on strengthening hospital quality and safety. Journal of health organization and management 2017; 31: 696-712.

13. socialdesign ag, Wetz S, Jörg R, Ruflin R, ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken,. Evaluation of the current status of ANQ rehabilitation reviews [Internet], 2018. Bern, Switzerland: https://www.anq.ch/wp-content/uploads/2018/05/ANQ_Reha_Evaluation-Report_April18.pdf Accessed [2020 July 31].

14. Grill E, Quittan M, Fialka-Moser V, Muller M, Strobl R, Kostanjsek N, Stucki G. Brief ICF Core Sets for the acute hospital. Journal of rehabilitation medicine 2011; 43: 123-30.

15. Grill E, Strobl R, Muller M, Quittan M, Kostanjsek N, Stucki G. ICF Core Sets for early post-acute rehabilitation facilities. Journal of rehabilitation medicine 2011; 43: 131-8.

16. Cieza A, Fayed N, Bickenbach J, Prodinger B. Refinements of the ICF Linking Rules to strengthen their potential for establishing comparability of health information. Disability and rehabilitation 2019; 41: 574-83.

17. Ballert CS, Hopfe M, Kus S, Mader L, Prodinger B. Using the refined ICF Linking Rules to compare the content of existing instruments and assessments: a systematic review and exemplary analysis of instruments measuring participation. Disability and rehabilitation 2019; 41: 584-600.

18. Maritz R, Pongpipatpaiboon K, Melvin JL, Graves DE, Prodinger B. Content comparison of the Spinal Cord Injury Model System Database to the ICF Generic Sets and Core Sets for spinal cord injury. Spinal cord 2019; 57: 1023-30.

19. Maritz R, Baptiste S, Darzins SW, Magasi S, Weleschuk C, Prodinger B. Linking occupational therapy models and assessments to the ICF to enable standardized documentation of functioning. Canadian journal of occupational therapy 2018; 85: 330-41.

20. Prodinger B, Tennant A, Stucki G, Cieza A, Ustun TB. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. Journal of health services research & policy 2016; 21: 223-8.

21. Andrich D. Rating scales and Rasch measurement. Expert review of pharmacoeconomics & outcomes research 2011; 11: 571-85.

22. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? Journal of rehabilitation medicine 2012; 44: 97-8.

23. Australasian Rehabilitation Outcomes Centre, AROC. AROC clinical datasets and supporting documents [Internet], 2020. Wollongong NSW, Australia https://www.uow.edu.au/ahsri/aroc/dataset/inpatient-dataset/ Accessed [2020 July 31].

24. Canadian Institute for Health Information, CIHI. Rehabilitaion [Internet], 2020. Ottawa, Canada: https://www.cihi.ca/en/rehabilitation Accessed [2020 July 31].

25. Model System Knowledge Translation Center, MSKTC. Model Systems Centers Directory [Internet], 2020. Arlington, United States: https://msktc.org/tbi/model-system-centers Accessed [2020 July 31].

26. Rauch A, Cieza A, Stucki G. How to apply the International Classification of Functioning, Disability and Health (ICF) for rehabilitation management in clinical practice. European journal of physical and rehabilitation medicine 2008; 44: 329-42.

27. Verband der führenden Rehabilitationskliniken der Schweiz, SW!SS REHA. Der nationale Messplan Rehabilitation wird von SW!SS REHA und von relevanten nationalen Fachgesellschaften begrüsst und unterstützt [Internet] [German], Bern, Switzerland: https://www.swiss-reha.com/de/anq.html Accessed [2020 July 31].

28. Stucki G, Bickenbach J, Gutenbrunner C, Melvin J. Rehabilitation: The health strategy of the 21st century. Journal of rehabilitation medicine 2018; 50: 309-16.

29. Stucki G, Bickenbach J, Melvin J. Strengthening Rehabilitation in Health Systems Worldwide by Integrating Information on Functioning in National Health Information Systems. American journal of physical medicine & rehabilitation 2017; 96: 677-81.

30. Selb M, Escorpizo R, Kostanjsek N, Stucki G, Ustun B, Cieza A. A guide on how to develop an International Classification of Functioning, Disability and Health Core Set. European journal of physical and rehabilitation medicine 2015; 51: 105-17.

31. Deutsch A, Heinemann AW, Cook KF, Foster L, Miskovic A, Goldsmith A, Cella D. Inpatient Rehabilitation Quality of Care From the Patient's Perspective: Effect of Data Collection Timing and Patient Characteristics. Archives of physical medicine and rehabilitation 2019; 100: 1032-41.

32. Deutsch A, Pardasaney P, Iriondo-Perez J, Ingber MJ, Porter KA, McMullen T. Development of a Risk-adjustment Model for the Inpatient Rehabilitation Facility Discharge Self-care Functional Status Quality Measure. Medical care 2017; 55: 706-15.

33. Dümbgen L, Institute of Mathematical Statistics and Actuarial Science, University of Bern. Technical Report 78: (Ab)Using Regression for Data Adjustment [Internet], 2016. Bern, Switzerland: https://www.anq.ch/wp-content/uploads/2018/12/Technical_Report.pdf Accessed [2020 July 31].

34. Lynn J, Baily MA, Bottrell M, Jennings B, Levine RJ, Davidoff F, Casarett D, Corrigan J, Fox E, Wynia MK, Agich GJ, O'Kane M, Speroff T, Schyve P, Batalden P, Tunis S, Berlinger N, Cronenwett L, Fitzmaurice JM, Dubler NN, James B. The ethics of using quality improvement methods in health care. Annals of internal medicine 2007; 146: 666-73.

35. Fiscella K, Tobin JN, Carroll JK, He H, Ogedegbe G. Ethical oversight in quality improvement and quality improvement research: new approaches to promote a learning health care system. BMC medical ethics 2015; 16: 63.

36. McLennan S, Maritz R, Shaw D, Elger B. The inconsistent ethical oversight of healthcare quality data in Switzerland. Swiss medical weekly 2018; 148: w14637.

37. Berwick DM, James B, Coye MJ. Connections between quality measurement and improvement. Medical care 2003; 41: I30-8.

38. Prang KH, Canaway R, Bismark M, Dunt D, Miller JA, Kelaher M. Public performance reporting and hospital choice: a cross-sectional study of patients undergoing cancer surgery in the Australian private healthcare sector. BMJ open 2018; 8: e020644.

39. Canaway R, Prang KH, Bismark M, Dunt D, Kelaher M. Public disclosure of hospital clinicians. Australian health review 2019; 44: 228-33.

40. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ (Clinical research ed) 2005; 330: 765.

41. Culyer AJ, Lomas J. Deliberative processes and evidence-informed decision making in healthcare: do they work and how might we know? Evidence & policy 2006; 2: 357-71.

42. Burns AS, Yee J, Flett HM, Guy K, Cournoyea N. Impact of benchmarking and clinical decision making tools on rehabilitation length of stay following spinal cord injury. Spinal cord 2013; 51: 165-9.

43. Hodel J, Ehrmann C, Stucki G, Bickenbach JE, Prodinger B. Examining the complexity of functioning in persons with spinal cord injury attending first rehabilitation in Switzerland using structural equation modelling. Spinal cord 2020; 58: 570-80.

44. Stucki G, Pollock A, Engkasan JP, Selb M. How to use the International Classification of Functioning, Disability and Health as a reference system for comparative evaluation and standardized reporting of rehabilitation interventions. European journal of physical and rehabilitation medicine 2019; 55: 384-94.

45. Prodinger B, Stamm T, Peterson D, Stucki G, Tennant A. Toward a Standardized Reporting of Outcomes in Hand Osteoarthritis: Developing a Common Metric of Outcome Measures Commonly Used to Assess Functioning. Arthritis care & research 2016; 68: 1115-27.

46. Andrich D, Sheridan B, Luo G, Rumm Laboratory Pty Ltd. RUMM2030 Version 5.4 for Windows [Internet], 2012. Crawley WA, Australia: http://www.rummlab.com.au/ Accessed [2020 July 31].

47. Andrich D. The Polytomous Rasch Model and the Equating of Two Instruments. In: Christensen K, Kreiner S, Mesbah M, editors. Rasch Models in Health. Hoboken, United States: ISTE Ltd and John Wiley and Sons, Icn.; 2013. p. 163-96.

48. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, ANQ. Entwicklung und Validierung eines ADL-Überführungsalgorithmus auf Basis von FIM® und EBI [Internet] [German], 2018. Bern, Switzerland:

https://www.anq.ch/wp-content/uploads/2018/03/ANQ_Reha_Abschlussbericht_ADL_Algorithmen.pdf Accessed [2020 July 31].

49. Adroher ND, Kreiner S, Young C, Mills R, Tennant A. Test equating sleep scales: applying the Leunbach's model. BMC Med Res Methodol 2019; 19: 141.

50. Averill RF, Fuller RL, McCullough EC, Hughes JS. Rethinking Medicare Payment Adjustments for Quality. The Journal of ambulatory care management 2016; 39: 98-107.

51. Kondo KK, Damberg CL, Mendelson A, Motu'apuaka M, Freeman M, O'Neil M, Relevo R, Low A, Kansagara D. Implementation Processes and Pay for Performance in Healthcare: A Systematic Review. Journal of general internal medicine 2016; 31 Suppl 1: 61-9.

52. Donabedian A. Volume I The Definition of Quality and Approaches to its Assessment. Ann Arbor, United States: Health Administration Press, 1980.

53. Lavis JN, Boyko JA, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 14: Organising and using policy dialogues to support evidence-informed policymaking. Health Research Policy and Systems 2009; 7: S14.

54. Rubinelli S, von Groote PM. Stakeholder Dialogue as Deliberation for Decision Making in Health Policy and Systems: The Approach from Argumentation Theory. American journal of physical medicine & rehabilitation 2017; 96: S17-S22.

55. Swiss Learning Health System, SLHS. Stakeholder Dialogue [Internet], 2018. Luzern, Switzerland: https://www.slhs.ch/en/learning-cycles/stakeholder-dialogues Accessed [2020 July 31].

56. World Health Organization, WHO. Health information systems and rehabilitation. Rehabilitation 2030: A Call for Action [Internet], 2017. Geneva, Switzerland: https://www.who.int/disabilities/care/rehab-2030/en/ Accessed [2020 July 31].

57. World Health Organization, WHO. Rehabilitation in health systems: guide for action [Internet], 2019. Geneva, Switzerland: https://www.who.int/rehabilitation/rehabilitation-guide-for-action/en/ Accessed [2020 July 31].

58. Houlden H, Edwards M, McNeil J, Greenwood R. Use of the Barthel Index and the Functional Independence Measure during early inpatient rehabilitation after single incident brain injury. Clinical rehabilitation 2006; 20: 153-9.

59. Marolf M, Vaney C, König N, Schenk T, Prosiegel M. Evaluation of disability in multiple sclerosis patients: a comparative study of the Functional Independence Measure, the Extended Barthel Index and the Expanded Disability Status Scale. Clinical rehabilitation 1996; 10: 309-13.

60. Oude Voshaar MAH, Das Gupta Z, Bijlsma JWJ, Boonen A, Chau J, Courvoisier DS, Curtis JR, Ellis B, Ernestam S, Gossec L, Hale C, Hornjeff J, Leung KYY, Lidar M, Mease P, Michaud K, Mody GM, Ndosi M, Opava CH, Pinheiro GRC, Salt M, Soriano ER, Taylor WJ, Voshaar MJH, Weel A, de Wit M, Wulffraat N, van de Laar M, Vonkeman HE. International Consortium for Health Outcome Measurement Set of Outcomes That Matter to People Living With Inflammatory Arthritis: Consensus From an International Working Group. Arthritis care & research 2019; 71: 1556-65.

61. Šimkovic M, Träuble B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. PloS one 2019; 14: e0220889.

62. World Health Organization. International Classification of Functioning, Disability and Health - ICF. Geneva, Switzerland: WHO, 2001.

63. Stucki G, Ewert T, Cieza A. Value and application of the ICF in rehabilitation medicine. Disability and rehabilitation 2002; 24: 932-8.

64. Blackmore AR, Leonard J, Madayag R, Bourg PW. Using the Trauma Quality Improvement Program Metrics Data to Enhance Clinical Practice. Journal of trauma nursing : the official journal of the Society of Trauma Nurses 2019; 26: 121-7.

65. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. Journal of Clinical Epidemiology 1996; 49: 711-7.

66. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? British journal of rheumatology 1996; 35: 574-8.

67. Graham JE, Middleton A, Roberts P, Mallinson T, Prvu-Bettger J. Health Services Research in Rehabilitation and Disability-The Time is Now. Archives of physical medicine and rehabilitation 2018; 99: 198-203.

68. World Health Organization, WHO. Towards a Common Language for Functioning, Disability and Health – ICF [Internet], 2002. Geneva, Switzerland: https://www.who.int/classifications/icf/icfbeginnersguide.pdf?ua=1 Accessed [2020 July 31].

69. World Health Organization, WHO. How to use the ICF - A Practical Manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for comment [Internet], 2013. Geneva, Switzerland: http://www.who.int/classifications/drafticfpracticalmanual2.pdf Accessed [2020 July 31].

70. National Academies of Sciences - Engineering - Medicine. Crossing the global quality chasm: Improving health care worldwide. In: National Academies of Sciences - Engineering - Medicine, editor. Washington DC, United States: National Academies Press; 2018.

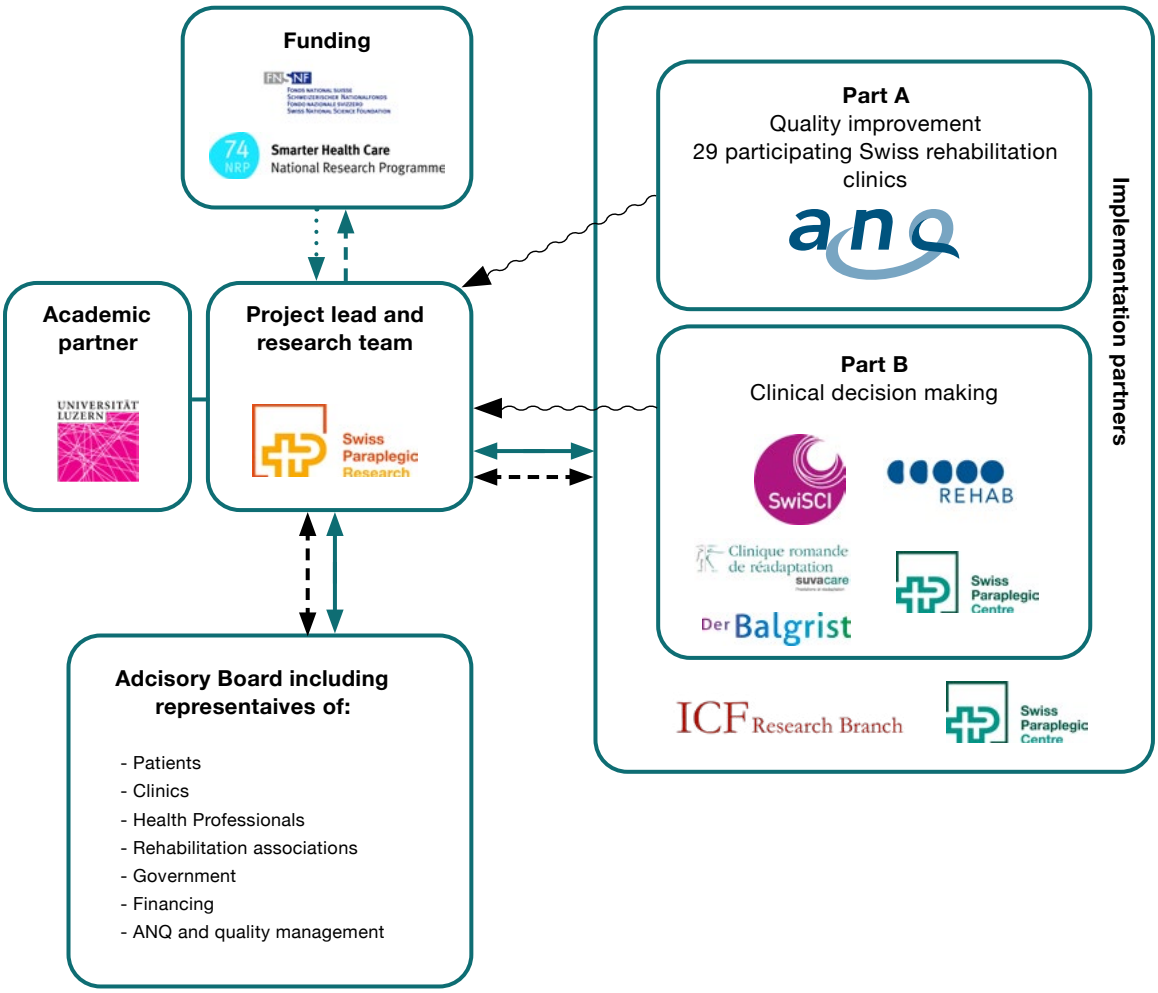
71. ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Auswertungskonzept Rehabilitation (Modul 2 & 3) [Internet] [German], 2018. Bern, Switzerland: https://www.anq.ch/wp-content/uploads/2017/12/ANQ_Module_23_Auswertungskonzept.pdf Accessed [2020 July 31].

72. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. European journal of physical and rehabilitation medicine 2017; 53: 134-8.

Supplementary material

Appendix 1

NRP74 StARS project organisation



Funding
Reporting
Data transfer
Consultation & Feedback
Interpretation of results and implementation

Appendix 2

Abstract stakeholder brief

Extract of the NRP74 StARS Stakeholder Brief (page 3)

Abstract in English

Problem: Functioning information is important at different levels of the health system. However, this information is collected with a wide variety of assessment tools, whereby impeding aggregation and comparison.

Proposed solution: An ICF-based (International Classification of Functioning, Disability and Health) and interval-scaled Standardized Assessment and Reporting System (StARS), which increases usefulness and comparability of functioning information.

Project: In the NRP74 StARS project, two concrete examples of a StARS for functioning information are being worked out: one for the Swiss national rehabilitation quality reviews (part A), and one for supporting clinical decision making in the rehabilitation of patients with a spinal cord injury (part B).

Goal of the stakeholder dialogue: The discussion and elaboration of the next steps towards the implementation of StARS in four potential application areas in Switzerland.

Areas of application: Four application areas are at the centre of the stakeholder dialogue. A StARS for functioning information in:

1) Swiss national quality reviews in rehabilitation (ANQ reviews)
2) the general assessment and reporting in rehabilitation in the Swiss clinical context
3) the planning and performance mandates of the Swiss cantons
4) Swiss national health statistics

Appendix 3

Participant list stakeholder dialogue

Extract of the NRP74 StARS Stakeholder Brief (page 20), stakeholder names are not displayed according to the Chatham House Rules

Representation	Institution	Function/Department
Federal offices & cantons	Federal Office of Public Health	Scientific employee, Department of health and accident insurance, section quality and processes
	Federal Office of Statistics	Director, Section of health and population
	Federal Office of Statistics	Scientific employee, Section of health and population
	Health Department Canton of Basel-Stadt	Director, quality monitoring
	Health Department Canton of Zürich	Director of rehabilitation, Department of health services
Finances	Commission for medical tariffs, accident insurance law	Commission for medical tariffs, accident insurance law, Department of inpatient tariffs
	SwissDRG AG	Project manager rehabilitation tariffs, Department of economics
	Santésuisse	Project manager eHealth
	curafututura	Medical officer, KPT health insurance
	Swiss Insurance Association (SVV)	Chief physician, department of health and accident insurances
Clinics and Swiss rehabilitation areas	Plateforme-Reha.ch	Member of the committee
	Reha Ticino	Director of the board and coordinator
	H+ (association of Swiss hospitals)	Project lead, quality and patient safety
Rehabilitation specialists	Swiss Medical Association (FMH)	Scientific employee, Department of data, demographics and quality
	Swiss Society for Physical Medicine and Rehabilitation (SGPMR)	President of the committee
	Swiss Association of Rehabilitation (SAR)	President of the committee
	Swiss Paraplegic Centre	Director research rehabilitation quality management (RQM)
	Swiss Society for Paraplegia (SSoP)	President of the committee
	Rehabilitation clinic Institution de Lavigny	Medical director and chief physician neurological rehabilitation, member of ANQ Quality Committee for rehabilitation
	Swiss Society for neurological rehabilitation (SGNR)	Member of the committee
Patient perspective	Swiss Paraplegic Association (SPV)	Member of the executive management
	Umbrella association Swiss patient centres	President of the committee
	Swiss Patient Security Foundation	President of the foundation council
Health care quality	National Association for Quality Development in Hospitals and Clinics (ANQ)	Director Department of rehabilitation
	Charité - Universitätsmedizin Berlin (scientific council and reporting institute of the ANQ)	Officer ANQ rehabilitation data collection and reports
	Swiss Society for Quality Management in Healthcare (sQmh)	Former Vice-president of the committee
Research and knowledge transfer	NRP74 Smarter Health Care	Head of knowledge transfer

Appendix 4

Agenda stakeholder dialogue, 25 November 2019

Extract of the NRP74 StARS Stakeholder Brief (page 18)

Location: University of Lucerne, Frohburgstrasse 3, Lucerne, Room 3.B48,

Lead and moderator of the entire day: Sarah Mantwill, Coordinator Swiss Learning Health System

08:15 – 09:00	**Welcome Coffee**
09:00 – 09:45	**Welcome & Introduction** Functioning in the learning health system & the four application areas of a StARS for functioning outcomes *Gerold Stucki, Principal Investigator NRP74 StARS project* *Roxanne Maritz, PhD student NRP74 StARS project part A Quality Improvement* *Jsabel Hodel, PhD student NRP74 StARS project part B Clinical Decision Making*
09:45 – 10:30	**Panel discussion for application areas 1) StARS for functioning outcomes national rehabilitation quality reviews and 2) general assessment and reporting in the clinical context in Swiss rehabilitation** Presentation discussion and clarification of open questions in regard to the application areas and the corresponding SWOT analyses in the policy brief *ANQ representative for application area 1* *Swiss Paraplegic Centre representative for application area 2* *Gerold Stucki. Roxanne Maritz, Jsabel Hodel (NRP74 StARS project team)*
10:30 – 11:00	**Coffee break**
11:00 – 12:30	**Group discussions application area 1** **Group discussions application area 2**
12:30 – 13:45	**Lunch break**
13:45 – 14:30	**Plenary - Creation of the implementation agenda** Presentation of the group discussions, plenary discussion to elaborate concrete next steps towards implementation of the NRP74 StARS projects outputs
14:30 – 14:45	**Coffee break**
14:45 – 15:45	**Plenary discussion application areas 3) a StARS for functioning outcomes for the planning and performance mandates of Swiss cantons and 4) a StARS for functioning outcomes for national health statistics.** Assessment of current situation and needs. Opening comments from: *Representative of the Canton of Basel-Stadt, Representative of the Canton of Zürich, Representative of the Federal Office of Statistics, Representative of the Federal Office of Public Health*
15:45 – 16:00	**Concluding feedback round, outlook and closure** *Gerold Stucki Principal Investigator NRP74 StARS project*
16:00 – 17:00	**Apéro riche**

Appendix 5

SWOT analysis

Extract of the NRP74 StARS Stakeholder Brief (page 11)

SWOT analysis for application area 1: StARS for national quality reviews in rehabilitation
Yes/No: Please indicate whether you agree with the respective point (yes) or not (no)
Comments/additions: Please indicate possible comments or additions to the points already mentioned

Strengths	Yes/No:	Weaknesses	Yes/No:
1) The developed StARS can be directly used for reporting, without changing the data collection method.	Y ¤ N ¤	1) Currently, only two assessment tools are included in the corresponding StARS for the ANQ. An extension is possible, but resource intensive.	Y ¤ N ¤
2) The developed StARS is based on the ICF, the international recognized WHO classification for functioning.	Y ¤ N ¤	2) The developed StARS must be extended to the other ANQ rehabilitation areas and instruments, so that the ANQ evaluation logic is the same across all areas.	Y ¤ N ¤
3) The developed StARS is based on interval scaling. This allows the valid calculation of the difference between admission and discharge scores.	Y ¤ N ¤	3) The StARS project includes a transformation algorithm of EBI and FIM™ scores, although such algorithms have already been developed using a different methodology (ANQ ADL score, SwissDRG transformation).	Y ¤ N ¤
4) The developed StARS makes the outcomes assessed with EBI and FIM™ directly comparable and takes the operational range (coverage of the difficulty levels) of the two integrated assessment tools into account. Comments/additions:	Y ¤ N ¤	Comments/additions:	
Opportunities	**Yes/No:**	**Threats**	**Yes/No:**
1) The developed StARS can be adapted and supplemented with further assessment tools.	Y ¤ N ¤	1) Users of the ANQ reports do currently not have an intuition how to interpret the interval scale of instruments (compared to their ordinal scale to which they are used to).	Y ¤ N ¤
2) The StARS' interval scaling does not lead to an over- or underestimation of the achieved outcome quality, in contrast to the current practice with calculations based on ordinal scaling.	Y ¤ N ¤	2) Although a StARS would allow cross-sectoral comparisons, caution must be taken to compare like with like.	Y ¤ N ¤
3) The ICF basis of the StARS shows gaps in the current ANQ review system and therefore potential further development. Comments/additions:	Y ¤ N ¤	3) On an international level, ordinal scales are often reported. Thus for international comparability, ordinal scaled reporting would still be necessary. This might lead to confusion. Comments/additions:	Y ¤ N ¤

Colophon

The work presented in this thesis was conducted at Swiss Paraplegic Research, Nottwil, Switzerland and the University of Lucerne, Lucerne, Switzerland.
The research and related studies were funded by the Swiss National Science Foundation´s National Research Programme Smarter Health Care (NRP 74) within the project „Enhancing continuous quality improvement and supported clinical decision making by standardized reporting of functioning“ [grant number: 407440_167412/1].

Furthermore, this thesis was supported by a PhD mobility grant from the Swiss National Science Foundation for a six-month research fellowship in the Evaluation and Implementation Science Group at the Centre for Health Policy at the University of Melbourne, Australia.

Cover/Layout Rio Basel, Switzerland
Print bc medien ag, Münchenstein, Switzerland
Printing of this thesis was financially supported by Swiss Paraplegic Research