Aus dem Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der Ludwig-Maximilians-Universität München

Lehrstuhl für Public Health und Versorgungsforschung

IBE Direktor: Prof. Dr. Ulrich Mansmann

Addressing the challenge of health measurement: The development of a metric of health to validly and reliably follow up the health of populations

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von

Cornelia Oberhauser

aus

München

2014

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. rer. biol. hum. Dipl.- Psych. Alarcos Cieza, MPH
Mitberichterstatter:	Prof. Dr. Eva Grill, MPH
	Priv.-Doz. Dr. Barbara Thorand
	Prof. Martha Merrow, PhD
Mitbetreuung durch den promovierten Mitarbeiter:	_____
Dekan:	Prof. Dr. med. Dr. h.c. M. Reiser, FACR, FRCR
Tag der mündlichen Prüfung:	10.03.2014

Danksagung

Hiermit möchte ich mich bei all denjenigen bedanken, die mich in den letzten Jahren begleitet, gefordert, gefördert und / oder unterstützt haben, und somit direkt oder indirekt zum Gelingen dieser Arbeit beigetragen haben:

- Alarcos Cieza, die mich stets sowohl gefordert als auch gefördert hat und mit der ich zusammen viele Herausforderungen gemeistert habe, und die aus einem ausschließlich in R programmierenden Statistiker einen SAS Programmierer und Psychometriker gemacht hat und mir damit viele neue Wege erschlossen hat. Vielen Dank für alles, was Du für mich getan hast!
- Meinem Team in Großhadern, mit dem ich über all die Jahre zusammenarbeiten durfte, und das über viele meiner Eigenheiten schmunzelnd hinweggesehen hat. Danke, dass Ihr mir ein zweites Zuhause gegeben habt!
- Darunter insbesondere Sara Hogger (ehemals Wadle) und Sarah Brockhaus (ehemals Maierhofer), die als studentische Hilfskräfte eng mit mir zusammengearbeitet haben, und die als Statistikerinnen dazu beigetragen haben, dass ich mir in einem Team aus lauter Nicht-Statistikern nicht ganz so exotisch vorgekommen bin. Vielen Dank für Eure Unterstützung!
- Ebenso Cristina Bostan, mit der ich über viele Jahre erfolgreich zusammengearbeitet habe und mit der ich viele fruchtbare Diskussionen führen durfte.
- Helmut Küchenhoff, mit dem ich zusammen zahlreiche Lehrveranstaltungen gemeistert habe, zwei Bücher bis zur Druckreife begleiten durfte, und der immer für fachliche Diskussionen zur Verfügung stand. Und der außerdem immer ein paar Gelder für mich übrig hatte, wenn es mit der Finanzierung meiner Stelle mal schwierig war. Vielen Dank für das Vertrauen, dass Du in mich gesetzt hast!
- Meinem Team im Statistik-Institut, mit dem ich zusammen zahlreiche Lehrveranstaltungen gehalten habe, und in dem sich immer ein Ansprechpartner für ausgefallene Statistik-Fragen gefunden hat. Vielen Dank für die Zusammenarbeit!
- Darunter insbesondere Monia Mahling, mit der ich viele Jahre ein Büro geteilt habe, zahlreiche Fortbildungen zur Verbesserung der Lehre besucht habe, und fast alle Lehrveranstaltungen zusammen gestaltet habe. Wir waren ein wirklich gutes Team!
- Allen meinen Studenten, die im Rahmen des Anfängerpraktikums unter meiner Betreuung viele der in Großhadern gesammelten Gesundheitsdaten deskriptiv analysiert haben, und mir auf diese Weise einen guten Einblick in die Vielfalt, aber auch die relativ konstanten Strukturen ermöglicht haben, selbst für die Daten, die ich letztendlich nie selbst weiter analysiert habe.
- Und zuletzt meiner Familie, die mich stets unterstützt hat, und meinem Freund, der mich auch bei größtem Stress wieder zum Lachen gebracht hat, für eine gemeinsame Vergangenheit und Zukunft. Danke!

Contents

List of abbreviations

ADL	Activities of Daily Living
ELSA	English Longitudinal Study of Ageing
GHS	German National Health Interview and Examination Survey
IADL	Instrumental Activities of Daily Living
ICF	International Classification of Functioning, Disability and Health
IRT	Item Response Theory
NHANES	United States National Health and Nutrition Examination Survey
PCM	Partial Credit Model
WHO	World Health Organization
WHS	WHO World Health Survey

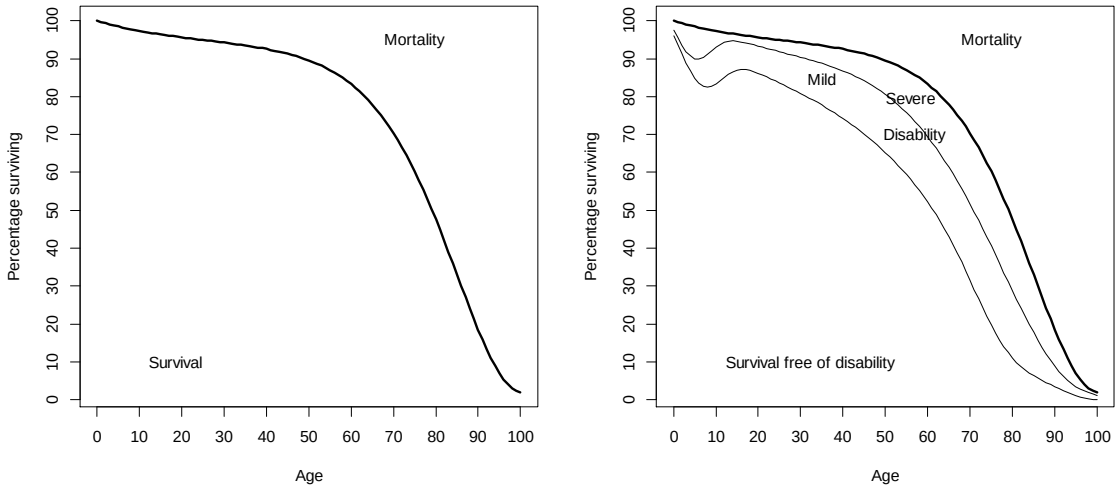
Background

Assessing the health of populations is important. Measuring health is essential to estimate the overall burden of ill health, to compare the relative impact of specific health problems [1], to monitor the effectiveness of health care [2] and, generally, to provide evidence for setting goals, implementing and monitoring health policy [3].
Different approaches have been implemented when collecting information about the health of populations. In general terms **two approaches** can be distinguished: 1) direct measures that rely on responses of a sample to survey questions, and 2) indirect measures that rely on summary statistics available only at population level.

Indirect measures assess and compare health only at the population level [4]. They are primarily used for health policy purposes and resource allocation. In contrast to the direct measures they do not directly rely on information collected from individuals but on existing information at the population level, such as mortality and morbidity statistics. These measures can be further subdivided into **two classes**: 1) health expectancies and 2) health gaps.

Health expectancy is a generic term for all population indicators that estimate the average time (in years) that a person could expect to live in various states of health [5]. Health expectancies relate to the area under the survival curve [4]. The survivorship function provides the proportion of survivors at each age for a specified group, e.g. a country's population. An example of a survivorship function for a hypothetical population is presented in Figure 1. The survivors can be further differentiated with regards to their health state, whether being in full health or in less than full health. Those in less than full health can be further differentiated based on the severity of their health states, e.g. mild and severe disability as visualized in the right part of Figure 1.

Figure 1: Survivorship function for a hypothetical population



Left: Survivorship function without further differentiation between survivors; Right: Survivorship function with survivors differentiated by severity of disability: no, mild and severe disability

In general terms, life expectancy is composed of the time lived in full health and the time lived at each age in a health state less than full health. Depending on the weight given to the component "less than full health", a different health expectancy measure is obtained. If a weight of 0 is used the resulting measure is called `disability-free life expectancy´, while for a weight of 1 the resulting measure is `life expectancy at birth´. In this case, only time lived is considered independently of the health level. If several states of less than full health are differentiated and weighted differently, `health-adjusted´ or `disability-adjusted life expectancies´ are obtained. Therefore, the prevalence of different health conditions is weighted based on their severity.

Measuring population health based on health expectancies has both advantages and disadvantages. Health expectancies are an easily interpretable measure for health comparisons, as they are measured in a well understood unit – years of living. They are appropriate to compare overall population health across communities and over time [6]. However, they do not provide information on the leading causes of death or non-fatal health status in populations [5]. It is not possible to disaggregate these measures in an additive fashion by cause or to identify the leading risk factors [5]. Therefore, they are not especially useful for health policy purposes.

Health gaps quantify the difference between the actual health of a population and some stated norm or goal [4]. Health gaps are composed as the sum of premature mortality and a proportion of the time lived at each age in a health

state less than ideal health. Depending on the survivorship norms used, four classes are distinguished [7]: 1) `potential years of life lost´; 2) `period expected years of life lost´; 3) `cohort expected years of life lost´; and 4) `standard expected years of life lost´. The normative survival goal of `potential years of life lost´ takes a constant value, with ages ranging from 60 to 85 being proposed. The goal for `period expected years of life lost´ is defined in relation to a period life table. In this kind of table life expectancy at each age is the estimated duration of life expected at each age, if the current age-specific mortality patterns were to hold in the future. The goal for `cohort expected years of life lost´ is defined in relation to cohort life expectancy, which is the estimated average duration of life a cohort would actually experience. Finally, the reference norm of the `standard expected years of life lost´ is defined in relation to a standard expectation of life at each age, of which the `disability-adjusted life years´ (DALYs) is the best known example [4]. It represents a combination of `years of life lost due to premature mortality´ (YLLs) and `years lived with disability´ (YLDs) [7].

Measuring population health based on health gaps also has both advantages and disadvantages. Health gaps are slightly more difficult to understand than health expectancies, but they permit to analyse the contribution of different causes to loss or gain in overall population health [8]. They are especially useful for policy decisions, as they permit to identify the leading causes and risk factors of disability and therefore implicitly indicate the most advantageous actions to be taken for improved overall health, as for example anti-smoking campaigns or immunization.

There is a need for these measures at the population level, especially for policy decisions. Since they use reference values of the whole population (i.e. mortality and morbidity statistics), the reference frame remains the entire population. That makes possible to carry out calculations that inform about the most beneficial interventions for the population of a country or the largest expected benefit. However, no information on health at the individual level can be obtained from these measures.

Direct measures of health rely on data directly collected from individuals without taking into account reference values of the entire population. Based on these data, comparisons at the individual or group level are possible but without having the reference of the entire population [2]. There exist **two main approaches**: One approach consists of measuring health based on a single response – using a single general health question. The other approach consists of assessing a profile of domains of health based on various questions on sub-domains, such as affect, pain, mobility and cognition.

The first approach – a **single general health question** – is **frequently used** both in population health surveys and also in clinical settings. A Medline search

for the terms “self-rated health” or “self-assessed health” or “self-reported health” or “self-perceived health” among publications from the year 2002 identified 1,991 reports [9]. Most commonly respondents were asked to rate their health on a five-point scale ranging from “very good” to “very bad” [10], or from “excellent” to “poor” [11]. Less frequently, scales with a different number of response options were used, or even continuous scales like a visual analogue scale [12].

Measuring health based on a single question has **several advantages**. A single question is intuitive and easy to assess. As it is based on self-report, it can be used in any data collection mode, whether direct interview, telephone interview or in writing. Furthermore, it constitutes a very economic approach to obtain information about the general health status of a sample. Finally, it was found to be a strong predictor of morbidity and mortality, even when accounting for socio-demographic characteristics and risk factors [13, 14].

However, this approach also has **several disadvantages**. Salomon et al. [15] have shown that a single general health question is not appropriate to compare health status over time. The authors undertook a comprehensive comparative analysis of self-rated health in four nationally representative longitudinal US surveys and observed widely discrepant results.
In addition, such a question seems to be inappropriate to compare health across populations, as official statistics suggest. Ratings of a five-point self-report general health questions were found to show an enormously high variation between European countries, which seems implausible, given other major health indicators [16]. A Eurostat publication from 1997 [17] reported large differences between the percentage of respondents rating their health with a specific response option for 12 countries in the European Union. Ratings of “very bad” or “bad” ranged from 5% in Ireland to 19% in Portugal, while a rating of “very good” ranged from 8% in Portugal to 53% in Denmark. More recently, an OECD publication [18] reported the percentage of adults rating their health as “very good” or “good” for 31 European countries in 2008, with similarly diverging results, ranging from 45% in Latvia over 65% in Germany to 87% in Switzerland. Given other major health indicators, it seems unlikely that real differences in the true health level of respondents are the only reason for these diverging patterns. More likely, also translational issues and cultural differences, different definitions of health, a different understanding of response options and different norms and expectations largely contribute to these differences [16].
Also, several qualitative studies found that respondents considered different aspects or dimensions when answering the question on self-assessed health [19-21]. These include 1) physical aspects, as chronic illness and physical problems, 2) a functional dimension, i.e. the extent to which they are able to perform, 3) a coping dimension, i.e. the extent to which they adapted to, or their attitude towards an existing illness, 4) a wellbeing dimension, i.e. the way they feel, and sometimes 5) a behavioural dimension, i.e. health behavior and lifestyle factors [19]. Interestingly, mental health, e.g. depression, anxiety or mental diseases, is hardly ever referred to. Also, the dimensions considered differed by sex and age

[19], which might in part explain why this question appears to be inappropriate for comparison purposes.

Despite these major shortcomings – that respondents consider different aspects for their responses and that both comparisons over time and across populations seem invalid – a single general health question continues to be used because it is available at very low cost. In addition, current practices recommend introducing health sections in multi-topic surveys by a self-rated health question to minimize potential order effects [13]. However, based on the evidence presented above, it seems that the general health question is not the optimal approach to assess health in surveys or studies.

The second approach for direct measurement - based on a **set of domains,** such as walking, self-care, memory and pain, - also constitutes a **frequently used approach** in health surveys. In fact, measuring health based on a set of domains is often combined with a single general health question, usually used as an introductory question to the health section.
Measuring health based on a set of domains is for example followed by World Health Organization (WHO) throughout all their surveys, as for example the WHO Multi-country Survey Study on Health and Responsiveness (2000-2001) [22] or the World Health Survey (2002-2004) [23]. The eight domains included are: mobility, self-care, pain and discomfort, cognition, interpersonal activities, vision, sleep and energy, and affect.
Also, this approach is implemented in many national surveys, e. g. in the German National Health Interview and Examination Survey from 1998 (GHS) [24] or in the United States National Health and Nutrition Examination Survey (NHANES) [25], and in studies on the aging population, e.g. the Health and Retirement Study (HRS) in the United States [26], the English Longitudinal Study of Aging (ELSA) [27] or the Survey of Health, Aging and Retirement in Europe (SHARE) [28].
Finally, this approach also constitutes the basis of several widely used instruments, as for example the MOS 36-item short-form health survey (SF-36) [29], or the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) [30].
As this approach is so frequently implemented, there seems to be an overall agreement of its usefulness.

Measuring health based on a set of domains also has **several advantages.** It provides more comprehensive information about the health of individuals. It results in a detailed profile of person's limitations, not only with regards to bodily impairments, as for example in hearing or cognition, but also with regards to limitations in activities as mobility and self-care, and restrictions in participation as in work or community life. Therefore, detailed information is obtained on where exactly individual's problems lie. In comparison to just listing the health conditions of a person, a profile on a set of domains implicitly informs about the

severity of these health conditions by transferring the impact of these health conditions into limitations in daily life.
Because of the more comprehensive information contained in this set of domains, it can be used for a wide range of applications. Individual needs can be identified, as for example for technical aids or personal assistance. For health policy purposes, this detailed level of information permits the planning, implementation and monitoring of more focused interventions.
Finally, as each domain pertains to a narrower construct, as for example mobility, the responses elicited by the questions will arguably refer to the same construct.

Several challenges have, however, also to be addressed when measuring health based on a set of domains:

According to WHO a selected set of domains **needs to fulfill** the following **criteria** [31]. The domains should be 1) valid in terms of intuitive, clinical and epidemiological concepts of health, 2) linked to the conceptual framework of the International Classification of Functioning, Disability and Health (ICF), 3) amenable to self-report, observation or measurements, 4) comprehensive enough to capture all important aspects of health states that people value, and 5) comparable across populations [31].

Also, this set of domains has to be **parsimonious** for practicability reasons [2, 31]. It must be as exhaustive as possible to contain a maximal amount of information – especially including the domains that are most important to people when assessing their overall health levels – , but as parsimonious as necessary to be practicably implemented – at low cost – in surveys and studies. Therefore, overlapping or even redundant domains should be excluded from the selected set of domains [2, 31].

In addition, measuring health based on a set of domains is **complex to standardize**, especially when using questions. First, a detailed instruction for measurement is needed for every domain. When questions are used, the exact wording of the question, the response options, and probably some framing (including the time frame and the context of the question) must be defined. Second, the measurement procedure must be implemented in a way that it can be used for any person all over the world in a comparable way. Hindrances to this might be cultural differences, e.g. men not doing any kind of housework in some cultures, or differences in the lived experience of persons, e.g. persons living on flat islands never climbing stairs throughout all their lives. Third, it must be possible to translate everything related to data collection in any language so that the same meaning is transferred in every language, which is especially relevant for questions and even more for the corresponding response options. So there is a lot of need for standardization with regards to the selected domains.

When using questions for data collection, an additional challenge must be addressed, namely that **response options** (even when transferring the same meaning) **might be differently interpreted by respondents**, as for example by males and females, persons from different age groups, or those from different countries. Even though each single domain, and especially each question constitutes a much narrower concept and therefore responses elicited by these questions will arguably refer to this construct, this does not imply that the same categorical response corresponds to the same amount of limitations in that domain for every respondent [32]. On the contrary, different individuals might use the categorical response scales in different ways, so that this response category cut-point shift might inhibit valid direct comparisons [31].

Finally, even if direct comparisons of responses on the separate domains were valid, as for example can be assumed for measured tests, this information would **only** make possible the **direct comparison for each very specific domain**, e.g. walking a certain distance. It would not permit, however, the direct comparison of the overall health status of individuals, neither between individuals nor for the same individuals over time. So it would not directly permit to say whether, on overall, individual A is healthier than individual B, or whether group C is healthier than group D, or whether population X is healthier than population Y. However, these kinds of comparisons are needed for several purposes, e.g. for analyzing health inequalities among and between subgroups of populations and for judging on whether one intervention is better than another one.

Three questions have, thus, still to be answered with regards to measuring and comparing the health of populations based on a set of domains.

The first open question refers to **which domains should be captured**, as at the moment there is a large variation between the domains used. These domains should be meaningful both for the general population and in clinical settings. However, for reasons of practicability, the selected set of domains should be extremely brief.

The second open question refers to **how these domains should be operationalized in a standardized way**, so that this operationalization can be integrated in any survey or study. In principle, health data can be collected based on questionnaires, through measured tests, and by observation. However, not all these measures can be obtained in any data collection mode, as for example based on computer-assisted telephone interviews or through self-administered postal forms, which might be used for general population surveys. This already limits the kinds of measures that potentially can be used. For the measures that in fact can be used in any data collection mode, an exact strategy for data collection must be defined. For example, for questions the exact wording of the question

itself, the response options and the phrasing of a possible introductory sentence to the question must be defined.

The next question is then **how to obtain a summary measure of overall health** based on a set of domains, especially in the context of potential response category cut-point shifts. With regards to a pre-selected minimal set of domains the third open question is **whether a summary measure based on a standardized minimal set of domains is useful**. Does it contain sufficient information to be reliable and valid? Is it sensitive to change, so that it can be used to monitor health over time? Is it in fact comparable across populations?

WHO has partly addressed these open questions.

With regards to the **selection of a set of domains** preparatory work towards the standardization of domains was done through the development of the **International Classification of Functioning, Disability and Health (ICF)** [33], and its endorsement in 2001.
The ICF constitutes the most sophisticated list of health and health-related domains ever developed. It provides a unified and standard language and framework for the description of health and health-related states.
It is composed of four components: body functions, body structures, activities and participation, and environmental factors. Within each component there is a hierarchy of so-called `ICF categories´, which form the most detailed units of functioning, e.g. emotional functions, sensation of pain, washing oneself or walking. In total, the classification contains 1424 ICF categories [34].
However, the ICF cannot directly be used for data collection purposes. It only contains the ICF categories and for each a short description, but no operationalization as needed for surveys or studies. In addition, due to its extremely large size, it cannot practically be applied. Thus, even though the ICF constitutes the standard language and framework to describe health and health related states, it does not advise on what are the most important domains for health measurement.

WHO has also proposed a **set of eight domains** to be considered for international health measurement and comparisons. This proposition was operationalized in the **World Health Survey (WHS)** [2]. The proposed domains are: **mobility**, **self-care**, **pain and discomfort**, **cognition**, **interpersonal activities**, **vision**, **sleep and energy**, and **affect** (see Appendix of the thesis for the wording of the questions) [35].
The development of this set of domains was based on extensive, sophisticated and multi-method studies carried out at WHO over a five year period [36, 37]. The development process began with an extensive review of the available items in common use in health and patient experience instruments [38], based on which a first draft was proposed. The instrument underwent several rounds of large scale international pilot testing based on general population surveys (specifically

designed for this purpose), followed by analysis, expert consultations, revision and reductions [23]. The WHO Multi-country Survey Study on Health and Responsiveness, carried out in 61 countries with a total of 188 307 respondents, served as the largest among these pilot studies [22, 39].
There has been no other international effort of this magnitude. At the same time, there have been no attempts outside of WHO to investigate the relevance of this set of domains across different populations, neither for the general population nor for subpopulations as for example specific clinical populations. Therefore, studies investigating the adequateness of this set of domains for health comparisons are needed for both kinds of populations. As the WHS domains so far constitute the most sophisticated proposition of domains of health, new proposals should be contrasted to them.

With regards to addressing the second open question – the **standardized operationalization of domains** – WHO also made a proposal for the eight domains included in the WHS. The questions on Health State Descriptions used in the WHS can be found in the Appendix of this thesis and their concrete operationalization including response options is available from the questionnaire [40]. There already exist further standardized question sets, as for example the short and extended question sets proposed by the Washington Group on Disability Statistics [41], or those used in instruments as the SF-36 [29] or the WHODAS 2.0 [30], from which questions for identified domains could be taken.

With regards to addressing the third question – **the creation of a summary measure** of overall health based on a set of domains – WHO examined two different strategies:

The first strategy relies on creating a metric of health using a **single-factor factor analysis**. This method assumes that one underlying latent construct, non-fatal health, exists, and that all preselected items contribute to it [42]. For comparison purposes, it is additionally hypothesized that the latent construct is similar across populations. As only a single factor is assumed, a person score can be obtained for each person based on the prediction from a multiple regression model, in which the different variables on health constitute the independent variables and the factor loadings for each of these variables serve as regression coefficients [42].
This method was applied to the functioning data from the WHS to compare mean health scores across disease states and demographic variables [43]. However, some years before, Sadana et al. [16, 42] had created such a health score based on individual data from 64 household interview surveys with nationally representative samples of the non-institutionalized civilian population from 46 countries. The kind and number of questions included in the analysis varied considerably across countries. The authors attempted to equalize the health scales across populations by transforming each scale on a range from 0 (worst

health) to 100 (best health). They concluded that valid comparison of existing data from household interview surveys across countries is limited, as biases in the self-report of health status prevent a meaningful comparison of non-fatal health status across populations. They stated that this even applies to surveys in which the survey methodologies and data collection approaches had been standardized.
So it is doubtful whether factor analysis constitutes the best approach to create a metric of health based on data collected for a set of domains.

The second strategy for creating a metric of health consists of applying a **unidimensional Item Response Theory (IRT) model** to functioning data from health surveys [44]. Unidimensional IRT models assume that there is a unidimensional latent construct to be measured, i.e. that a scale exists on which both persons and items can be located [45]. Information about the latent construct is obtained based on persons' responses to items, e.g. questions with yes-no or ordered response options from a questionnaire. The metric of health is defined by the locations each person is assigned to at this latent scale.
IRT methods are increasingly used in the literature for general population surveys, especially for the WHS. For this data source the health metric was originally developed based on a special IRT model – the Rasch Rating Scale model – for the World Report on Disability [1], with some technical details provided in the Appendix [35]. Thereafter, Hosseinpoor et al. [46, 47] used this score to investigate health differences between men and women in the context of other socio-demographic factors, and Chatterji et al. [48] used this score to compare the health of two populations, for China and India. As IRT models permit to detect systematic shifts in persons' interpretation of response options and provide a simple means to correct for it, they constitute the preferred method to create a metric of health that is cross-population comparable.
The resulting metric can then be evaluated with regards to reliability, validity and, if longitudinal data is available, sensitivity to change. In addition, it can be used to compare the health of populations.

The assessment, monitoring and comparison of health of populations is recognized as one of the most important open questions in health sciences [49, 50]. Especially for health policy purposes, i.e. for monitoring and evaluation, it is necessary to have an instrument at hand that is comparable over time, across communities within a population, and across populations [51]. Only when being able to compare the health of individuals, subgroups of populations, or populations as a whole, and over time, it is possible to evaluate the efficacy of health systems, to judge on the effectiveness of health interventions and to identify discrepancies within or between subgroups of populations [52].

In this doctoral thesis I propose an approach to address the open questions related to measuring health based on a set of domains:

First, I will identify a core set of domains that can be argued to be appropriate to capture health. The proposed set of domains will not only be based on general population data, as it has been used for the selection of the WHS domains, but also on data from a large clinical study. This set will be presented and discussed using the WHS domains as a reference. In addition, a brief proposition on where to look for an operationalization of the selected domains will be made.

Second, I will use data on this identified set of domains to investigate whether it can be used to develop a reliable and valid metric of health, and whether this metric proves sensitive to change and can therefore be used to monitor the health of populations over time.

If this metric – based on a brief set of preselected domains – proves useful, i.e. if it has sound psychometric properties, it has a large potential for future use for comparisons across populations.

Research Objectives

The overall objective of this doctoral thesis is to develop a valid, reliable and sensitive metric of health that permits to monitor the health of populations over time, and which provides the basis for the comparisons of health across different populations.

The specific aims are:

1) To identify a minimal generic set of domains suitable for measuring health both in the general population and in clinical populations, and

2) To determine whether the information collected in these domains can be integrated in a sound psychometric measure, based on which the health of populations can be assessed and monitored over time.

To achieve these two specific aims, I carried out two psychometric studies that will be presented in the following sections. The first is entitled “Towards a Minimal Generic Set of Domains” and the second “Development of a metric of health”.

Towards a Minimal Generic Set of Domains

Objective and specific aims

The objective of this study is to identify a minimal generic set of domains of functioning suitable for measuring health, both at the individual and population levels.

The specific aims are, first to determine whether the domains of functioning of the WHS are relevant for both clinical populations and the general population; and second, as clinical populations were not considered for the selection of the WHS domains, to determine whether additional domains need to be added to the WHS to facilitate comparability across clinical populations.

Methods

The most advanced proposition for the domains to be considered for international health measurement and comparisons was made by WHO for the **World Health Survey (WHS)** [2]. The eight domains included are: **mobility**, **self-care**, **pain and discomfort**, **cognition**, **interpersonal activities**, **vision**, **sleep and energy**, and **affect** (see Appendix for the wording of the questions) [35]. As mentioned before, the development of this set of domains was based on extensive, sophisticated and multi-method studies carried out at WHO over a five year period [36, 37], with several rounds of pilot testing based on general population data. As the WHS domains so far constitute the most advanced proposition of domains to be used, they will serve as a reference for the minimal generic set of domains to be developed in this study.

Data

This is a psychometric study using data from **three sources**, two national general population surveys and one large scale survey with clinical populations. In this study, both population based data and clinical data is used, because content valid domains must be applicable to both kinds of populations, and even more importantly to clinical ones. In addition, clinical data was not used in the development process of the WHS domains, nor data from nationally fielded general population surveys. These two weaknesses are therefore compensated here.

The three data sources used are: 1) the German National Health Interview and Examination Survey 1998 (GHS); 2) the United States National Health and Nutrition Examination Survey 2007/2008 (NHANES); and 3) the ICF Core Set studies.

The **GHS** was the first German Health Survey covering the former East and West Germany together in one survey. It was carried out between 1997 and 1999 by the Robert Koch Institute, the central federal institution responsible for disease control and prevention in Germany. The data available for public use includes information of 7124 adults from a representative sample of the residential population in Germany [24]. The GHS data are available on demand for scientific purposes from the Robert Koch Institute [53].

The **NHANES** is a survey of the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention designed to assess the health and nutritional status of adults and children in the United States [25]. The data used for this study are from 6228 persons aged 18 years and older from the 2007-2008 wave. NHANES data are openly available from the corresponding webpage [54].

The **ICF Core Set studies** are a series of 22 studies carried out at the ICF Research Branch of the WHO Collaborating Centre for the Family of International Classifications in Germany from 2004 to 2010 in collaboration with institutions in 44 countries in clinical settings ranging from early post-acute over primary care to rehabilitation [34]. Each of these studies aimed at developing or validating a so-called `ICF Core Set´ for a specific setting or health condition.
An ICF Core Set is a selection of ICF categories, i.e. a comparably short list compared to the complete classification, that was judged to be relevant for a specific condition based on expert consensus, taking into account evidence from further studies. These studies included literature review, qualitative patient interviews or focus groups, an expert survey and results from structured patient interviews.
Each study considered here involved a clinical population with one of the following health conditions as the main diagnosis: ankylosing spondylitis, breast cancer, chronic widespread pain, depression, diabetes mellitus, head and neck cancer, chronic ischemic heart disease, hand conditions, low back pain, multiple sclerosis, osteoarthritis, obesity, osteoporosis, obstructive pulmonary disease, rheumatoid arthritis, spinal cord injury (early post-acute and chronic), sleep disorders, stroke, traumatic brain injury, and low vision. Data in the vocational rehabilitation setting with persons with different diagnoses was also collected.
In all but one study the data on health and functioning was recorded using the ICF qualifiers [33], i.e. 0=no problem, 1=mild problem, 2=moderate problem, 3=severe problem, and 4=complete problem. Only in the study on hand conditions the data was collected using a visual analogue scale with values ranging from 0 to 100. Due to the different coding scheme, this data was not comparable to the other datasets and was therefore not considered for the analyses. In total, the data of 9863 persons were available.
Even though the ICF Core Sets include ICF categories of all the components of the ICF, only categories of the components of functioning (body functions and structures, activities and participation) were further considered. The data on the ICF Core Set studies is also publicly available upon request from the ICF Research Branch [55].

Preprocessing

To make the data from all three sources comparable, the health and health-related questions contained in the GHS and NHANES datasets were linked to the categories of the ICF by established linking rules [56]. For example, the variable "DPQ040 [Over the last 2 weeks, how often have you been bothered by the following problems:] feeling tired or having little energy?" of the NHANES was linked to the ICF category `b130 Energy and drive functions´. Only the data of those questions that could be unequivocally linked to a single ICF category of the components body functions and structures, activities and participation were further considered. The data from the ICF Core Set studies had directly been collected using the language of the ICF. Therefore, no linking was necessary for these studies.

At this stage in the study, the number of variables from each data source available for further steps was: 1) 25 questions from the GHS; 2) 28 questions from NHANES; and 3) 204 ICF categories from the ICF Core Set studies.

To make sure that all relevant, and only relevant variables are included in the analyses, the next step consisted of selecting variables using information sources as filters:

1) the questions used in the WHS to address its eight domains [35, 40],
2) the 17 questions used in the Washington City Group extended set [41]
3) the questions contained in 3 out of the 6 most commonly used health status measures [57], and
4) the ICF categories of the dimension functioning common in at least 11 of the 22 ICF Core Sets.

The questions of 1) and 2) were also linked to the ICF using the same rules as for the GHS and NHANES. The questions of 3) had already been linked to the ICF in a previous work [57]. For 3) and 4) a 50% cut-off was used, since it captures the majority of relevant ICF categories. However, any cut-off threshold is in a sense arbitrary. As both the variables from each data source and the filters were expressed in the standard language of the ICF, the selection of variables could be performed using the filters. Variables related to at least one of these four sets were considered for further analysis. This selection resulted in 14 variables from GHS, 20 from NHANES and 56 from the ICF Core Set studies.

Analysis

Descriptive statistics were used to characterize the study populations of all three data sources in terms of age, gender, and percent of people living alone.
Regression methodologies were applied using the self-reported general health question common to the ICF Core Set studies and the two surveys -- "In general, would you say your health is (excellent / very good / good / fair / poor)?" -- as dependent variable. In accordance with previous investigations the response options were transformed as follows: excellent = 5.0, very good = 4.4, good = 3.4, fair = 2.0, and poor = 1.0 [58, 59]. The variables linked to the ICF and preselected by the application of the filters served as independent variables.

Two regression methodologies were applied for the sake of robustness – Random Forests and Group Lasso regression [60-63] – to the data from the ICF Core Set studies, the GHS and the NHANES and separately for the ICF categories contained in the ICF components of body functions and structures, and activities and participation.
Random Forests is a non-parametric regression technique that can be used to obtain a rank of the explanatory relevance of the independent variables, based on a so-called variable importance measure assigned to each independent variable [64]. **Group Lasso** regression is a parametric regression technique that allows for the selection of the ordinal independent variables that explain most of the variance of a dependent variable by taking their ordinal structure into account. Group Lasso can also be used to rank independent variables according to their level of explanatory relevance, defined through the maximal size of the penalty for which the variable is first selected into the model [65, 66].
ICF categories are designated as relevant independent variables when they rank among the top 50% in both regression methodologies for at least one data source.

The results from these two methods, i.e. the finally identified domains based on the three data sources, were compared to the domains of the WHS, as these constitute the most advanced selection of domains available so far and therefore form an appropriate reference. A WHS domain was considered valid for both, the general and clinical populations when ICF categories addressing this domain were above the 50% cut off in both clinical and general population. An ICF category above the 50% cut off in the clinical population is proposed to be added to the WHS domains when functioning and health is assessed in clinical populations.

The descriptive statistics, the Random Forests and the Group Lasso regression were performed with R version 2.11.1 [67].

Results

The number of cases for which the dependent variable was available consisted of 6224 in the GHS, 4436 in the NHANES, and 9264 in the ICF Core Set studies. The age, gender and percentage of persons living alone in all three samples are presented in Table 1.

Table 1: Demographics of the study populations of the three datasets used for the regression analyses

	GHS (n=6224)	**NHANES** (n=4436)	**ICF Core Set studies** (n=9264)
Males %	48.6	48.8	44.6
Age: years mean (sd)	45.8 (15.9)	48.5 (17.3)	53.1 (15.9)
Living alone %	29.6	12.5	18.7

The identified functioning-related variables from GHS, NHANES and the ICF Core Set studies are listed as ICF categories and organized by the components of the ICF in Tables 2 and 3 across the three data sets. Table 2 contains the results regarding body functions and structures, while Table 3 contains those related to activities and participation. For each data set there are two columns, one for each of the two regression methodologies, containing the ranking obtained from the respective method. Where the ranking is missing, no variable related to the specific ICF category could be identified in the dataset. The ICF categories most associated with the self-report of health are those with the highest ranks across the different data sets. The smaller the associated number, the more relevant is the variable, i.e. a rank of 1 identifies the most important variable. Within each study, ICF categories above the 50% cut off are marked in bold.

Using the 50% cut off for both methodologies within each data source, 10 ICF body functions and 18 activity and participation ICF categories were identified as most associated with self-reported general health.

Table 2: List of ICF body functions categories from the GHS, the NHANES and the ICF Core Set studies datasets included in the analyses, rank order resulted from Random Forest and Group Lasso indicating the level of association with the general health question, cut off rank for the different datasets. Those categories with a rank below or equal to the cut-off point for both regression methodologies in at least one dataset were considered confirmed and selected for comparison with the World Health Survey domains of functioning.

ICF code	**Title**	**GHS***		**NHANES***		**ICF Core Set studies***	
		Random Forest	**Group Lasso**	**Random Forest**	**Group Lasso**	**Random Forest**	**Group Lasso**
b126	Temperament and personality functions	8	8			13	**9.5**
b130	Energy and drive functions	**2**	**3**	**2**	**1**	**6**	**4.5**
b134	Sleep functions	7	**5**	4	**2**	**3**	**2**
b140	Attention functions			6	6	15	17
b144	Memory functions			5	5	17	19
b152	Emotional functions	**4**	6	**3**	**3**	**5**	**6**
b180	Experience of self and time functions					19	15.5
b210	Seeing functions	**5**	**4**			16	14
b230	Hearing functions	6	7	**1**	4	18	18
b280	Sensation of pain	**1**	**1**			**1**	**1**
b455	Exercise tolerance functions					**2**	**4.5**
b530	Weight maintenance functions	9	9			11	11
b640	Sexual functions					**7**	**8**
b710	Mobility of joint functions					**8**	**7**
b730	Muscle power functions					**4**	**3**
b740	Muscle endurance functions					**10**	15.5
b780	Sensations related to muscles and movement functions	**3**	**2**			**9**	12
s750	Structure of lower extremity					14	13
s760	Structure of trunk					12	**9.5**
Cut off point (top 50% of ranking)		5	5	3	3	10	10

* The ICF categories containing a rank number in these columns were included in the analyses with data of this study

Table 3: List of ICF activities and participation categories from the GHS, the NHANES and the ICF Core Set studies datasets included in the analyses, rank order resulted from Random Forest and Group Lasso indicating the level of association with the general health question, cut off rank for the different datasets. Those categories with a rank below or equal to the cut off point for both regressions methodologies in at least one dataset were considered confirmed and selected for comparison with the World Health Survey domains of functioning.

ICF code	**Title**	**GHS***		**NHANES***		**ICF Core Set studies***	
		Random Forest	**Group Lasso**	**Random Forest**	**Group Lasso**	**Random Forest**	**Group Lasso**
d110	Watching			**1**	**2**	36	35.5
d115	Listening					37	35.5
d160	Focusing attention					33	31
d175	Solving problems					31	**15.5**
d230	Carrying out daily routine	**1**	**1**			**14**	**18**
d240	Handling stress and other psychological demands					**3**	**7**
d310	Communicating with - receiving - spoken messages					30	19.5
d335	Producing nonverbal messages					35	35.5
d410	Changing basic body position	**2**	**3**	**7**	**5**	**16**	31
d415	Maintaining a body position			**4**	**3**	23	31
d430	Lifting and carrying objects	4	5	**5**	8	**19**	19.5
d440	Fine hand use			9	12	28	22
d445	Hand and arm use			**6**	**4**	27	22
d450	Walking	5	4	**3**	**6**	**8**	**5**
d455	Moving around	**3**	**2**	11	9	**6**	**3**
d465	Moving around using equipment					29	25.5
d470	Using transportation					**13**	**12**
d475	Driving					33	**13.5**
d510	Washing oneself					**2**	**4**
d520	Caring for body parts					20	35.5
d530	Toileting					25	31
d540	Dressing			12	11	**5**	**6**

d550	Eating			14	13.5	26	27.5
d570	Looking after one's health					**11**	**9**
d620	Acquisition of goods and services					22	24
d630	Preparing meals			13	13.5	**18**	27.5
d640	Doing housework			10	10	**4**	**2**
d660	Assisting others					**8**	**8**
d710	Basic interpersonal interactions					**10**	**17**
d760	Family relationships					21	**13.5**
d770	Intimate relationships					**12**	**10**
d830	Higher education					32	25.5
d845	Acquiring, keeping and terminating a job					**17**	22
d850	Remunerative employment			**2**	**1**	**15**	**11**
d870	Economic self-sufficiency					24	**15.5**
d910	Community life					**7**	31
d920	Recreation and leisure			8	**7**	**1**	**1**
Cut off point (top 50% of ranking)		3	3	7	7	19	19

* The ICF categories containing a rank number in these columns were included in the analyses with data of this study

In Table 4 these 28 ICF categories are rearranged in three sections: those ICF categories considered valid A) for both types of populations, B) only for the general population, and C) only for the clinical population. Each section is arranged by the 8 WHS domains of functioning, linked with the specific ICF categories.

Section A of Table 4 shows, which domains of the WHS are considered valid for both the clinical population and the general population (mobility, pain and discomfort, sleep and energy, and affect). The table also shows the specific ICF categories that confirm those WHS domains: d450 Walking, d455 Moving around, b280 Sensation of pain, b130 Energy and drive functions, and b152 Emotional functions.

Section A of Table 4 also shows that `d230 Carrying out daily routine´ and `d850 Remunerative employment´ are relevant to self-perceived health in both general and clinical populations. The five above mentioned ICF categories and these two are proposed to make up the minimal generic set of ICF categories suitable for describing functioning both at the individual and population levels.

Section B of Table 4 gives the WHS domains only relevant for the general population. It shows that the WHS domain of vision has been confirmed for the general population based on the ICF categories of `b210 Seeing functions´ and `d110 Watching´. It also shows the ICF categories that confirmed the relevance of the WHS domain of mobility for the general population alone.
Section C of Table 4 gives the WHS domains only relevant for the clinical populations. It shows the ICF categories that confirmed the WHS domains of mobility, self-care, interpersonal activities, and sleep and energy for clinical populations. In addition, five ICF categories not contained in the WHS domains were identified as relevant to self-perceived health in clinical populations: b640 Sexual functions, d770 Intimate relationships, d240 Handling stress and other psychological demands, d640 Doing housework, and d660 Assisting others.

Table 4: WHS domains of functioning and ICF categories found explanatory for self-perceived health

WHS domains of functioning	**Specific ICF Categories**		**GHS**	**NHANES**	**ICF Core Set studies**
	ICF Code	Title			
Section A: ICF categories found explanatory for self-perceived health both in the general and clinical population studies					
Mobility	d450	Walking	-	✓	✓
	d455	Moving around	✓	-	✓
Self Care					
Pain and Discomfort	b280	Sensation of pain	✓		✓
Cognition					
Interpersonal Activities					
Vision					
Sleep and Energy	b130	Energy and drive functions	✓	✓	✓
Affect	b152	Emotional functions	-	✓	✓
	d230	Carrying out daily routine	✓		✓
	d850	Remunerative employment		✓	✓
Section B: ICF categories found explanatory for self-perceived health only in the general population studies					
Mobility	b780	Sensations related to muscles and movement functions	✓		-
	d410	Changing basic body position	✓	✓	-
	d415	Maintaining a body position		✓	-
	d445	Hand and arm use		✓	-
Self Care					
Pain and Discomfort					
Cognition					
Interpersonal Activities					
Vision	b210	Seeing functions	✓		-
	d110	Watching		✓	-
Sleep and Energy					
Affect					

Section C: ICF categories found explanatory for self-perceived health only in the clinical population studies					
Mobility	b455	Exercise tolerance functions			✓
	b710	Mobility of joint functions			✓
	b730	Muscle power functions			✓
	d470	Using transportation			✓
Self Care	d510	Washing oneself			✓
	d540	Dressing		-	✓
	d570	Looking after one's health			✓
Pain and Discomfort					
Cognition					
Interpersonal Activities	d710	Basic interpersonal interactions			✓
	d920	Recreation and leisure		-	✓
Vision					
Sleep and Energy	b134	Sleep functions	-	-	✓
Affect					
	b640	Sexual functions			✓
	d770	Intimate relationships			✓
	d240	Handling stress and other psychological demands			✓
	d640	Doing housework		-	✓
	d660	Assisting others			✓

Legend: ✓ means that data on the ICF category were available and the ICF category was confirmed for the corresponding dataset. - means that data on the category were available but the ICF category was not confirmed based on the 50% cut off criterion for the corresponding dataset. Space means that no data on the category were available for the corresponding dataset. Empty lines mean that for the corresponding WHS domain no ICF category could be confirmed by the corresponding combination of datasets.

Discussion

This study has proposed the following set of ICF categories as a minimal generic set of functioning and health:

b130	Energy and drive functions
b152	Emotional functions
b280	Sensation of pain
d230	Carrying out daily routine
d450	Walking
d455	Moving around
d850	Remunerative employment

Based on the criteria of relevance used in this study, four of the eight domains of functioning of the WHS were sufficiently explanatory for self-perceived health both in the general and in clinical populations. The other WHS domains not represented in the proposed minimal generic set are vision, which was only confirmed with data of the general population, self-care and interpersonal activities, which were only confirmed with data of the clinical population and cognition, which could not be confirmed at all.
The ICF categories of carrying out daily routine and remunerative employment also fulfilled the inclusion criteria, though not related to any of the eight WHS domains. However, the WHS questionnaire on Health State Descriptions (see Appendix) is introduced through an `Overall Health´ section [40], containing the general health question and a question on difficulty with work and household activities. The latter's content is closely related to both remunerative employment and carrying out daily routine.

The construction of a minimal generic set requires hard decisions and there will always be good reasons for and against each proposed ICF category. In this study, ICF categories were selected based on statistical evidence involving a large international clinical sample and two national general population samples. So a lot of evidence was provided for the selection of relevant ICF categories. The non-inclusion of ICF categories related to the WHS domains of vision, cognition, self-care and interpersonal activities might partly be explained by the inclusion of carrying out daily routine, for which vision and cognition are a prerequisite, and of which self-care and interpersonal activities form an integral part. So excluding any ICF category must not be interpreted as saying that the ICF category is irrelevant.

The proposed minimal generic set of ICF categories of functioning and health can always be augmented for specific applications. This study provides some evidence for the decision about what other ICF categories to add. As shown in Table 4 (section B), in general population studies additional mobility ICF categories can be included. Also, the inclusion of ICF categories for vision or watching is recommended.

An additional set of ICF categories is also proposed for clinical populations, as shown in section C. It contains additional ICF categories related to the WHS domains confirmed, i.e. more detailed ICF categories on mobility, and sleep functions. In addition, it contains ICF categories related to self-care and interpersonal activities, which are WHS domains that were not confirmed based on the criteria used within this study. Finally, five ICF categories not related to any of the WHS domains were identified.

The ICF categories identified as relevant for clinical populations are, to an enormous extent, in agreement with results from a previous study [59]. Therein, the authors applied a complex, multi-stage selection process involving linear regression on ICF data from a clinical sample of 1039 German patients with 12 different chronic health conditions, with the same general health question as dependent variable.
For the ICF components of body functions and structures, they identified energy and drive function, emotional functions, sensation of pain and muscle power functions in accordance with the findings presented here. In addition, they identified vestibular functions, i.e. sensory functions of the inner ear related to position, balance and movement according to the ICF [33], which did not fulfill the filter criteria applied in this study.
For the ICF component of activities and participation, the authors identified walking, remunerative employment, recreation and leisure, doing housework and assisting others in accordance with the findings presented here. In addition, they only identified acquisition of goods and services, for which the obtained ranking was just slightly above the 50% cut-off. Therefore, the ICF categories identified in this study for clinical populations proved valid and can be recommended for further use in clinical settings.

When designing a disability survey, countries can also take advantage of the results of this study. Section C of Table 4 presents those ICF categories relevant exclusively for persons with health conditions, who experience disability or who are at risk of doing so. Disability surveys usually target these persons with the objective of describing their problems or their needs in different areas of life. It is always difficult to decide which relevant domains will help to achieve that objective. A recent comparison of over 100 disability surveys showed that, despite some attempts at harmonization [41], disability surveys are extremely diverse in the domains they address [68]. The set of ICF categories presented in Section C of Table 4 can be seen as a proposal of ICF categories relevant to capture disability. This proposal has been taken into account in a current project conducted by the WHO and the World Bank to develop a Model Disability Survey. All those categories of the minimal generic set as well as those that might be called the “disability set” are captured in the model disability survey.

To ensure a wide applicability of the minimal generic set, its implementation should be amenable to different data collection modes. In clinical settings, these include patient interviews conducted by health professionals, and self-

administered forms. In general population surveys these are face-to-face interviews, computer-assisted telephone interviews and postal self-administered forms, as have been used in the WHO Multi-country Survey Study [39].

Therefore, the minimal generic set should be operationalized with self-report questions. For the 4 WHS domains of mobility, pain and discomfort, sleep and energy, and affect, the WHS itself provides public-domain questions that have been extensively and psychometrically studied [22] and widely used around the world [43, 46, 48, 69-71]. For the operationalization of the two additional categories, `d230 Carrying out daily routine´ and `d850 Remunerative employment´, there also exist good candidate questions from the many widely used health status measures that have already been linked to the ICF [72].

There are several limitations of this study. The general population data used came from high-resource western countries, which are not representative of the general population worldwide. This fact affected the choice of 'remunerative employment' rather than the more general term 'work'. As well, the data comes from the adult, non-institutionalized population and might have been different if children and institutionalized populations were included. Data from many questions and ICF categories came exclusively from clinical populations rather than the general population. Therefore, it is not sure that the same ICF categories would have been found as highly explanatory for both the general and clinical population if more general population data had been included. Relying on the self-reported general health question as the only dependent variable may also be a limitation since, in the literature, implausible response patterns where identified across countries [17, 18, 42]. However, in this study both the general health question and the questions linked to the ICF were answered by the same person and are therefore likely to be exposed to the same `cultural´ bias, so that an analysis of the relationship between the two remains valid. In addition, self-rated general health questions have been shown to be strong predictors of functioning and disability and are sensitive to the full spectrum of health conditions [13].

The WHO group responsible for the selection of the WHS domains guided their work according to five criteria [2]: These domains must be 1) valid in terms of intuitive, clinical, and epidemiological concepts of health; 2) linked to the conceptual framework of the ICF; 3) amenable to self-report, observation, or direct measurement; 4) comprehensive enough to capture the most important aspects of health states that people value; and 5) comparable across populations.
The process implemented here was guided by these criteria as well. The seven ICF categories of the proposed minimal generic set can be assumed to satisfy the first three criteria. The next essential step for future research would be to identify the extent to which these ICF categories satisfy the last two criteria, namely capturing the aspects of health that people value and being cross-population comparable.

These two criteria are essential for the next and most important challenge yet to be resolved in health assessment, namely, to develop a common metric of health to link information from the general population to information about sub-populations, such as clinical and institutional populations. Such a metric would be useful for assessing and comparing levels and patterns in the functional trajectory of a person's life, and thus permit to compare the health of populations and to analyse trends in population health.

The minimal generic set proposed in this study is the starting point to address one of the most important challenges in health measurement, namely the comparability of data across time, studies and countries. It also represents the first step for developing a common metric of health to link information from the general population to information about sub-populations, such as clinical and institutional populations.

Development of a metric of health

Objective and specific aims

In the previous study, a **minimal generic set** of functioning and health was proposed. The domains of the minimal generic set are: 1) **energy and drive functions**, 2) **emotional functions**, 3) **sensation of pain**, 4) **carrying out daily routine**, 5) **walking and moving around** and 6) **remunerative employment**. One important question left open is whether the information collected in these domains can be integrated in a sound psychometric measure, based on which the health of populations can be assessed and monitored over time.

The specific aims are to evaluate the psychometric properties of the health metric, including 1) internal consistency reliability, 2) construct validity and 3) sensitivity to change, which implies that the health metric can be used to track the health of populations over time.

Methods

Data

Data from the **English Longitudinal Study of Ageing (ELSA)** was used for the analysis. ELSA is a biannual, longitudinal and nationally representative survey that focuses on adults aged 50 and over, living in private households in England [27]. Their partners are also interviewed, irrespective of age. Therefore, the sample also contains information on a small number of persons aged less than 50 years.
More concretely, data from waves 3 and 4 collected in 2006/07 and 2008/09 were analysed (N=9779 and 11050). Depending on the needs for each specific aim, the **combined data**, **wave 4 data**, or the **overlapping data** (i.e. persons on whom data was available for both wave 3 and 4) was used. In wave 4 a new cohort entered the sample and several wave 3 members were no longer available, leading to an overlapping sample of 7908 persons.

Table 5 shows the questions of ELSA that operationalize the domains of the minimal generic set and were, therefore, selected for the construction of the metric.

Table 5: Questions used to operationalize the six domains of the minimal generic set

Domain	**Question**
Energy and drive functions	(Much of the time during the past week), you felt that everything you did was an effort?
	(Much of the time during the past week), you could not get going?
	Here is a list of statements that people have used to describe their lives or how they feel. How often, do you feel like this? - I feel full of energy these days
Emotional functions	(Much of the time during the past week), you felt depressed?
	(Much of the time during the past week), you felt sad?
	(Much of the time during the past week), you were happy?
Sensation of pain	Are you often troubled with pain?
	How bad is the pain most of the time? Is it mild, moderate, or, severe?
Carrying out daily routine	*Please tell me if any difficulty with these because of a physical, mental, emotional or memory problem. Again exclude any difficulties you expect to last less than three months. Because of a health or memory problem, have difficulty doing any of the activities on this card? -*
	Dressing, including putting on shoes and socks
	Bathing or showering
	Eating, such as cutting up food
	Getting in or out of bed
	Using the toilet, including getting up or down
	Using a map to figure out how to get around in a strange place
	Preparing a hot meal
	Shopping for groceries
	Taking medications
	Doing work around the house or garden
	Managing money, such as paying bills and keeping track of expenses
Walking and moving around	By and without using any special equipment, how much difficulty do you have walking for a quarter of a mile?
	Because of a physical or health problem, do you have difficulty doing any of the activities on this card? Exclude any difficulties that you expect to last less than three months. - Walking 100 yards.
Remunerative employment	Do you have any health problem or disability that limits the kind or amount of paid work you could do, should you want to?

For energy and drive functions, emotional functions, walking and moving around, and remunerative employment the questions directly reflect the content from the respective domains. Therefore, the original variables were used in the analysis. For sensation of pain, the first question ("often troubled with pain") served as a filter for the second (severity of the pain). This means that the second question was only asked if the first was answered "yes". Therefore, these two questions were summarized into one variable with response options "not often troubled with pain", "mild", "moderate" and "severe" pain.
For carrying out daily routine a different strategy was needed, as the single variables alone did not reflect that domain. Therefore, two sum scores were created: one for activities of daily living (ADLs) – including dressing, washing, eating, getting in and out of bed and using the toilet – with values indicating none to five limitations and the other one for instrumental activities of daily living (IADLs) – including difficulty using a map, preparing a hot meal, shopping for groceries, taking medications, doing housework and managing money – with values indicating none to six limitations.
The response options for these selected variables were coded or recoded so that higher values indicated worse health.

Analysis

Descriptive statistics

Descriptive statistics were used to characterize the study population. They are presented for the complete wave 3 sample, the complete wave 4 sample, and their overlap.

Development of the health metric

To develop the metric of health the **Partial Credit Model (PCM)** is applied [45, 73]. The PCM or Polytomous Rasch Model is a unidimensional Item Response Theory (IRT) Model that can be applied to a set of ordinal, polytomous items [74]. Unidimensional IRT models assume that there is a unidimensional latent construct to be measured, i.e. that a scale exists on which both persons and items can be located. Information about the latent construct is obtained based on persons' responses to items, e.g. questions with yes-no or ordered response options from a questionnaire.

Based on the model, information is obtained both for persons and items. For each person the so-called **person ability** is obtained, i.e. the location of the person on the scale. For each item the so-called **item location** is obtained, i.e. the overall difficulty of the item on the same scale. In addition, **item thresholds** are available for each item. For an item with k response options, there are k-1 thresholds. These indicate the location on the latent trait where the item best discriminates between persons. At each item threshold, the probability of a person with this ability is defined to be 0.5 to have a response below or above the corresponding threshold. Persons with higher ability are more likely to give a

response above the threshold, while persons with lower ability are more likely to give a response below the threshold.
In IRT models an additional parameter exists for items, i.e. the **item discrimination**. The higher the discrimination parameter for an item, the better it discriminates between persons, especially between persons with abilities close to its thresholds. In the PCM, the discrimination parameter is fixed to one for all items. As a consequence, all the items in the model are considered equally important and equally contribute to the scale.

Before the PCM was applied, **model assumptions** were evaluated: unidimensionality, local independency and monotonicity.

Unidimensionality means that a person's response to an item that measures a construct is accounted for by his/her level on that trait, and not by other factors [75]. It was probed with bifactor analysis [76-78]. Bifactor analysis assumes the presence of a single general factor and multiple independent group factors. If all items load high on the general factor, and the factor loadings on the general factor exceed those of the group factors, an underlying unidimensional latent trait can be assumed. The number of factors considered in the bifactor analysis was determined based on permuted parallel analysis [79]. Based on this method, the number of factors is defined as the number of eigenvalues resulting from the observed data exceeding the 95% quantile of the eigenvalues resulting from several permutations of the observed data.
Bifactor analysis was applied on the polychoric correlation matrix [80, 81]. The polychoric correlation coefficient is a measure of association between two ordinal variables. It is based on the assumption of an underlying joint continuous distribution of the two variables. Categories of the two ordinal variables correspond to intervals of the respective continuous variables. The polychoric correlation coefficient then constitutes a measure of the correlation between these two underlying continuous variables.

Local independence means that there should be no significant association among item responses after the dominant factor influencing a person's response to an item was controlled for [75]. It was examined based on the residual correlations among items resulting from a single-factor factor analysis [82]. The PCM was then estimated with and without the flagged possible local dependent items (residual correlations higher than 0.2) to see if results were robust to questions' dependencies [83]. If the item thresholds fundamentally change when considering local dependent items in the same model, all but one of them needs to be excluded.

Monotonicity means that the probability of selecting an item response indicating higher ability on the latent trait (here better health) should increase as the underlying person's level of ability (here person's health) increases [75]. It was studied for each item by examining graphs of the item's distribution

conditional on mean "rest-scores", calculated for each person as the total raw score of all the remaining non-missing items divided by their number. Usually, in the case of non-missing data, the "rest-score" (i.e. the total raw score minus the item score) is used [75]. In the case of missing data, the mean "rest-score" as described above can be obtained for all persons (even when data on some items is missing) and is a less biased measure than would be the "rest-score" calculated based on the non-missing data for each person. If there is a consistent trend that persons with higher mean rest-scores are more likely to have more problems in the selected item, monotonicity can be assumed. Items violating the monotonicity assumption need to be excluded from the model.

After the evaluation of the model assumptions, the PCM was fitted. In case of unordered thresholds, the response options of the affected items were collapsed until all thresholds were in the correct order. To examine whether persons from different groups with the same (latent) health level have a different probability of giving a certain response to an item, **differential item functioning (DIF)** was tested for gender and age groups (<=64 and >64) using iterative hybrid ordinal logistic regression with change in McFadden's pseudo R-squared measure (above 0.02) as DIF criterion [84, 85]. For items showing DIF, the item must be split into two separate items for the two groups and the model re-estimated. For the final PCM, item locations and item thresholds are presented. Furthermore, the persons' health level is presented on the same scale.
Finally, persons' health level was linearly transformed to a (**health**) **scale** ranging from 0 (worst health) to 100 (best health), as on this scale differences are easier to interpret. Based on this new scale it is easier to judge on the relevance of differences between groups, e.g. with and without health conditions, and change over time. This is especially relevant as due to the large sample size statistical significance (e.g. p-values below 0.05) does not necessarily indicate meaningful differences.

For this health scale the following **psychometric properties** were evaluated: 1) internal consistency reliability, 2) construct validity and 3) sensitivity to change.

Internal consistency reliability

Reliability in research can be interpreted as repeatability or consistency [86]. **Internal consistency reliability** is the type of reliability that is estimated to assess the consistency of results across items within a test. In contrast to other measures of reliability (inter-rater, test-retest, parallel-forms) it is based on the data obtained with a single instrument from a group of people at one time point.

In this study, internal consistency reliability was assessed based on different measures. **Inter-item correlation** [86] indicates the strength of the correlation

between items. **Item-to-total correlation** [86] indicates the strength of the correlation of each item with the total score, represented here by the mean score of non-missing items for each person (following the same strategy as was already done for monotonicity). Polychoric correlations were used for both these types of correlation.

Cronbach's alpha [87] corresponds to the average of all possible split-half estimates of reliability, which would be obtained from dividing the items into two sets each containing half of the items, and then calculating the correlation between the two total scores. Cronbach's alpha is then the average of all possible resulting correlations and this way indicates in how far the total scores from two randomly created subsets of items are expected to be correlated.

Cronbach's alpha is the most commonly used measure for internal consistency reliability. However, two additional measures are recommended to be provided: McDonald's omega hierarchical and McDonald's omega total [88]. **McDonald's omega hierarchical** [89] measures the general factor saturation in bi-factor analysis, thereby providing the proportion of test variance due to the general factor. This indicates the extent to which total scores can be generalized to the latent variable common to all test items [88]. **McDonald's omega total** estimates the proportion of test variance due to all common factors [90].

All these measures can range from 0 to 1, with higher values indicating higher reliability.

Construct validity

Validity is related to generalizing [86]. **Construct validity** involves generalizing from a measure to the concept of this measure, or expressed the other way round, translating any construct into its operationalization. Construct validity can be interpreted as the approximate truth of the conclusion that an operationalization or measure accurately reflects its construct [86]. Here the question in case is whether the so-called health score in fact constitutes a measure of health.

Construct validity can be assessed based on **four different criteria** [86]:

Convergent validity is understood as the degree to which a measure is similar to other operationalizations it theoretically should be similar to. It is analysed based on the **Spearman correlation** of the health scale with other health-related variables, as the general health question and a question on long-standing limiting illnesses. A high correlation indicates high convergent validity.

Discriminant validity is understood as the degree to which a measure is not similar to other operationalizations that it theoretically should not be similar to. It is analysed based on the **Spearman correlation** of the health scale with less

health-related variables, as life satisfaction, the number of falls in the last year, and age. A low correlation indicates high discriminant validity.

Concurrent validity is defined as the measure's ability to distinguish between groups that it should theoretically be able to distinguish between. It is assessed based on a **linear additive model** [91], which predicts the value on the health scale based on sex, age, education, income and health conditions as independent variables. Age is modeled in a flexible, non-parametric way using P-splines. Concurrent validity can be judged as high if persons with health conditions have lower expected levels of health compared to those without the respective health condition, and persons with severe health conditions on average have lower expected health levels than those with very mild health conditions.

Predictive validity is defined as the measure's ability to predict something it should theoretically be able to predict. It is analysed based on predicting mortality in 2008 to 2012 based on wave 4 data. For this purpose four **additive logit-models** [91] are compared, each containing the covariates sex, age, education, income and health conditions (as above). Model 1 contains only these independent variables. Model 2 in addition contains the health scale, while model 3 in addition contains the general health question. Model 4 contains all the covariates, and both the general health question and the health scale. Where contained, age and the health scale are modeled in a flexible, non-parametric way using P-splines. For all these models different model fit criteria are compared: the adjusted R-square, the percentage of explained deviance, and the Akaike information criterion (AIC). If the inclusion of the health scale improves model fit this indicates predictive validity.

Sensitivity to change

Sensitivity to change is the ability of a measure to detect changes over time, such as an improvement or deterioration in the health state of a person [92]. The measure must detect meaningful change when it has occurred, and it must remain stable when no change has occurred. Sensitivity to change was evaluated in two ways for the subsample on which data was available for both wave 3 and 4.

First, the change in the health scale between the two waves was compared to the 'change' in the responses to the general health question in the two waves. Change in the health scale was defined as the difference between the values from the two waves, and calculated as the value of the health scale in wave 4 minus the value in wave 3. Therefore, positive differences indicate improvement in health, while negative differences indicate deterioration. The distribution of these differences was visualized through **boxplots** for each combination of responses to the general health question in the two waves. Unfortunately, the response options of the general health question differed in the two waves (ranging from "very good" to "very bad" in wave 3, and from "excellent" to "poor" in wave 4), which slightly

complicated the comparison. Sensitivity to change is high if both measures (the general health question and the health scale) show the same tendencies, i.e. both should be stable, indicate improved health or deterioration.

Second, a **linear additive model** [91] was fitted with the value of the health scale in wave 4 as dependent variable and new incidence of health conditions since wave 3 as independent variables, while controlling for the value of the health scale in wave 3 and additional covariates. If the incidence of severe health conditions has a high negative impact on the expected value of the health scale, while the incidence of less severe health conditions has a smaller effect, the health scale shows high sensitivity to change.

The complete analyses was performed with R version 2.15.2 [93].

Results

Descriptive statistics

Descriptive statistics of the study population are provided in Table 6 for wave 3 and 4 data and their overlap. The response options for the general health question differed for the two waves, leading to a very different response pattern.

Development of the health metric

When testing the **IRT model assumptions** for the combined dataset, permuted parallel analysis indicated the presence of two factors. The bifactor analysis showed high factor loadings on the general factor for all items (ranging from 0.55 to 0.85), all of which exceeded the factor loadings of the group factors, supporting the assumption of unidimensionality. High residual correlations were found for "feeling everything was an effort" and "could not get going" in the domain energy and drive functions, and for feeling "depressed", "sad" or being "happy" in emotional functions. When keeping only one of the local dependent variables for each domain ("feeling everything was an effort" and "depressed"), sensitivity analyses showed similar thresholds compared to the model with all items included (Pearson correlation of 0.99). This indicates that all items can be kept in the final model. Monotonicity was graphically confirmed by all items.

When fitting the PCM on the combined dataset the thresholds of four items were disordered and had to be collapsed: For pain and walking a quarter of a mile "mild" and "moderate" problems were collapsed. For the two scores on ADL and IADL the response options one and two limitations were collapsed, and three and more. None of the items showed DIF by gender or age based on the selected DIF criterion.

Table 6: Descriptive statistics of wave 3 and wave 4 data, and their overlap

	Wave 3 (N=9779)	**Wave 4** (N=11050)	**Overlap of wave 3 and 4 - Wave 4 values** (N=7908)
Age: mean (median)	64.56 (63)	65.24 (64)	66.40 (65)
Gender: female (%)	0.56	0.55	0.57
Education: low (%) #	-	0.42	0.42
Education: medium (%) #	-	0.27	0.27
Education: high (%) #	-	0.31	0.31
Income: low (%) +	-	0.31	0.32
Income: medium (%) +	-	0.33	0.33
Income: high (%) +	-	0.36	0.35
General health			
w3: very good / w4: excellent (%)	0.26	0.13	0.12
w3: good / w4: very good (%)	0.43	0.29	0.29
w3: fair / w4: good (%)	0.24	0.32	0.33
w3: bad / w4: fair (%)	0.06	0.19	0.19
w3: very bad / w4: poor (%)	0.01	0.07	0.07

The education division is from a level lower than "O-level" or equivalent (typically 0-11 years of schooling), qualified to a level lower than "A-level" or equivalent (typically 12-13 years), and a higher qualification (typically >13 years).

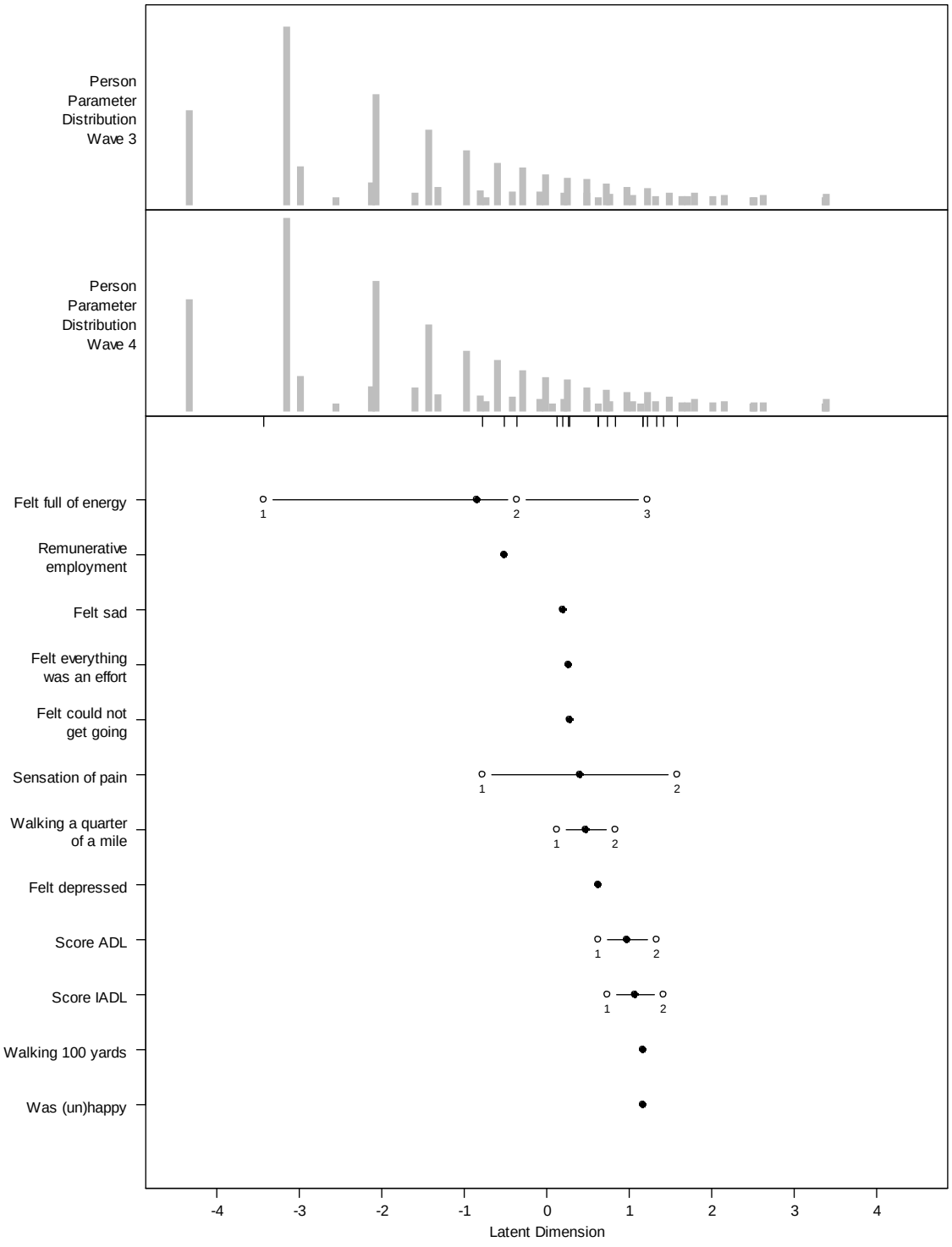
+ Income groups were formed by dividing equivalised total income into three approximately equally sized groups based on the sample.

- Information on education and income was incomplete for wave 3 data and is therefore not included in the table.

w3 and w4 are abbreviations for wave 3 and wave 4, respectively.

The results for the **final PCM** are visualized in the person-item map in Figure 2. In the top part of the figure, the distribution of persons' health levels are shown separately for wave 3 and wave 4. The pattern of persons' levels is quite similar in the two waves, with values ranging from -4.33 to 4.21. Item locations and item thresholds are presented in the bottom part of the figure. Item locations (bullets) range from -0.85 to 1.17, while item thresholds (circles) range from -3.43 to 1.58. The items are well suited to differentiate between persons' levels in the medium range of health. They do however not well differentiate between the large proportion of very healthy persons (to the left), nor for the small proportion of extremely unhealthy persons (to the right). To facilitate the comparison of item thresholds with persons' ability, the item thresholds are additionally plotted below the persons' distribution (of wave 4) by small vertical lines.

Figure 2: Person-item map for the final PCM: The top part displays the distribution of persons' health levels separately for wave 3 and wave 4. The bottom part shows the item locations (bullets) and item thresholds (circles).



Internal consistency reliability

Table 7 shows the results on **internal consistency reliability**. The values of the different measures yield consistent results when calculated for each of the two waves separately and for the combined dataset. Inter-item correlation is high, but has high variability. Item-to-total correlation is higher, with less variation. Cronbach's alpha and McDonald's omega total are quite close to 1. McDonald's omega hierarchical is lower (with values around 0.60), but of reasonable size for general factor saturation. Therefore, all values indicate high internal consistency reliability.

Table 7: Results on internal consistency reliability

Measure	**Wave 3**	**Wave 4**	**Wave 3 and 4 combined**
Inter-item correlation: mean [min; max]	0.54 [0.23; 0.90]	0.53 [0.24; 0.92]	0.53 [0.25; 0.91]
Item-to-total correlation: mean [min; max]	0.76 [0.61; 0.84]	0.75 [0.59; 0.85]	0.75 [0.60; 0.84]
Cronbach's alpha	0.93	0.93	0.93
McDonald's omega hierarchical	0.60	0.61	0.61
McDonald's omega total	0.96	0.95	0.96

Construct validity

Table 8 shows the results on **convergent** and **discriminant validity**. It contains the Spearman correlations for wave 4 data between the following six variables: 1) the health scale, 2) the self-rated general health question, 3) a variable on long-standing limiting illness, disability or infirmity (response options: "no", "yes, but not limiting", "yes and limiting"), 4) life satisfaction (with seven response options), 5) the number of falls within the last year, and 6) age.

Table 8: Spearman correlation matrix for the health scale and additional health-related variables

	Health scale	General health	Long-standing illness	Life satisfaction	Number of falls	Age
Health scale	1.00	0.64	0.59	0.36	0.25	0.23
General health	0.64	1.00	0.56	0.26	0.17	0.19
Long-standing illness	0.59	0.56	1.00	0.18	0.19	0.19
Life satisfaction	0.36	0.26	0.18	1.00	0.08	-0.06
Number of falls	0.25	0.17	0.19	0.08	1.00	0.10
Age	0.23	0.19	0.19	-0.06	0.10	1.00

The correlation of the health scale with general health and the question on long-standing illness is comparably high (0.64 and 0.59), indicating high convergent validity. Both these correlations exceed the correlation between general health and longstanding illness (0.56). The correlation of the health scale with life satisfaction is lower, and lowest for the number of falls and age. These concepts are distinct from health, covering mostly only very specific aspects of it. Therefore, their low correlations with the health scale indicate high discriminant validity.

The linear additive model indicates high **concurrent validity**. Table 9 and Figure 3 present the results from the linear additive model on wave 4 data with the health scale as dependent variable and sex, age, education, income and health conditions as independent variables. As expected, all the listed health conditions have a negative effect on the health score. The health conditions with the largest negative impact on health are dementia, Parkinson's disease, heart failure, arthritis and lung disease, followed by psychiatric conditions and stroke. The lowest effect on health is observed for high cholesterol, heart attack and heart murmur. This ordering corresponds to the subjective severity of these health conditions and is, where comparison is possible, in agreement with the disability weights estimated for 220 unique health states within the Global Burden of Disease study 2010 [94].

Table 9: Regression coefficients, standard errors (SE) and p-values resulting from the linear additive model predicting the value of the health scale for wave 4 data. For the health conditions, the number of cases (N) having the respective health condition is provided.

	N	Coefficient	SE	p-value
Intercept		73.61	0.44	<0.0001
Gender (female)		-0.56	0.34	0.0965
Education (middle)		3.47	0.41	<0.0001
Education (high)		4.67	0.41	<0.0001
Income (middle)		2.08	0.40	<0.0001
Income (high)		5.43	0.42	<0.0001
High cholesterol	3546	-0.58	0.36	0.1108
High blood pressure	4214	-2.44	0.35	<0.0001
Angina	885	-3.31	0.80	<0.0001
Heart attack	741	-1.25	0.86	0.1459
Heart failure	65	-12.38	2.17	<0.0001
Heart murmur	423	-1.26	0.84	0.1336
Abnormal heart rhythm	820	-2.97	0.63	<0.0001
Other heart disease	303	-3.97	1.02	<0.0001
Diabetes	1063	-6.36	0.56	<0.0001
Stroke	481	-8.44	0.81	<0.0001
Lung disease	544	-11.06	0.76	<0.0001

Asthma	1260	-3.35	0.51	<0.0001
Arthritis	3816	-11.09	0.35	<0.0001
Osteoporosis	753	-7.46	0.65	<0.0001
Cancer	571	-3.45	0.72	<0.0001
Parkinson‘s disease	79	-19.20	1.98	<0.0001
Psychiatric condition	971	-9.92	0.57	<0.0001
Dementia	154	-19.44	1.65	<0.0001

The reference categories are male, low education, low income, and not having the respective health condition.

Figure 3 shows the nonlinear effect of age on the health score together with its 95% credible intervals. As there are only a small number of observations below an age of 50, there is a lot of uncertainty in the estimation in this range. From an age of 68 on, age has an increasingly negative impact on health levels.

Figure 3: Nonlinear effect of age (solid line) for wave 4 resulting from the linear additive model and 95% credible intervals (dashed lines)

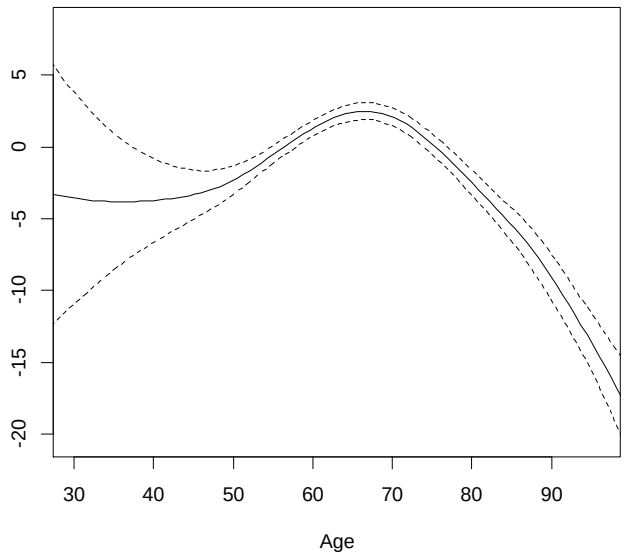


Table 10 shows the results regarding the **predictive validity** of the health scale. For wave 4 data, four different additive logit-models predicting mortality are compared based on three model fit criteria. To permit a fair comparison of criteria, the same subset of data with complete responses in all the variables considered over the four models was used. This particularly meant that all cases with missing responses to the general health question were excluded.

Table 10: Comparison of model fit criteria for four different models

	Adjusted R-square	**Percentage of deviance explained**	**AIC**
Model 1 including only covariates	15.2	23.6	3362
Model 2 including covariates and the health metric	17.5	26.2	3251
Model 3 including covariates and the general health question	16.7	25.5	3285
Model 4 including covariates, the general health question and the health metric	17.9	26.6	3240

When comparing model 1 and 2, which differ by the inclusion of the health scale as independent variable in model 2, all three criteria indicate higher model fit for model 2 (higher adjusted R-square, higher percentage of deviance explained, smaller AIC). This means that the health scale contributes to model fit in addition to all the other covariates, which already contain a list of health conditions.

When comparing model 2 and 3, which differ by the inclusion of the health scale in model 2 and the general health question in model 3, all three criteria indicate higher model fit for model 2. This means that adding the health scale to the covariates improves the model more than adding the general health question instead. This means that the health scale is more appropriate for predicting mortality than the general health question.

When comparing model 4 with models 2 and 3, model fit increases more when adding the health scale to the model with the general health question already included compared to vice versa. Adding the general health question to the model with the health scale already included only minimally improves model fit.

These results indicate that the health scale has a higher predictive validity for mortality than the general health question, and that it still has predictive validity when included in addition to the general health question.

In fact, the health metric is even more superior to the general health question than shown in the table, as the general health question cannot be assessed based on proxy interviews, while the value of the health metric can be deduced based on proxy's responses to objective limitations, like those in self-care or mobility. These persons are typically assigned a bad or very bad level of health based on the metric, and correctly classified with high risk of mortality, causing additional increase in model fit.

Sensitivity to change

Sensitivity to change was analysed based on two methods. Figure 4 shows **boxplots** of the differences in the health scale for each combination of responses to the general health question in wave 3 and 4. Positive differences in the health scale indicate improvement in health, while negative differences indicate deterioration. When the same response (in wording) to the general health question was provided for both waves (e.g. very good-very good, good-good, etc.), the difference in the health scale on average is close to 0. This also applies for slight improvement in wording (e.g. very good-excellent, good-very good, etc.), which might be explained by the fact that in respondents' ratings not only the wording of the response option, but also its rank among the alternatives might be taken into account, and that therefore providing the equally ranked alternative among the response options (e.g. the second) might in fact reflect similar health status. When the general health question indicated larger improvement in health (e.g. from fair to very good), the health scale on average also indicated improvement. The same applied to deterioration in health. This means that on average the expected trends were observed, indicating high sensitivity to change. However, the difference in the health scale showed a high variation.

Figure 4: Boxplots of the differences in the health scale for each combination of responses to the general health question in wave 3 and wave 4. In order to additionally provide information about the amount of uncertainty related to each boxplot, information on the sample size in each group, i.e. for each combination of responses to the general health question, is integrated. The width of each box is proportional to the square-root of the number of observations in each group, i.e. the wider the box, the higher is the corresponding sample size.

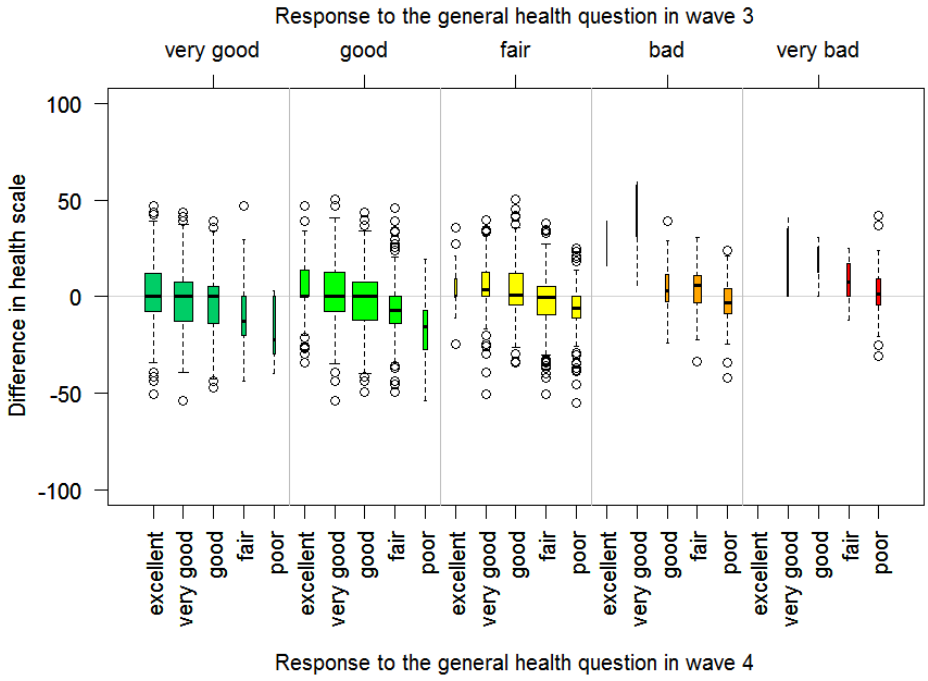


Table 11 and Figure 5 show the results from the **linear additive model** predicting the value of the health scale in wave 4 based on the incidence of health conditions within the last two years, when controlling for the value of the health scale in wave 3 and other covariates. Table 11 shows that the incidence of any of these health conditions has a negative effect on the value of the health scale. Dementia has the highest negative impact, followed by heart failure, psychiatric conditions, Parkinson's disease, lung disease, stroke, cancer and arthritis. High cholesterol, angina and heart attack have the smallest effect. As the health conditions with large effects are severe, often without effective therapy and fast progressing, while those with small effects are mild or with effective therapy, these results indicate high sensitivity to change. Figure 5 shows the nonlinear effect of age from this model. However, age was only used as a control variable here.

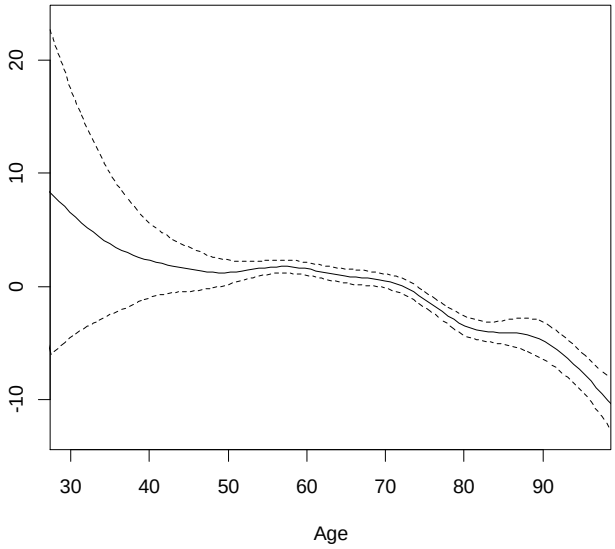
Table 11: Regression coefficients, standard errors (SE) and p-values resulting from the linear additive model predicting the value of the health scale in wave 4 based on the incidence of health conditions within the last two years, when controlling for the value of the health scale in wave 3 and other covariates. For the health conditions, the number of cases (N) with incidence in the last two years is provided.

	N	Coefficient	SE	p-value
Intercept		18.98	0.65	<0.0001
Health scale in wave 3		0.71	0.01	<0.0001
Gender (female)		-0.86	0.30	0.0047
Education (middle)		1.42	0.38	0.0002
Education (high)		1.22	0.38	0.0012
Income (middle)		0.86	0.37	0.0200
Income (high)		1.79	0.39	<0.0001
Incidence of:				
High cholesterol	503	-0.57	0.61	0.3563
High blood pressure	325	-1.46	0.75	0.0502
Angina	180	-0.73	1.00	0.4655
Heart attack	265	-0.94	0.83	0.2571
Heart failure	14	-9.62	3.90	0.0138
Heart murmur	57	-3.06	1.73	0.0771
Abnormal heart rhythm	137	-2.27	1.15	0.0481
Other heart disease	126	-1.94	1.19	0.1019
Diabetes	138	-2.95	1.13	0.0093
Stroke	91	-4.73	1.41	0.0008
Lung disease	91	-5.90	1.39	<0.0001
Asthma	91	-2.72	1,40	0.0514
Arthritis	361	-4.06	0.71	<0.0001
Osteoporosis	121	-1.17	1.21	0.3323
Cancer	138	-4.15	1.12	0.0002

Parkinson's disease	13	-6.51	3.56	0.0679
Psychiatric condition	110	-8.00	1.29	<0.0001
Dementia	66	-16.62	1.96	<0.0001

The reference categories are male, low education, low income, and no incidence of the respective health condition within the last two years.

Figure 5: Nonlinear effect of age (solid line) resulting from the linear additive model and 95% credible intervals (dashed lines)



Discussion

In this study, a **metric of health** was developed. This metric was based on a small set of questions from the domains of the minimal generic set: energy and drive functions, emotional functions, sensation of pain, carrying out daily routine, mobility and remunerative employment. It showed high internal consistency reliability, high construct validity and high sensitivity to change. Therefore, it can be considered an appropriate measure of population health.

In this study, the **ELSA** data was used for several reasons. It constituted one of very few general population datasets containing questions to operationalize all six domains from the minimal generic set. Also, the data was publicly available. In addition to the questions on functioning it contained information on socio-economic status, health related variables (e.g. life satisfaction and the number of falls) and detailed information on health conditions. Finally, longitudinal data was available, which permitted to create the health metric for two waves, so that comparisons over time were possible. Because of all these properties, the ELSA data enabled all the analyses necessary to examine the psychometric properties of the developed health metric.

The **applicability of the PCM** on the data was examined in a rigorous manner. All three model assumptions of IRT analysis, i.e. unidimensionality, local independency and monotonicity, were formally investigated and the results provided. Results from testing IRT model assumptions are commonly reported for clinical studies [75, 95, 96], but very rarely for general population surveys.

The developed metric **differentiated** well between persons in the medium range of health, and less precisely at the lower and upper end.
Its creation constitutes a trade-off between two conflicting aims: On the one hand, maximal measurement precision is desired to permit fine distinctions in persons' health levels and this way to increase its potential uses, both for group comparisons and comparisons over time. On the other hand, the metric should be based on an extremely parsimonious set of domains that can be integrated at low cost, both with regards to financial resources and (interviewing) time, into any already existing or newly developed survey or questionnaire. Only if the metric is based on a parsimonious set of domains, its domains will be frequently implemented in studies. This in turn constitutes the prerequisite for the subsequent application of the metric for comparison purposes. Therefore, loss in measurement precision, especially at the margins of the continuum, must be accepted for reasons of practicability.
In addition, the ability to differentiate between persons at the margins of the continuum is not as informative as differentiating between them in the medium range. Persons in the medium range of health are at risk for further deterioration, while little change in health states are expected for the extreme groups. The very unhealthy persons are already (known to be) in need of care, while the very healthy persons are unlikely to need extensive health services resources within the next few years. So monitoring changes in the medium range is more meaningful for both health care providers and health policy.
However, this does not imply that only these proposed domains should be used in any study. For specific purposes, e.g. in clinical studies, additional domains should be added, as have been proposed in the previous study.

The health metric showed **sound psychometric properties**. The absolute values of the measures used to examine internal consistency reliability were all sufficiently large, especially when compared to those resulting from other datasets [88]. The assessment of construct validity completely relied on the comparison of statistics (Spearman correlations, regression coefficients and model fit criteria), without the need of absolute thresholds to judge on them. The same applied to the methods used to demonstrate sensitivity to change.

Besides the results already mentioned, the **validity** of the developed metric is reinforced by additional findings. As can be seen from the linear additive model to evaluate concurrent validity (Table 9), in addition to the plausible effects of health conditions on the health metric [94], all of the well-known gradients of health – age, education and income levels – are captured by the health metric

[47]. The results show that from an age of 68 on, the expected level of health dramatically decreases, and that higher levels of education and income are associated with higher levels of health. Also, the well-documented but variable differential in health between men and women is captured by the metric [47, 97], with women being slightly worse off than men, even when several covariates are taken into account.

In the analyses and results presented **two different kinds of comparisons** were implicitly illustrated: comparing the health of groups of populations 1) at the same time point and 2) over time.
With regards to the comparison between groups **at the same time point**, subgroups of the sample can implicitly be compared based on the results from the linear additive model used to examine concurrent validity. For example, females were on average less healthy by 0.55 points compared to men. Persons with high education were on average 4.67 points healthier compared to persons with low education, and persons with high income were on average 5.43 points healthier than persons with low income. Also, person groups with different health conditions could be compared.
With regards to observing health **over time**, the distribution of health levels for the – overlapping, but different – complete wave 3 and wave 4 samples is visualized in the top part of the person-item map (Figure 2), and can thereby be compared. Differences in the values of the health metric for the overlapping sample, i.e. changes in health between the two waves, are visualized in the boxplots used to examine sensitivity to change. Finally, in the linear additive model used to examine sensitivity to change, health status in wave 3 is used to explain health status in wave 4, when taking the incidence of specific health conditions into account.

These two kinds of health comparisons already allow for a **wide range of possible applications**: monitoring changes over time in the health status of a given population; evaluating the effectiveness of health interventions; identifying and quantifying overall health inequalities between subgroups of populations, e.g. across different ethnic groups; and providing appropriate and balanced attention to the effects of non-fatal health outcomes on overall population health [98].

There are some **limitations** of this study. Only one exemplary dataset, ELSA, was analysed. England is a high-resource western country and not representative for the general population worldwide. In addition, the focus of ELSA was on persons aged 50 and above, and not the general population without age restrictions. In general population health surveys children and institutionalized populations are often not covered, which can be considered a general limitation of the methodology applied with regards to the generalizability of results.

As the analyses were restricted to a single data source, also the investigated set of questions was limited. If slightly different content was asked in the questions, or different response options were used, results might differ. Therefore, the results shown need to be validated in further studies, with data from different populations with regards to country, age group, and setting.

The health metric developed in this study – based on the domains of the minimal generic set – proved useful for a wide range of health comparisons, especially for groups of persons, and both at one point in time and over time. Monitoring health over time is especially informative both for health care providers and health policy, and both in clinical settings and in the general population.

Discussion

The **overall objective** of this doctoral thesis was to develop a valid, reliable and sensitive metric of health that permits to monitor the health of populations over time, and which provides the basis for the comparisons of health across different populations.

To achieve this objective, two consecutive psychometric studies were carried out.

In the **first study**, a minimal generic set of domains suitable for measuring health both in the general population and in clinical populations was identified. This set of domains was presented and discussed using the WHS domains as a reference. In addition, a brief proposition on where to look for an operationalization of the selected domains was made.

In the **second study**, information collected on the domains from the minimal generic set were integrated in a sound psychometric measure, based on which the health of populations can be assessed. It was shown that this measure constitutes a reliable and valid metric of health, and that it proves sensitive to change and can therefore be used to monitor the health of populations over time.

The **significance** of the developed health metric arises in the light of decreasing mortality rates all over the world [99], leading to increased life expectancies [6] and thereby to a steadily growing older population. These changes in the aging structure of populations create the need to monitor ongoing changes in the health of populations, in order to predict health services needs, as for example for institutional care, and to allocate resources for them.
The health metric has the potential to address these needs by producing accurate estimates of individuals' health, which can be aggregated to any group level, including a country's total population. These estimates permit to monitor health over time for individuals and subgroups of populations, including those in clinical and institutionalized settings, or a country's total population.

However, the **largest potential** of the developed approach lies in the **comparison** of the health **of two different populations**. These populations can be two general population samples, two clinical populations, or a clinical and a general population sample.

When two general population samples, e.g. from two different countries, are compared – in combination with longitudinal data – this provides essential input for the evaluation of the performance of different health systems, along with information on health inequalities, responsiveness, and fairness in financing [98]. Comparative judgements also provide the dependent variable in analyses of the

independent variables that contribute to health differences across populations [98]. Furthermore, such comparisons might inform debates on priorities for health service delivery and planning, and also for research and development, improve curricula for professional training in public health, and help to analyse the benefits of health interventions in studies on cost-effectiveness [98].

The comparison of two clinical populations permits to analyse the effect of two different treatments on the overall health level. It also enables to differentiate patients with the same health condition with regards to their severity. In contrast to condition specific instruments, the health metric provides a sound picture of the overall health level – especially for persons with comorbidities.

Finally, based on the comparison of a clinical and a general population sample it is possible to obtain standardized reference values for the clinical sample based on the general population data. In addition, as general population surveys typically exclude institutionalized persons, measuring health for both parts of the population on the same scale permits to obtain a summary measure for the complete population of a country. This in turn permits fairer comparisons between the health levels of different countries, even if the health care system, and therefore the degree of institutionalization, is extremely different.

This comparison of health across populations can be performed by applying the **same methods** as used in the second study. The only prerequisite is that a sufficient amount of similar questions with similar response options were used in both studies. I already carried out a psychometric study to compare the health of the English with that of the Americans using the data from the English Longitudinal Study of Ageing (ELSA) and the Health and Retirement Study (HRS) in the United States and all the health domains comparable across these two studies [100].

In the context of this doctoral thesis the developed metric of health proved useful to compare the health of groups of persons at the same time point, and to follow up the health of populations over time.
These two kinds of health comparisons already allow for a wide range of possible applications: monitoring changes over time in the health status of a given population; evaluating the effectiveness of health interventions; identifying and quantifying overall health inequalities between subgroups of populations, e.g. across different ethnic groups; and providing appropriate and balanced attention to the effects of non-fatal health outcomes on overall population health [98].

Conclusion

The minimal generic set proposed in the first study is the starting point to address one of the most important challenges in health measurement, namely the comparability of data across studies and countries. It also represents the first step for developing a common metric of health to link information from the general population to information about sub-populations, such as clinical and institutional populations.

The health metric developed in the second study – based on the domains of the minimal generic set – proved useful for a wide range of health comparisons, especially for groups of persons, and both at one point in time and over time. Monitoring health over time is especially informative both for health care providers and health policy, and both in clinical settings and in the general population.

This health metric can likely not only be used to monitor health over time, but also to compare health across populations and even countries. This provides essential input for the evaluation of the performance of different health systems, along with information on health inequalities, responsiveness, and fairness in financing [98]. When the same health metric is obtained both for the general population and for institutionalized populations, an overall summary measure can be obtained for the total population of a country, which permits fairer comparisons across countries with different degrees of institutionalization.

Therefore, the developed health metric can be seen as the starting point for a wide range of health comparisons, between individuals, groups of persons and populations as a whole, and both at one point in time and over time. It opens up a wide range of possible applications for both health care providers and health policy, and both in clinical settings and in the general population.

Summary

Assessing the health of populations is important for various reasons, especially for health policy purposes. Therefore, there exists a substantial need for health comparisons between populations, including the comparison of individuals, groups of persons, or even populations from different countries, at one point in time and over time.

Two fundamentally different approaches exist to assess the health of populations. The first approach relies on indirect measures of health, which are based on mortality and morbidity statistics, and which are therefore only available at the population level. The second approach relies on direct measures of health, which are collected – based on health surveys – at the individual level.

Based on the needs for comparisons, indirect measures appear to be less appropriate, as they are only available at the population level, but not at the individual or group level. Direct measures, however, are originally obtained at the individual level, and can then be aggregated to any group level, even to the population level. Therefore, direct measures seem to be more appropriate for these comparison purposes.

The open question is then how to compare overall health based on data collected within health surveys. At first glance, a single general health question seems to be appealing. However, studies have shown that this kind of question is not appropriate to compare health over time, nor across populations. Qualitative studies found that respondents even consider very different aspects of health when responding to such a question.

A more appropriate approach seems to be the use of data on several domains of health, as for example mobility, self-care and pain. Anyway, measuring health based on a set of domains is an extremely frequent approach. It provides more comprehensive information and can therefore be used for a wider range of possible applications.

However, three open questions must be addressed when measuring health based on a set of domains. First, a parsimonious set of domains must be selected. Second, health measurement based on this set of domains must be operationalized in a standardized way. Third, this information must be aggregated into a summary measure of health, thereby taking into account that categorical responses to survey questions could be differently interpreted by respondents, and are not necessarily directly comparable. These open questions are addressed in this doctoral thesis.

The overall **objective** of this doctoral thesis is to develop a valid, reliable and sensitive metric of health – based on data collected on a set of domains – that permits to monitor the health of populations over time, and which provides the basis for the comparisons of health across different populations. To achieve this aim two psychometric studies were carried out, entitled "Towards a Minimal Generic Set of Domains" and "Development of a metric of health".

In the **first study** a minimal generic set of domains suitable for measuring health both in the general population and in clinical populations was identified, and contrasted to the domains of the World Health Survey (WHS).

The eight domains of the WHS – mobility, self-care, pain and discomfort, cognition, interpersonal activities, vision, sleep and energy, and affect – were used as a reference, as this set – developed by the World Health Organization (WHO) – so far constitutes the most advanced proposal of what to measure for international health comparisons.

To propose the domains for the minimal generic set, two different regression methodologies – Random Forest and Group Lasso – were applied for the sake of robustness to three different data sources, two national general population surveys and one large international clinical study: the German National Health Interview and Examination Survey 1998, the United States National Health and Nutrition Examination Survey 2007/2008, and the ICF Core Set studies. A domain was selected when it was sufficiently explanatory for self-perceived health.

Based on the analyses the following set of domains, systematically named based on their respective categories within the International Classification of Functioning, Disability and Health (ICF), was proposed as a minimal generic set:

b130	Energy and drive functions
b152	Emotional functions
b280	Sensation of pain
d230	Carrying out daily routine
d450	Walking
d455	Moving around
d850	Remunerative employment

Based on this set, four of the eight domains of the WHS were confirmed both in the general and in clinical populations: mobility, pain and discomfort, sleep and energy, and affect. The other WHS domains not represented in the proposed minimal generic set are vision, which was only confirmed with data of the general population, self-care and interpersonal activities, which were only

confirmed with data of the clinical population and cognition, which could not be confirmed at all.
The ICF categories of `carrying out daily routine´ and `remunerative employment´ also fulfilled the inclusion criteria, though not directly related to any of the eight WHS domains.

This minimal generic set can be used as the starting point to address one of the most important challenges in health measurement, namely the comparability of data across studies and countries. It also represents the first step for developing a common metric of health to link information from the general population to information about sub-populations, such as clinical and institutional populations, e.g. persons living in nursing homes.

In the **second study** a sound psychometric measure was developed based on information collected on the domains of the minimal generic set: energy and drive functions, emotional functions, sensation of pain, carrying out daily routine, mobility and remunerative employment. It was demonstrated that this metric can be used to assess the health of populations and also to monitor health over time.

To develop this metric of health, data from two successive waves of the English Longitudinal Study of Ageing (ELSA) was used. A specific Item Response Theory (IRT) model, the Partial Credit Model (PCM), was applied on 12 items representing the 6 domains from the minimal generic set. All three IRT model assumptions – unidimensionality, local independency and monotonicity – were examined and found to be fulfilled.

The developed metric showed sound psychometric properties: high internal consistency reliability, high construct validity and high sensitivity to change. Therefore, it can be considered an appropriate measure of population health.

Furthermore, it was demonstrated how the health of populations can be compared based on this metric, for subgroups of populations, and over time. Finally, it was outlined how this metric can be used as the basis for comparing health across different populations, as for example from two different countries.

The developed health metric can be seen as the starting point for a wide range of health comparisons, between individuals, groups of persons and populations as a whole, and both at one point in time and over time. It opens up a wide range of possible applications for both health care providers and health policy, and both in clinical settings and in the general population.

Zusammenfassung

Die Messung der Gesundheit von Populationen ist aus verschiedenen Gründen von Bedeutung, insbesondere für gesundheitspolitische Zwecke. Es existiert ein grundlegendes Bedürfnis, Populationen bezüglich ihrer Gesundheit vergleichen zu können. Diese Vergleiche beinhalten den Vergleich von Individuen, Personengruppen oder sogar Bevölkerungen ganzer Ländern, sowohl zum selben Zeitpunkt als auch im Verlauf der Zeit.

Es existieren zwei äußerst unterschiedliche Ansätze zum Vergleich der Gesundheit von Populationen. Der erste Ansatz bezieht sich auf indirekte Maßzahlen, die auf Statistiken zu Mortalität und Krankheit beruhen, und nur auf der Ebene der Gesamtbevölkerung eines Landes verfügbar sind. Der zweite Ansatz bezieht sich auf direkte Maßzahlen, die im Rahmen von Gesundheitssurveys für die einzelnen teilnehmenden Personen erhoben werden.

In Bezug auf die verschiedenen Bedürfnisse im Hinblick auf Vergleiche erscheinen indirekte Maßzahlen weniger geeignet, da sie nur auf der Ebene der Gesamtbevölkerung verfügbar sind, aber nicht auf der individuellen Ebene oder auf Gruppenebene. Direkte Maßzahlen werden hingegen auf individueller Ebene erhoben und können zu jeglichem Gruppenlevel aggregiert werden, einschließlich der Ebene der Gesamtbevölkerung. Daher erscheinen direkte Maßzahlen besser geeignet, um alle Bedürfnisse an möglichen Vergleichen abzudecken.

Aufgrund dieser Vorüberlegungen stellt sich die Frage, wie Gesundheit am besten mit Hilfe von Daten aus Gesundheitssurveys verglichen werden kann. Auf den ersten Blick erscheint eine einzelne Frage zur allgemeinen Gesundheit attraktiv. Allerdings haben Studien gezeigt, dass diese Frage weder dazu geeignet ist, um Gesundheit über die Zeit zu vergleichen, noch zwischen Populationen. Wie qualitative Studien zeigen, berücksichtigen Befragte außerdem äußerst unterschiedliche Aspekte von Gesundheit, wenn sie diese Frage beantworten.

Somit erscheint es vorteilhaft, Informationen zu verschiedenen Gesundheitsbereichen zu berücksichtigen, wie z.B. zu Mobilität, Selbstversorgung oder Schmerz. Die Erhebung von Daten zu verschiedenen Gesundheitsbereichen stellt eine extrem häufige Vorgehensweise dar. Damit erhält man umfassendere Informationen, die für ein breiteres Spektrum an Anwendungen nützlich sind.

Allerdings müssen drei Fragen beantwortet werden, wenn man Gesundheit basierend auf Informationen zu verschiedenen Gesundheitsbereichen messen will. Die erste Frage ist, welche Bereiche auf jeden Fall berücksichtigt werden müssen. Die zweite Frage ist, wie genau Daten zu diesen vorausgewählten Bereichen auf eine standardisierte Art erhoben werden können. Die dritte Frage

ist, wie diese Informationen in eine Maßzahl zusammengefasst werden können, vor allem wenn man berücksichtigt, dass Befragte in einer Umfrage Antwortkategorien unterschiedlich interpretieren könnten, und Antworten deshalb nicht zwingend direkt vergleichbar sind. Diese Fragen werden in der vorliegenden Doktorarbeit adressiert.

Das **Gesamtziel** dieser Doktorarbeit ist es, eine valide, reliable und sensitive Gesundheitsskala zu entwickeln – und zwar basierend auf Daten aus verschiedenen Gesundheitsbereichen – die es ermöglicht, Gesundheit über die Zeit zu verfolgen, und die eine Basis für den Vergleich von Gesundheit über verschiedene Populationen liefert. Um dieses Ziel zu erreichen, wurden zwei psychometrische Studien durchgeführt, mit den Titeln „Auswahl einer minimalen und allgemein anwendbaren Menge von Gesundheitsbereichen“ und „Entwicklung einer Gesundheitsskala“.

In der **ersten Studie** wurde eine minimale und allgemein anwendbare Menge von Gesundheitsbereichen identifiziert, die geeignet ist, Gesundheit sowohl in der Allgemeinbevölkerung als auch in klinischen Populationen zu messen. Diese Auswahl wurde den Gesundheitsbereichen aus dem Weltgesundheitssurvey gegenübergestellt.

Die acht Gesundheitsbereiche des Weltgesundheitssurveys – Mobilität, Selbstversorgung, Schmerz und Unannehmlichkeiten, Kognition, soziale Aktivitäten, Sehen, Schlaf und Energie, und Affekt – wurden als Referenz verwendet, da diese Auswahl, die von der Weltgesundheitsorganisation entwickelt wurde, bis jetzt den am besten ausgearbeiteten Vorschlag darstellt, was zum Zweck von internationalen Gesundheitsvergleichen gemessen werden sollte.

Um eine minimale und allgemein anwendbare Menge von Gesundheitsbereichen vorzuschlagen, wurden zwei verschiedene Regressionsmethoden angewendet: Random Forest und Group Lasso. Diese wurden vor dem Hintergrund der Robustheit auf drei Datensätze angewendet, auf zwei nationale allgemeine Bevölkerungsumfragen und auf Daten aus einer großen internationalen klinischen Studie: den deutschen Bundesgesundheitssurvey von 1998, den National Health and Nutrition Examination Survey 2007/2008 aus den USA und die ICF Core Set Studie. Ein Gesundheitsbereich wurde ausgewählt, wenn er ausreichend erklärend für die Selbsteinschätzung von Gesundheit war.

Basierend auf den Analysen wurden die folgenden Gesundheitsbereiche, systematisch benannt durch die zugehörigen Kategorien der Internationalen Klassifikation für Funktionsfähigkeit, Behinderung und Gesundheit (ICF), als minimale und allgemein anwendbare Menge von Gesundheitsbereichen vorgeschlagen:

b130 Funktionen der psychischen Energie und des Antriebs
b152 Emotionale Funktionen
b280 Schmerz
d230 Die tägliche Routine durchführen
d450 Gehen
d455 Sich auf andere Weise fortbewegen
d850 Bezahlte Tätigkeit

Basierend auf dieser Auswahl wurden vier der acht Gesundheitsbereiche des Weltgesundheitssurveys sowohl in der Allgemeinbevölkerung als auch in klinischen Populationen bestätigt: Mobilität, Schmerz und Unannehmlichkeiten, Schlaf und Energie sowie Affekt. Die anderen Gesundheitsbereiche, die nicht in dieser Auswahl enthalten sind, sind Sehen, welches nur mit Daten aus der Allgemeinbevölkerung bestätigt wurde, Selbstversorgung und soziale Aktivitäten, welche nur mit Daten aus der klinischen Population bestätigt wurden, und Kognition, welche durch keine der beiden Datenquellen bestätigt wurde.
Die ICF Kategorien „Die tägliche Routine durchführen“ und „Bezahlte Tätigkeit“ erfüllten ebenfalls die Einschlusskriterien, auch wenn sie nicht direkt in Beziehung zu einer der Gesundheitsbereiche aus dem Weltgesundheitssurvey stehen.

Diese minimale und allgemein anwendbare Menge von Gesundheitsbereichen kann als Ausgangspunkt für die Beantwortung einer der größten Herausforderungen in der Messung von Gesundheit verwendet werden, nämlich der studien- und länderübergreifenden Vergleichbarkeit von Gesundheit. Diese Auswahl kann auch als erster Schritt gesehen werden, um eine Gesundheitsskala zu entwickeln, die es ermöglicht, Gesundheitsangaben aus der Allgemeinbevölkerung mit der von Untergruppen zu verknüpfen, wie z.B. klinischen Populationen oder Personen in Institutionen wie z.B. in Altenheimen.

In der **zweiten Studie** wurde eine Gesundheitsskala mit gut fundierten psychometrischen Eigenschaften auf Basis der vorher definierten minimalen und allgemein anwendbaren Menge von Gesundheitsbereichen entwickelt: „Funktionen der psychischen Energie und des Antriebs“, „Emotionale Funktionen“, „Schmerz“, „Die tägliche Routine durchführen“, „Mobilität“, und „Bezahlte Tätigkeit“. Es wurde gezeigt, dass diese Gesundheitsskala verwendet werden kann, um die Gesundheit von Populationen zu messen, und auch, um ihre Entwicklung über die Zeit zu verfolgen.

Um diese Gesundheitsskala zu entwickeln, wurden Daten von zwei aufeinanderfolgenden Befragungswellen der English Longitudinal Study of Ageing (ELSA) verwendet. Das Partial Credit Modell, ein spezielles Modell aus der Item Response Theorie, wurde auf 12 Items angewendet, die Informationen

zu den sechs vorher identifizierten Gesundheitsbereichen enthalten. Alle drei Modellannahmen der Item Response Theorie – Eindimensionalität, lokale Unabhängigkeit und Monotonie – wurden untersucht und für erfüllt befunden.

Die entwickelte Gesundheitsskala zeigte gut fundierte psychometrische Eigenschaften: hohe interne Reliabilität, hohe Konstruktvalidität, und starke Sensitivität für Veränderungen über die Zeit. Somit kann sie als geeignetes Maß für die Messung der Gesundheit von Populationen betrachtet werden.

Darüber hinaus wurde gezeigt, wie die Gesundheit von Populationen basierend auf dieser Gesundheitsskala verglichen werden kann, sowohl für Subgruppen als auch über die Zeit. Schlussendlich wurde ausgeführt, wie diese Gesundheitsskala verwendet werden kann, um die Gesundheit von verschiedenen Populationen zu vergleichen, wie z.B. die von zwei verschiedenen Ländern.

Die entwickelte Gesundheitsskala kann als Ausgangspunkt für zahlreiche Gesundheitsvergleiche angesehen werden, zwischen Individuen, Personengruppen oder ganzen Populationen, zu einem Zeitpunkt und im Verlauf der Zeit. Sie ermöglicht eine Vielzahl von potenziellen Anwendungen sowohl für Leistungserbringer im Gesundheitssystem als auch für die Gesundheitspolitik, und sowohl im klinischen Rahmen wie in der Allgemeinbevölkerung.

References

1. World Health Organization: **World report on disability**: World Health Organization; 2011.
2. Salomon JA, Mathers CD, Chatterji S, Sadana R, Üstün TB, Murray CJL: **Quantifying Individual Levels of Health: definitions, concepts, and measurement issues**. In: *Health systems performace assessment: debates, methods and empiricism*. edn. Edited by Murray CJL, Evans DB. Geneva: WHO; 2003: 301-318.
3. US Department of Health and Human Services: **Healthy People 2010**. Washington, DC: US Department of Health and Human Services; 2000.
4. Mathers CD, Salomon J, Murray CJ, Lopez A, Murray C, Evans D: **Alternative summary measures of average population health**. In: *Health systems performance assessment: debates, methods and empiricism*. edn. Edited by Murray CJL, Evans DB. Geneva: WHO; 2003: 319-334.
5. Mathers CD: **Health expectancies: An overview and critical appraisal**. In: *Summary measures of population health: concepts, ethics, measurement and application*. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.
6. Salomon JA, Wang H, Freeman MK, Vos T, Flaxman AD, Lopez AD, Murray CJ: **Healthy life expectancy for 187 countries, 1990-2010: a systematic analysis for the Global Burden Disease Study 2010**. *Lancet* 2012, **380**(9859):2144-2162.
7. Murray CJL, Mathers CD, Salomon JA, Lopez AD: **Health gaps: an overview and critical appraisal**. In: *Summary measures of population health: concepts, ethics, measurement and application*. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.
8. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, Ezzati M, Shibuya K, Salomon JA, Abdalla S *et al*: **Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010**. *Lancet* 2012, **380**(9859):2197-2223.
9. Mantzavinis GD, Pappas N, Dimoliatis ID, Ioannidis JP: **Multivariate models of self-reported health often neglected essential candidate determinants and methodological issues**. *Journal of clinical epidemiology* 2005, **58**(5):436-443.
10. **European Health Interview Survey (EHIS) Questionnaire - English Version -** [http://ec.europa.eu/health/ph_information/implement/wp/systems/docs/ev_20070315_ehis_en.pdf] (last accessed July 28, 2013)
11. McHorney CA, Ware JE, Jr., Raczek AE: **The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs**. *Medical care* 1993, **31**(3):247-263.
12. Rabin R, de Charro F: **EQ-5D: a measure of health status from the EuroQol Group**. *Annals of medicine* 2001, **33**(5):337-343.
13. Lee S, Grant D: **The effect of question order on self-rated general health status in a multilingual survey context**. *American journal of epidemiology* 2009, **169**(12):1525-1530.
14. Idler EL, Benyamini Y: **Self-rated health and mortality: a review of twenty-seven community studies**. *Journal of health and social behavior* 1997, **38**(1):21-37.

15. Salomon JA, Nordhagen S, Oza S, Murray CJ: **Are Americans feeling less healthy? The puzzle of trends in self-rated health**. *American journal of epidemiology* 2009, **170**(3):343-351.
16. Sadana R, Mathers CD, Lopez AD, Murray CJL, Iburg KM: **Comparative analyses of more than 50 household surveys on health status**. In: *Summary measures of population health: concepts, ethics, measurement and application*. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.
17. Eurostat: **Self-reported health in the European Community**. In: *Statistics in focus, population and social conditions*. Luxembourg: Eurostat; 1997.
18. OECD: **Health at a Glance: Europe 2010**. *OECD Publishing* 2010.
19. Simon JG, De Boer JB, Joung IM, Bosma H, Mackenbach JP: **How is your health in general? A qualitative study on self-assessed health**. *European journal of public health* 2005, **15**(2):200-208.
20. Krause NM, Jay GM: **What do global self-rated health items measure?** *Medical care* 1994, **32**(9):930-942.
21. Manderbacka K: **Examining what self-rated health question is understood to mean by respondents**. *Scandinavian journal of social medicine* 1998, **26**(2):145-153.
22. Üstün TB, Chatterji S, Villanueva M, Bendib L, Çelik C, Sadana R, Valentine NB, Ortiz JP, Tandon A, Salomon JA *et al*: **WHO Multi-country Survey Study on Health and Responsiveness 2000–2001**. In: *GPE Discussion Paper 37*. Geneva: World Health Organization/Global Programme on Evidence for Health Policy; 2001.
23. Üstün TB, Chatterji S, Mechbal A, Murray CJL, Groups WC: **The World Health Surveys**. In: *Health systems performace assessment: debates, methods and empiricism*. edn. Edited by Murray CJL, Evans DB. Geneva: WHO; 2003.
24. **Public Use File BGS98, German National Health Interview and Examination Survey 1998**. In. Berlin (Germany): Robert Koch Institute; 2000.
25. **About the National Health and Nutrition Examination Survey** [http://www.cdc.gov/nchs/nhanes/about_nhanes.htm] (last accessed July 28, 2013)
26. **Health and Retirement Study: About the Health and Retirement Study** [http://hrsonline.isr.umich.edu/] (last accessed July 28, 2013)
27. **English Longitudinal Study of Ageing: Insight into a maturing population** [http://www.ifs.org.uk/ELSA] (last accessed July 28, 2013)
28. **SHARE - Survey of Health, Ageing and Retirement in Europe** [http://www.share-project.org/] (last accessed July 28, 2013)
29. Ware JE, Jr., Sherbourne CD: **The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection**. *Medical care* 1992, **30**(6):473-483.
30. Ustun TB, Chatterji S, Kostanjsek N, Rehm J, Kennedy C, Epping-Jordan J, Saxena S, von Korff M, Pull C: **Developing the World Health Organization Disability Assessment Schedule 2.0**. *Bulletin of the World Health Organization* 2010, **88**(11):815-823.
31. Murray CJL, Salomon JA, Mathers CD, Lopez AD: **Summary measures of population health: conclusions and recommendations**. In: *Summary measures of population health: concepts, ethics, measurement*

and application. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.

32. Iburg KM, Salomon JA, Tandon A, Murray CJL: **Cross-population comparability of physician-assessed and self-reported measures of health**. In: *Summary measures of population health: concepts, ethics, measurement and application*. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.
33. World Health Organization: **International Classification of Functioning, Disability and Health: ICF**. Geneva: World Health Organization; 2001.
34. Stucki G, Kostanjsek N, Ustun B, Cieza A: **ICF-based classification and measurement of functioning**. *European journal of physical and rehabilitation medicine* 2008, **44**(3):315-328.
35. World Health Organization: **Technical appendix C: Design and implementation of the World Health Survey**. In: *World report on disability*. edn. Edited by World Health Organization TWB. Malta: World Health Organization; 2011.
36. Murray CJL, Salomon JA, Mathers CD, Lopez AD (eds.): **Summary measures of population health: concepts, ethics, measurement and applications**. Geneva: WHO; 2002.
37. Murray CJL, Evans DB: **Health systems performance assessment: debates, methods and empiricism**. Geneva: World Health Organization; 2003.
38. Sadana R: **Development of standardized health state descriptions**. In: *Summary measures of population health: concepts, ethics, measurement and application*. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.
39. Üstün TB, Chatterji S, Villanueva M, Bendib L, Çelik C, Sadana R, Valentine NB, Ortiz JP, Tandon A, Salomon JA *et al*: **WHO Multi-country Survey Study on Health and Responsiveness 2000–2001** In: *Health systems performance assessment: debates, methods and empiricism*. edn. Edited by Murray CJL, Evans DB. Geneva: WHO; 2003.
40. **World Health Survey Instruments and Related Documents** [http://www.who.int/healthinfo/survey/instruments/en/] (last accessed July 28, 2013)
41. **Washington Group on Disability Statistics** [http://unstats.un.org/unsd/methods/citygroup/washington.htm] (last accessed July 28, 2013)
42. Sadana R, Mathers CD, Lopez AD, Murray CJL, Iburg K: **Comparative analyses of more than 50 household surveys on health status**. In: *GPE Discussion Paper Series: No15*. Geneva: World Health Organization/Global Programme on Evidence for Health Policy; 2000.
43. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B: **Depression, chronic diseases, and decrements in health: results from the World Health Surveys**. *Lancet* 2007, **370**(9590):851-858.
44. Murray CJL, Tandon A, Salomon JA, Mathers CD, Sadana R: **Cross-Population Comparability of Evidence for Health Policy**. In: *Health systems performance assessment: debates, methods and empiricism*. edn. Edited by Murray CJL, Evans DB. Geneva: World Health Organization; 2003.

45. Bond TG, Fox CM: **Applying the Rasch model: Fundamental measurement in the human sciences**, 2nd edn. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.; 2007.
46. Hosseinpoor AR, Stewart Williams J, Amin A, Araujo de Carvalho I, Beard J, Boerma T, Kowal P, Naidoo N, Chatterji S: **Social determinants of self-reported health in women and men: understanding the role of gender in population health**. *PloS one* 2012, **7**(4):e34799.
47. Hosseinpoor AR, Williams JS, Jann B, Kowal P, Officer A, Posarac A, Chatterji S: **Social determinants of sex differences in disability among older adults: a multi-country decomposition analysis using the World Health Survey**. *International journal for equity in health* 2012, **11**:52.
48. Chatterji S, Kowal P, Mathers C, Naidoo N, Verdes E, Smith JP, Suzman R: **The health of aging populations in China and India**. *Health affairs (Project Hope)* 2008, **27**(4):1052-1063.
49. Murray CJ, Frenk J: **Health metrics and evaluation: strengthening the science**. *Lancet* 2008, **371**(9619):1191-1199.
50. Meijer E, Kapteyn A, Andreyeva T: **Internationally comparable health indices**. *Health economics* 2011, **20**(5):600-619.
51. Murray CJL, Mathers CD, Salomon JA: **Towards Evidence-Based Public Health**. In: *Health systems performance assessment: debates, methods and empiricism*. edn. Edited by Murray CJL, Evans DB. Geneva: World Health Organization; 2003.
52. Murray CJ, Salomon JA, Mathers C: **A critical examination of summary measures of population health**. *Bulletin of the World Health Organization* 2000, **78**(8):981-994.
53. **Public Use Files zu den RKI-Gesundheitssurveys** [http://www.rki.de/DE/Content/Gesundheitsmonitoring/PublicUseFiles/publicusefiles_node.html] (last accessed July 28, 2013)
54. **National Health and Nutrition Examination Survey** [http://www.cdc.gov/nchs/nhanes.htm] (last accessed July 28, 2013)
55. **ICF Core Sets Projects** [http://www.icf-research-branch.org/icf-core-sets-projects-sp-1641024398] (last accessed July 28, 2013)
56. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G: **ICF linking rules: an update based on lessons learned**. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* 2005, **37**(4):212-218.
57. Cieza A, Stucki G: **Content comparison of health-related quality of life (HRQOL) instruments based on the international classification of functioning, disability and health (ICF)**. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2005, **14**(5):1225-1237.
58. Ware JE, Kosinski M, Dewey JE: **How to score version 2 of the SF-36® health survey (Standards & Acute Forms)**, 2nd edn. Lincoln, RI: QualityMetric, Inc.; 2001.
59. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Üstün BT, Stucki G: **Identification of candidate categories of the International Classification of Functioning Disability and Health (ICF) for a Generic ICF Core Set based on regression modelling**. *BMC Medical Research Methodology* 2006, **6**(1):36.
60. Breiman L: **Random forests**. *Machine learning* 2001, **45**(1):5-32.

61. Tibshirani R: **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society Series B (Methodological)* 1996, **58**:267-288.
62. Yuan M, Lin Y: **Model selection and estimation in regression with grouped variables**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006, **68**(1):49-67.
63. Meier L, Van De Geer S, Bühlmann P: **The group lasso for logistic regression**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008, **70**(1):53-71.
64. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, 2nd edn: Springer; 2009.
65. Gertheiss J, Tutz G: **Penalized regression with ordinal predictors**. *International Statistical Review* 2009, **77**(3):345-365.
66. Gertheiss J, Hogger S, Oberhauser C, Tutz G: **Selection of ordinally scaled independent variables with applications to international classification of functioning core sets**. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2011, **60**(3):377-395.
67. R Development Core Team: **R: A language and environment for statistical computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2010.
68. **Repository of Disability Surveys and Censuses: Project Model Disability Survey** [http://disabilitysurvey.checkdesign.de/] (last accessed July 28, 2013)
69. Andreotti A, Minicuci N, Kowal P, Chatterji S: **Multidimensional profiles of health status: an application of the grade of membership model to the world health survey**. *PloS one* 2009, **4**(2):e4426.
70. Hosseinpoor AR, Stewart Williams JA, Itani L, Chatterji S: **Socioeconomic inequality in domains of health: results from the World Health Surveys**. *BMC public health* 2012, **12**:198.
71. Hosseinpoor AR, Stewart Williams JA, Gautam J, Posarac A, Officer A, Verdes E, Kostanjsek N, Chatterji S: **Socioeconomic inequality in disability among adults: a multicountry study using the world health survey**. *American journal of public health* 2013, **103**(7):1278-1286.
72. Fayed N, Cieza A, Bickenbach JE: **Linking health and health-related information to the ICF: a systematic review of the literature from 2001 to 2008**. *Disability and rehabilitation* 2011, **33**(21-22):1941-1951.
73. Mair P, Hatzinger R: **Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R**. *Journal of Statistical Software* 2007, **20**(9):1-20.
74. Masters GN: **A Rasch model for partial credit scoring**. *Psychometrika* 1982, **47**(2):149-174.
75. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK *et al*: **Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS)**. *Medical care* 2007, **45**(5 Suppl 1):S22-31.
76. Reise SP: **The rediscovery of bifactor measurement models**. *Multivariate Behavioral Research* 2012, **47**(5):667-696.

77. Jennrich RI, Bentler PM: **Exploratory Bi-factor Analysis**. *Psychometrika* 2011, **76**(4):537-549.
78. Reise SP, Morizot J, Hays RD: **The role of the bifactor model in resolving dimensionality issues in health outcomes measures**. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2007, **16 Suppl 1**:19-31.
79. Buja A, Eyuboglu N: **Remarks on parallel analysis**. *Multivariate behavioral research* 1992, **27**(4):509-540.
80. Roscino A, Pollice A: **A Generalization of the Polychoric Correlation Coefficient**. In: *Data Analysis, Classification and the Forward Search.* edn. Edited by Zani S, Cerioli A, Riani M, Vichi M: Springer Berlin Heidelberg; 2006: 135-142.
81. Ekström J: **A Generalized Definition of the Polychoric Correlation Coefficient.** *UC Los Angeles: Department of Statistics, UCLA* 2011.
82. Reeve BB, Fayers P: **Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers and R. D. Hays (Eds.), Assessing Quality of Life in Clinical Trials: Methods of Practice**. *2nd Edition Oxford University Press* 2005:55-73.
83. Pallant JF, Tennant A: **An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS)**. *The British journal of clinical psychology / the British Psychological Society* 2007, **46**(Pt 1):1-18.
84. Crane PK, Gibbons LE, Jolley L, van Belle G: **Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar**. *Medical care* 2006, **44**(11 Suppl 3):S115-123.
85. Choi SW, Gibbons LE, Crane PK: **Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations**. *Journal of statistical software* 2011, **39**(8):1.
86. **The Research Methods Knowledge Base** [http://www.socialresearchmethods.net/kb/] (last accessed July 28, 2013)
87. Cronbach LJ: **Coefficient alpha and the internal structure of tests**. *Psychometrika* 1951, **16**(3):297-334.
88. Revelle W, Zinbarg RE: **Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma**. *Psychometrika* 2009, **74**(1):145-154.
89. Zinbarg RE, Revelle W, Yovel I, Li W: **Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability**. *Psychometrika* 2005, **70**(1):123-133.
90. McDonald RP: **Test Theory: A unified Treatment**: Psychology Press; 1999.
91. Wood SN: **Generalized additive models: an introduction with R**, vol. 66: Chapman & Hall; 2006.
92. Preedy VR, Watson RR: **Handbook of disease burdens and quality of life measures**: Springer New York; 2010.
93. R Development Core Team: **R: A language and environment for statistical computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2013.
94. Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, Begum N, Shah R, Karyana M, Kosen S *et al*: **Common values in assessing health outcomes from disease and injury: disability weights**

measurement study for the Global Burden of Disease Study 2010. *Lancet* 2012, **380**(9859):2129-2143.

95. Hays RD, Liu H, Spritzer K, Cella D: **Item response theory analyses of physical functioning items in the medical outcomes study**. *Medical care* 2007, **45**(5 Suppl 1):S32-38.
96. Hahn EA, Devellis RF, Bode RK, Garcia SF, Castel LD, Eisen SV, Bosworth HB, Heinemann AW, Rothrock N, Cella D: **Measuring social health in the patient-reported outcomes measurement information system (PROMIS): item bank development and testing**. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2010, **19**(7):1035-1044.
97. Gorman BK, Read JG: **Gender disparities in adult health: an examination of three measures of morbidity**. *Journal of health and social behavior* 2006, **47**(2):95-110.
98. Murray CJL, Salomon JA, Mathers CD: **A critical examination of summary measures of population health**. In: *Summary measures of population health: concepts, ethics, measurement and applications*. edn. Edited by Murray CJL, Salomon JA, Mathers CD, Lopez AD. Geneva: WHO; 2002.
99. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY *et al*: **Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010**. *Lancet* 2012, **380**(9859):2095-2128.
100. Cieza A, Oberhauser C, Bickenbach J, Jones RN, Üstün TB, Kostanjsek N, Morris JN, Chatterji S: **The English are healthier than the Americans: Really? – A psychometric study**. (submitted to the British Medical Journal).

Appendix

Questions on Health State Descriptions used in the World Health Survey

Overall Health

The first questions are about your overall health, including both your physical and your mental health.

- In general, how would you rate your health today?
- Overall in the last 30 days, how much difficulty did you have with work or household activities?

Mobility

- Overall in the last 30 days, how much difficulty did you have with moving around?
- In the last 30 days, how much difficulty did you have in vigorous activities, such as running 3 km (or equivalent) or cycling?

Self Care

- Overall in the last 30 days, how much difficulty did you have with self-care, such as washing or dressing yourself?
- In the last 30 days, how much difficulty did you have in taking care of and maintaining your general appearance (e.g. grooming, looking neat and tidy etc.)

Pain and Discomfort

- Overall in the last 30 days, how much of bodily aches or pains did you have?
- In the last 30 days, how much bodily discomfort did you have?

Cognition

- Overall in the last 30 days, how much difficulty did you have with concentrating or remembering things?
- In the last 30 days, how much difficulty did you have in learning a new task (for example, learning how to get to a new place, learning a new game, learning a new recipe etc.)?

Interpersonal Activities

- Overall in the last 30 days, how much difficulty did you have with personal relationship or participation in the community?
- In the last 30 days, how much difficulty did you have in dealing with conflicts and tensions with others?

Vision

- Do you wear glasses or contact lenses?
  (If Respondent says YES to this question, preface the next 2 questions with "Please answer the following questions taking into account your glasses or contact lenses".)
- In the last 30 days, how much difficulty did you have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)?
- In the last 30 days, how much difficulty did you have in seeing and recognizing an object at arm's length or in reading?

Sleep and Energy

- Overall in the last 30 days, how much of a problem did you have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?
- In the last 30 days, how much of a problem did you have due to not feeling rested and refreshed during the day (e.g. feeling tired, not having energy)?

Affect

- Overall in the last 30 days, how much of a problem did you have with feeling sad, low or depressed?
- Overall in the last 30 days, how much of a problem did you have with worry or anxiety?

Curriculum Vitae

Persönliche Daten

Cornelia Oberhauser

Geburtsdatum: 17. September 1982
Geburtsort: München

Schulbildung

1989-1993
Grundschule an der Ostpreußenstraße, München

1993-2002
Theresia-Gerhardinger-Gymnasium am Anger, München

Studium

2002-2007 Studium der Statistik (Diplom) an der Ludwig-Maximilians-Universität München mit den Anwendungsgebieten Soziologie und Geographie

Vordiplom im Oktober 2004
Diplom im Juli 2007

Berufspraxis

von Februar 2005 bis Dezember 2006:
studentische Hilfskraft am Sonderforschungsbereich 386 (Fahrmeir) am Institut für Statistik der LMU

von September 2005 bis Juni 2007:
studentische Hilfskraft im Statistischen Beratungslabor (Küchenhoff) am Institut für Statistik der LMU

seit Juli 2007:
wissenschaftliche Mitarbeiterin am damaligen Institut für Gesundheits- und Rehabilitationswissenschaften der LMU, jetzt Lehrstuhl für Public Health und Versorgungsforschung am Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der LMU

seit November 2007:
zusätzlich wissenschaftliche Mitarbeiterin im Statistischen Beratungslabor am Institut für Statistik der LMU

Lehre am Institut für Statistik

SS 2007:	Übung Statistik für Kommunikationswissenschaftler
WS 2007/08:	Anfängerpraktikum (jeweils Vorlesungszeit und Ferien)
SS 2008:	Übung Statistik für Kommunikationswissenschaftler SAS-Kurs
WS 2008/09:	Anfängerpraktikum (jeweils Vorlesungszeit und Ferien)
SS 2009:	Vorlesung zur SAS-Vertiefung Übung zur SAS-Vertiefung
WS 2009/10:	Übung Statistik für Kommunikationswissenschaftler Anfängerpraktikum (jeweils Vorlesungszeit und Ferien)
SS 2010:	Vorlesung zur Statistischen Software (SAS) Übung zur Statistischen Software (SAS)
WS 2010/11:	Übung Statistik für Kommunikationswissenschaftler Vorlesung Einführung in die Statistische Software (SAS-Teil) Anfängerpraktikum (jeweils Vorlesungszeit und Ferien) Praxisprojekt (jeweils Vorlesungszeit und Ferien)
SS 2011:	Vorlesung zur Statistischen Software (SAS) Übung zur Statistischen Software (SAS) Vorlesung Einführung in die Statistische Software (SAS-Teil)
WS 2011/12:	Übung Statistik für Kommunikationswissenschaftler Vorlesung Einführung in die Statistische Software (SAS-Teil) Anfängerpraktikum (jeweils Vorlesungszeit und Ferien) Praxisprojekt (jeweils Vorlesungszeit und Ferien)
SS 2012:	Vorlesung zur Statistischen Software (SAS) Übung zur Statistischen Software (SAS) Anfängerpraktikum (Ferien)
WS 2012/13:	Übung Statistik für Kommunikationswissenschaftler Vorlesung Einführung in die Statistische Software (SAS-Teil) Anfängerpraktikum (jeweils Vorlesungszeit und Ferien) Praxisprojekt (jeweils Vorlesungszeit und Ferien)
SS 2013:	Vorlesung zur Statistischen Software (SAS) Übung zur Statistischen Software (SAS)

Sonstige Lehre

Juni 2009:	Zweitägiger SAS-Kurs für Mitarbeiter und Doktoranden am Technologie- und Förderzentrum im Kompetenzzentrum für Nachwachsende Rohstoffe (TFZ) in Straubing
Oktober 2012:	Viertägiger SAS-Kurs im Rahmen des Promotionsprogrammes an der TUM School of Management, Technische Universität München

Publikationen

Cieza A, Bostan C, Oberhauser C, Bickenbach JE: **Explaining functioning outcomes across musculoskeletal conditions: a multilevel modelling approach**. *Disability and rehabilitation* 2010, **32 Suppl 1**:S85-93.

Tschiesner U, Oberhauser C, Cieza A: **ICF Core Set for head and neck cancer: do the categories discriminate among clinically relevant subgroups of patients?** *International journal of rehabilitation research Internationale Zeitschrift für Rehabilitationsforschung Revue internationale de recherches de readaptation* 2011, **34**(2):121-130.

Gertheiss J, Hogger S, Oberhauser C, Tutz G: **Selection of ordinally scaled independent variables with applications to international classification of functioning core sets**. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2011, **60**(3):377-395.

Kus S, Oberhauser C, Cieza A: **Validation of the brief International Classification of Functioning, Disability, and Health (ICF) core set for hand conditions**. *Journal of hand therapy : official journal of the American Society of Hand Therapists* 2012, **25**(3):274-286; quiz 287.

Bostan C, Oberhauser C, Cieza A: **Investigating the dimension functioning from a condition-specific perspective and the qualifier scale of the International Classification of Functioning, Disability, and Health based on Rasch analyses**. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists* 2012, **91**(13 Suppl 1):S129-140.

Oberhauser C, Escorpizo R, Boonen A, Stucki G, Cieza A: **Statistical validation of the brief International Classification of Functioning, Disability and Health Core Set for osteoarthritis based on a large international sample of patients with osteoarthritis**. *Arthritis care & research* 2013, **65**(2):177-186.

Eingereichte Publikationen

Cieza A, Oberhauser C, Bickenbach J, Chatterji S, Stucki G: **Towards a Minimal Generic Set of Domains of Functioning and Health.** (*BMC Public Health, under review*)

Bostan C, Oberhauser C, Stucki G, Bickenbach J, Cieza A: **Biological health or lived health: Which predicts self-reported general health better?** (*European Journal of Public Health, under review*)

Cieza A, Oberhauser C, Bickenbach J, Jones RN, Üstün TB, Kostanjsek N, Morris JN, Chatterji S: **The English are healthier than the Americans: Really? – A psychometric study** (*British Medical Journal, submitted*)

Eidesstattliche Versicherung

Oberhauser, Cornelia

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Thema

Addressing the challenge of health measurement:
The development of a metric of health to validly and reliably follow up the health of populations

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 10.03.2014 — Ort, Datum

Cornelia Oberhauser — Unterschrift Doktorandin/Doktorand