DISSERTATION

Patient-reported outcome measures: Anwendungen und Herausforderungen im Bereich der Neurologie

Patient-reported outcome measures: applications and challenges in the field of neurology

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Ana Sofia Oliveira Gonçalves

Erstbetreuer: Prof. Dr. Dr. Tobias Kurth

Datum der Promotion: 28.02.2025

1. Table of Contents

I. List of Abbreviations

ALDVMM	Adjusted Limited Dependent Variable Mixture Model
AQoL	Assessment of Quality of Life
BIC	Bayesian information criterion
BMI	Body mass index
B_PROUD	Berlin_PRehospital Or Usual Delivery in stroke care
CHU9D	Child Health Utility instrument
CI	Confidence interval
CINAHL	Cumulative Index to Nursing and Allied Health Literature
EFA	Exploratory factor analysis
EMS	Emergency medical services
EORTC QLQ30	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30
FACT	Functional Assessment of Cancer Therapy
GDP	Gross Domestic Product
HAQ	Health Assessment Questionnaire
HIT-6	Headache Impact Test-6
HRQoL	Health-related quality of life
HSUV	Health state utility value
HUI	Health Utilities Index
ICD-10	International Statistical Classification of Diseases and Related Health Problems, Tenth Revision
ICER	Incremental cost-effectiveness ratio
IQR	Interquartile range
IQWIG	Institute for Quality and Efficiency in Health Care
ISPOR	Professional Society for Health Economics and Outcomes Research
QALY	Quality-adjusted life year
QWB	Quality of Well-Being index
MAE	Mean absolute error
MDK	Medical Service of the Health Insurance
MIC	Multi-Instrument Comparison

mRS	Modified Rankin scale
MSU	Mobile stroke unit
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PaRIS	Patient-Reported Indicator Surveys
PBM	Preference-based measure
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROM	Patient reported outcome measures
RCT	Randomised controlled trial
RMSE	Mean squared error
RMQ	Roland–Morris Disability Questionnaire
SD	Standard deviation
SF-12	Short Form-12
SF-36	Short Form-36
SRM	Standardised response mean
SMARTGEM	Smartphone assisted migraine therapy
tPA	Tissue plasminogen activator
UK	United Kingdom
US	United States
VAS	Visual Analogue Scale
WHO	World Health Organization

II. List of Figures

III. List of Tables

1. Abstract

Patient-reported outcome measurements (PROMs) are tools that can give a broad view of a patient's health, rather than just focusing on their clinical symptoms. They are part of a trend that aims to shift the health care paradigm towards a patient-centre care approach. Given their increasing importance, it is crucial that appropriate PROMs are administered, correct methods are employed to analyse their results, and their strengths and limitations are discussed.

In Thesis Article 1, I developed a mapping algorithm to convert values from a condition-specific PROM (Headache Impact Test-6) into German EQ-5D utility scores. We started by analysing the correlation between the two instruments, as well as their conceptual overlap. We then fitted several regression models. I showed that there might be no conceptual overlap between the HIT-6 and the EQ-5D-5L. Thus, mapping can't always be used to obtain utilities.

Despite a plethora of guidance on ways to conduct as well as report mapping studies, small attention has been devoted to how authors work with datasets with multiple observations per subject over time. In Thesis Article 2 I conducted a systematic review on the methodological challenges of this subject. I showed that when data sets with multiple observations are used, researchers often only employ one time point in the estimation data set and another time point for its validation, hence ignoring that health states with different degrees of severity may be present only at a specific time point.

Thesis Article 3 uses a PROM (EQ-5D-3L) and a clinician-reported measure of global disability (mRS – modified Ranking Scale) to conduct, respectively, a cost-utility and a cost-effectiveness analysis. The G-formula was employed to compute incremental costs and incremental outcomes due to a mobile stroke unit (MSU) mobilisation. We found that the additional MSU mobilisation yielded an incremental EUR 40,984 per quality-adjusted life year and an incremental EUR 81,491.49 per survival without symptoms/disability (using a dichotomised mRS).

This PhD dissertation showcases different methods within the health data sciences field, including the development of a mapping algorithm, the conduction of a systematic review and the conduction of an economic evaluation.

2. Zusammenfassung

Patient-reported outcome measurements (PROMs) sind Instrumente, die einen umfassenden Überblick über den Gesundheitszustand eines Patienten geben können, anstatt sich nur auf seine klinischen Symptome zu konzentrieren. Sie sind Teil eines Trends, der darauf abzielt, das Paradigma der Gesundheitsversorgung in Richtung eines patientenzentrierten Versorgungsansatzes zu verändern. In Anbetracht ihrer zunehmenden Bedeutung ist es von entscheidender Bedeutung, dass geeignete PROMs durchgeführt werden, korrekte Methoden zur Analyse ihrer Ergebnisse angewandt werden und ihre Stärken und Grenzen diskutiert werden.

In Artikel 1 meiner Dissertation habe ich einen Mapping-Algorithmus entwickelt, um Werte eines krankheitsspezifischen PROM (Headache Impact Test-6) in EQ-5D-Utility-Scores für Deutschland zu übersetzen. Zunächst analysierten wir die Korrelation zwischen den beiden Instrumenten sowie ihre konzeptionelle Überschneidung. Anschließend haben wir mehrere Regressionsmodelle berechnet. Dadurch konnte gezeigt werden, dass es möglicherweise keine konzeptionelle Überschneidung zwischen dem HIT-6 und dem EQ-5D-5L gibt. Daher kann das Mapping nicht immer zur Ermittlung des Utility Scores verwendet werden.

Trotz einer Vielzahl von Anleitungen zur Durchführung und Berichterstattung von Mapping-Studien wurde der Frage wenig Aufmerksamkeit gewidmet, wie Autoren mit Datensätzen mit wiederholten Messungen pro Person im Laufe der Zeit umgehen. In Artikel 2 meiner Dissertation habe ich eine systematische Übersicht über die methodischen Herausforderungen dieses Themas erstellt. Es konnte gezeigt werden, dass bei der Verwendung von Datensätzen mit wiederholten Messungen die Wissenschaftler oft nur einen Zeitpunkt für den Schätzdatensatz und einen weiteren für dessen Validierung verwenden und somit ignorieren, dass Gesundheitszustände mit unterschiedlichen Schweregraden nur zu einem bestimmten Zeitpunkt vorliegen können.

In Artikel 3 dieser Dissertation werden ein PROM (EQ-5D-3L) und ein vom Arzt angegebenes Maß für die globale Behinderung (mRS – modified Rankin Scale) verwendet, um eine Kosten-Nutzwert- bzw. eine Kosten-Wirksamkeits-Analyse durchzuführen. Die G-Formel wurde verwendet, um die zusätzlichen Kosten und die zusätzlichen Ergebnisse aufgrund der Entsendung einer mobilen Schlaganfallstation (MSU) zu schätzen. Es zeigte sich, dass der zusätzliche Einsatz der MSU zu zusätzlichen Kosten in Höhe von €40.984 pro qualitätsbereinigtem Lebensjahr und €81.491,49 pro Überleben ohne Behinderung führte (unter Verwendung einer dichotomisierten mRS).

In dieser Dissertation werden verschiedene Methoden im Bereich der Gesundheitsdatenwissenschaften vorgestellt, darunter die Entwicklung eines Mapping-

Algorithmus, die Durchführung einer systematischen Übersicht und die Durchführung einer wirtschaftlichen Bewertung.

3. Introduction

The definition of health has evolved over time, and it is not limited to the sole absence of disease. In fact, the World Health Organization (WHO) currently classifies health as ‘a state of complete physical, mental and social well-being’[1]. Furthermore, the WHO defines ‘people-centredness’ as a fundamental goal of any health system[2]. Thus, the WHO increasingly helped to shift the healthcare systems’ focus from purely medical aspects to a more people-centred approach to care.

Patient-reported outcomes (PROs) play a crucial role in patient-centred care, as they can provide a more holistic picture of a patient's health beyond their clinical symptoms[3]. PROs are tools that allow patients to give their own perspective on aspects such as their health, health-related quality of life (HRQoL), or functional status, instead of relying on clinical values or assessments by clinicians[4]. They are increasingly recognised by healthcare providers, payers, and policymakers[5]. PROMs (patient-reported outcome measures) are the tools employed to measure PROs and are mostly self-completed surveys[6]. Both the European Medicines Agency and the Food and Drug Administration have accepted and even encourage the use of PROMs as measures of treatment efficacy[7].

We can broadly distinguish disease-specific PROMs from generic PROMs. As the name indicates, disease-specific PROMs were created to measure the symptoms and impact on the function of a specific illness. Sometimes generic measures fail to account for the particularities of specific illnesses and might not be able to detect significant treatment effects if the effects are not very sizable[8]. While disease-specific measures display greater face validity and credibility, generic measures allow comparisons across conditions[9].

Furthermore, PROMs can also be grouped into preference-based measures (PBMs) and non-PBMs. PBMs assess subjects' health status and are then weighted using health state utility values (HSUVs). HSUVs are scores derived from samples from the general population (so-called 'societal preferences'), usually ranging from 0 (death) to 1 (perfect health). As PBM instruments assign different weights to each dimension or item based on general population preferences, they enable comparisons of effectiveness between healthcare interventions for various conditions[10]. These include the EQ-5D, Short Form-6D (SF-6D), Quality of Well-Being Scale (QWB), Health Utilities Index (HUI), Child Health Utility instrument (CHU9D), and others. PROMs that are not PBMs do not possess a scoring system to value the preference that individuals assign to a particular health state.

In multiple jurisdictions, policymakers require effects used in economic evaluations to be measured with PBMs. However, clinical trials that showed e.g. efficacy of a treatment did not always administer/use the instrument required by the jurisdiction’s policymakers. In order to overcome this issue, several researchers have started developing mapping algorithms. Mapping is a method that enables the prediction of HSUVs using information from disease-specific HRQoL instruments or from generic instruments that do not have a preference-based index score system. Researchers usually employ a dataset containing both participants' answers to surveys that do not have a scoring system, and a PBM, to develop mapping algorithms[11]. PROMs are usually administered at multiple time points (e.g. baseline and follow-up). Therefore, mapping algorithms must be developed and validated taking into consideration that when measurements from the same subjects over time are present, the hypothesis imposed by standard models, that observations are independent, is violated, and thus standard errors will be underestimated.

PROMs can play an important role in health economic evaluations. Health economic evaluations can assist decisionmakers and healthcare professionals to make evidence-based decisions on how to best allocate their usually limited resources. They constitute a standardised way to measure and value the costs and the effects of different choices. Healthcare expenditure — both public and private — comprises a substantial portion of the Gross Domestic Product (GDP) in numerous countries. In 2020, European Union Member States spent on average 10.90% of their GDP on healthcare. In Germany, this sector represented 12.82% of the GDP[12]. This highlights the necessity of evaluating resource allocation and ensuring that treatments are cost-effective and provide value for money.

Economic evaluations compare costs and outcomes of alternatives, and the category of outcome used in an economic evaluation defines its type. Different measurements can constitute an outcome used in an economic evaluation. In cost-effectiveness analyses, the outcomes can be clinical measurements (e.g. systolic blood pressure) and clinician-reported outcomes (e.g. Unified Parkinson's Disease Rating Scale (UPDRS)) which are usually condition-specific or PROMs that are not PBMs (e.g. Short-Form 12), which can be condition-specific or generic. In stroke trials, the modified Rankin Scale (mRS) is one of the most commonly used tools to assess functional outcomes[13]. Nevertheless, it is subject to shortcomings that should be considered[14,15]. Cost-utility analyses are economic evaluations where the outcome are utilities stemming from PBMs (e.g. EQ-5D-5L and Short-Form 6D (SF-6D)).

Although PROMs can have drawbacks and it may not even be possible to collect them directly (e.g. only through a proxy), their importance is undeniable. In the context of increasing patient involvement in healthcare decision making, PROMs remain a critical tool in our healthcare

research arsenal to place the patient at the heart of the healthcare process. Thus, it is crucial to overcome their methodological limitations.

This cumulative Thesis comprises three Research Projects that draw on each other to contribute towards a better understanding of how PROMs can be used successfully to inform decision-making in the healthcare sector.

- In Thesis Article 1 (Oliveira Goncalves et al, 2021[16]) we developed a mapping algorithm to convert values from a condition-specific PROM (HIT-6) to utility scores from a PBM (EQ-5D-5L) for Germany and raised concerns about the limitation of this process.
- Thesis Article (Gonçalves et al. 2022[17]) builds on the aspiration to broaden the knowledge in the field of mapping, specifically regarding the use of datasets with multiple measurements of the same subjects over time. Despite a plethora of guidance regarding the conduction and reporting of mapping algorithms, little attention has been paid to how authors handle datasets with multiple observations per subject over time.
- Thesis Article 3 (Gonçalves et al. 2023[18]) draws on knowledge gained from Thesis Articles 1 and 2. It uses a PBM (EQ-5D-3L) and a clinician-reported measure of global disability (mRS) to conduct a cost-utility and a cost-effectiveness analysis, respectively.

This cumulative dissertation is submitted to the Medical Faculty Charité – Universitätsmedizin Berlin within the PhD in Health Data Sciences programme.

4. Methods

4.1 Research Project 1 – Mapping algorithm development

4.1.1 Data

The data used in this project stem from the SMARTGEM study (DRKS-ID: DRKS00016328). SMARTGEM was a randomised controlled clinical trial (RCT) that sought to determine if an intervention involving the usage of an app for headache (M-sense) together with online consultations could reduce the occurrence of migraine. The intervention comprised a web-based instrument where participants could interact with other participants as well as with clinicians, together with a certified app where participants recorded attacks, the intake of medication, and trigger factors in an electronic diary. This app analysed this information and suggested individualised care plans. Participants were recruited between January 2019 and December 2020 and were followed up over a 12-month period. For this study, we used data from this trial collected until 7th May 2020.

Answers to both EQ-5D-5L and HIT-6 were collected at baseline and at 3, 6, 9, and 12 months. Further details on this trial can be found elsewhere[19].

4.1.2 Instruments

The EQ-5D-5L is a generic HRQoL and preference-based measure created by the EuroQol Group, which assesses HRQoL in five different dimensions (‘Mobility’, ‘Self-care’, ‘Usual activities’, ‘Pain and discomfort’, as well as ‘Anxiety and depression’)[20]. Each one is characterised by five corresponding response degrees of severity: ‘no problems’, ‘slight problems’, ‘moderate problems’, ‘severe problems’, and ‘unable to or extreme problems’. EQ-5D-5L health states can be represented by an index value, which indicates how excellent or terrible a health state is based on the preferences of a specific country or region's general population. In our study, the EQ-5D-5L index values were computed with the social health status preference valuation (value set, also called preference-based values, utilities, or weights) for the German population. It can range between -0.661 and 1, where 1 denotes ‘full health’, 0 corresponds to being dead and values below 0 relate to health states perceived as worse than being dead[21]. These value sets allow the conversion of each health state to a single score.

It also includes the EQ Visual Analogue Scale (VAS). The EQ VAS measures respondents’ self-rated health condition on a vertical VAS (range 0-100) where the upper and lower bounds are named ‘The best health you can imagine’ and ‘The worst health you can imagine’.[20]

In this study, the EQ-5D-5L constitutes the target measure for mapping.

HIT-6 is a HRQoL measure specific for headache, which comprises six questions, with five levels (‘never’, ‘rarely’, ‘sometimes’, ‘very often’, and ‘always’). Its score can range between 36 and 78, where higher scores indicate lower functioning levels, i.e. higher disability[22].

Since the HIT-6 is not a PBM, its score is not weighted with societal preferences, as in the EQ-5D-5L case.

4.1.3 Overlap between the two questionnaires

The degree to which EQ-5D and HIT-6 questions are associated was assessed by calculating correlation coefficients which took multiple observations into consideration.

By calculating standardised response mean(s) (SRM), we also assessed each tool's responsiveness — its ability to identify changes in HRQOL over time. SMR can be computed by dividing the mean score change by the change’s standard deviation. According to Cohen's criteria, SMR values can be categorised as follows: >0.2 is ‘small’, 0.2—0.5 is ‘moderate’, and >0.8 is ‘large’[23].

We carried out an exploratory factor analysis (EFA). The aim of the EFA was to investigate if the two questionnaires' underlying constructs overlapped. We assumed that the two questionnaires (EQ-5D and HIT-6) have the same underlying latent structure if the same factors have considerable loadings from both. Factor loadings over 0.3 were deemed ‘meaningful’[24].

When dealing with ordinal data, the appropriate way for determining how many factors shall be considered is to run parallel analyses using polychoric correlations rather than Pearson correlations[25]. Owing to the categorisation, it is thought that Pearson correlations may underestimate the link between ordered categorical data[26]. Additionally, Glorfeld et al. demonstrated that parallel analysis works with data that are not normally distributed[27]. Weighted least squares, which require no distributional assumptions and are suitable for ordinal data, were chosen as the factoring mode[28]. Factor loadings were interpreted using both an orthogonal rotation (varimax) and an oblique rotation (promax).

4.1.4 Modelling techniques

The development of mapping algorithms entailed estimating the relationship between the target (EQ-5D-5L) and the source instrument (HIT-6) using regression techniques.

Mapping guidelines do not advocate for a particular statistical model[11]. Instead, they highlight which factors shall be assessed to select a model, such as the existence of large spikes in the

utilities' distribution (which often occurs at the full/perfect health upper bound), the degree of skewness, gaps in the possible value range, and multimodality. These aspects can vary according to specific elements such as the utility measure to be mapped, the condition, and the patient population.

Thus, we computed different models. We fitted mixed-effects linear regression models, mixed-effects Tobit models, limited dependent variable mixture models, mixture beta regression models, as well as two-part models.

Linear regression models

Due to the multiple observations' nature of our data (individuals replied to the questionnaires multiple times over the study period), we had to account for the observations' interdependence. Thus, we included random-effects in mixed-effects linear regression models.

Tobit models

We carried out a mixed-effects Tobit model censored at the upper bound (corresponding to the value '1').

Two-part models

In the first stage of the two-part model, we fitted a mixed-effects logistic regression to predict the likelihood that a participant would be in full health. A mixed-effects linear regression that only included those who were not in full health was computed in the second stage. An expected value technique was employed to determine the overall expected EQ-5D index value[29].

$$E(EQ-5D) = \Pr(full\ health) \times (EQ-5D\ in\ full\ health) + (1 - \Pr(full\ health)) \times (EQ-5D\ not\ in\ full\ health)$$

where E stands for expected and Pr stands for probability.

Adjusted limited dependent variable mixture models and mixture beta regression models

We carried out adjusted limited dependent variable mixture models (ALDVMM) and mixture beta regression models. Stata commands for ALDVMMs and mixture beta regression models were created explicitly to work with the specificities of health utility data[30,31]. These models enable the dependent variable to be restricted to the EQ-5D-5L country-specific threshold, at the same time accounting for the break between perfect health (1) and the following possible score (in Germany this value corresponds to 0.974).

We fitted these models both with and without taking into account the truncation point, and with and without considering a probability mass at full together with the truncation point (only in the beta mixture models). A 'truncation point' corresponds to the 'next feasible value after full health'[32]. In what concerns the EQ-5D-5L for Germany, this corresponds to the value of 0.974.

As mixture models tend to have several optima, we employed the predicted parameters from a constant-only model in our regressions' models to identify the global maximum[30]. In contrast to the other models we developed, we were unable to incorporate random-effects to account for multiple observations. Nevertheless, we calculated robust cluster-corrected standard errors.

Response mapping

We also aimed to conduct response mapping. Response mapping includes two steps: first the probability of being in one of the five levels of the EQ-5D-5L's five domains must be estimated. For this purpose, we planned to use a random-effects generalised ordered probit (using the regoprob command in Stata), fitting five different regression models. This method does not require a parallel line assumption (the assumption that the coefficients for the independent variables across the different categories are the same) as do standard ordered probit or logit models and considers the ordinal nature of the EQ-5D levels (the multinomial logit model does not require the parallel line assumption but ignores the ordinal nature of data)[33]. The expected EQ-5D-5L index value for Germany would be then calculated using the expected value method[34]. This method overcomes possible biases that could arise from merely choosing the level with the highest associated probability[34].

$$E(EQ-5D) = 1 - (Prmo2 \times 0.026) - (Prmo3 \times 0.042) - (Prmo4 \times 0.139) - (Prmo5 \times 0.224) - (Prsc2 \times 0.05) - (Prsc3 \times 0.056) - (Prsc4 \times 0.169) - (Prsc5 \times 0.26) - (Prua2 \times 0.036) - (Prua3 \times 0.049) - (Prua4 \times 0.129) - (Prua5 \times 0.209) - (Prpd2 \times 0.057) - (Prpd3 \times 0.109) - (Prpd4 \times 0.404) - (Prpd5 \times 0.612) - (Prad2 \times 0.03) - (Prad3 \times 0.082) - (Prad4 \times 0.244) - (Prad5 \times 0.356)$$

where Pr stands for probability, mo stands for 'Mobility', sc stands for 'Self-care', ua stands for 'Usual activities', pd stands for 'Pain/discomfort', and ad stands for 'Anxiety/depression'. The number after each dimension's name corresponds to the level (1 stands for 'no problems', 2 for 'slight problems', 3 for 'moderate problems', 4 for 'severe problems', and 5 for 'unable to/extreme problems').

4.1.5 Model specification

We contrasted models in which independent variables were either the overall HIT-6 score or each of HIT-6 six questions.

We pre-defined age, sex, and chronification stage as variables that had to be part of the model. We only took into consideration sex and age as potential socio-demographic determinants because they are commonly gathered in studies and mapping techniques are designed to be used by other researchers. Age affects migraine symptoms, as evidenced by a decline in the

incidence of photophobia and phonophobia[35]. Three times as many women as men get migraines, and it is well recognised that changes in female hormones have a significant impact on this relationship[36,37]. Thus, we investigated if there was a sex-age interaction given that the effects of migraines vary with age, particularly in women[35]. Additionally, we took into account the patients' migraine chronification stage (chronic or episodic) and the interaction with the variable age. Exploring the addition of a variable of interaction between chronification stage and age was important since migraine features may change over time (e.g. from episodic to chronic migraine)[38]. The Stata package ‘Global Search Regression’ gsreg for automatic selection of variables was used[39]. We used BIC (Bayesian information criterion) as a criterion for variable selection, whereby the lower the BIC, the better the model performed.

All mentioned analyses were based on complete-cases with respect to the variables EQ-5D domains, HIT-6 domains, sex, age, as well as chronification stage.

4.1.6 Model validation

To analyse the performance of each model, we displayed the observed as well as the predicted EQ-5D-5L scores with scatterplots. Models’ validation was carried out using tenfold cross-validation. It was not possible to conduct external validation due to the lack of external data. However, for small samples, cross-validation performs well[40]. Mean absolute error (MAE), root mean squared error (RMSE), and R2, averaged across 10 cycles, were employed to evaluate the prediction ability of the distinct models.

Analyses were carried out using R 3.6.3[41] and Stata 15.

4.2 Research Project 2 – Systematic review of mapping algorithms

The PROSPERO registration number for this systematic literature review is CRD42020188130.

The methodology for this research project, complying with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, is described in the sections that follow.

4.2.1 Inclusion and exclusion criteria

Studies that created mapping models to predict generic HSUVs from both PBMs and non-PBMs were taken into consideration.

We focused on studies that employed statistical approaches to build novel mapping algorithms that other researchers may use (i.e. regression techniques).

Hence, the following studies were excluded: studies that employed ‘judgement-based mapping’; records only using and/or testing already developed mapping functions; records that directly predicted HSUVs that resulted from valuation methods like time trade-off or discrete choice experiments; records that mapped to disease-specific PBMs, and conference papers or abstracts without an accessible complete-text publication. We also disregarded systematic reviews, although they were utilised for manually searching for further relevant studies.

4.2.2 Data sources and search strategy

Our systematic search was conducted in MEDLINE (via Ovid), the Cumulative Index to Nursing and Allied Health Literature (CINAHL), the database of mapping studies from the University of Oxford, and the ScHARRHUD database from the School of Health and Related Research, University of Sheffield.

Initial searches were carried out on 1st March 2021. An update search was performed on 13th December 2021. The used search strings are shown in the Supplementary Material 1 of Oliveira Goncalves et al. 2021[17].

We further searched for additional records in reference lists included in already published mapping reviews[42,43] and on a specialised website (EuroQoL). We considered studies in English and German.

4.2.3 Selection process

Records were uploaded to the reference manager Paperpile® and were screened for eligibility in an independent way by two reviewers. First, titles as well as abstracts were inspected against our pre-defined criteria of inclusion and exclusion. In order to decide which research should be included in the systemic literature review, the full texts of all possibly relevant studies were subsequently reviewed. Disagreements were resolved through debate or by engaging an additional reviewer to establish a consensus.

4.2.4 Extraction of data

We then proceeded by extracting data from all qualifying studies using a standardised extraction template. Information from the included records was gathered by the same two reviewers. Multiple checklists guided the development of the extraction matrix: the ‘Mapping to estimate health-state utility from non-preference-based outcome measures’ report[11], and the MAPS checklist[44].

The data extraction template contained bibliographic details, the source instruments, the target PMBs, and the jurisdictions’ used weights. In what concerns information on methods, we retrieved data on the used mapping methods (direct or indirect mapping), regression methods, and whether validation (including the type of validation) was carried out. In terms of the dataset employed for estimation, we extracted details regarding the jurisdiction where the questionnaires were handed out, and the sample population. Given that our primary interest was to investigate how manuscripts take into account multiple observations when the outcome was measured more than once per individual, we retrieved details on whether mapping models were fitted employing exclusively data at one time point from a longitudinal study (e.g. just baseline) or the entire dataset; the number of considered time points; if specific time points from the dataset were just used for training or for validation; which statistical analysis approaches were utilised; and which (if any) adjustments were carried out for multiple observations. We further extracted data if the variance-covariance matrix was made available by authors, since standard errors will be different if researchers adjust for multiple observations. Moreover, we extracted data on whether the authors of these studies did not recommend mapping. We carried out descriptive statistical analysis with R 4.0.2[41].

4.2.5 Assessment of the quality and bias of included publications

The goal of this systematic literature review was to present an overview of common practice, namely how researchers handle multiple observations per subject over time, rather than to

evaluate the general mapping publications' quality. As a result, assessing the general quality of mapping studies was outside the purview of this study.

4.3 Research Project 3 – Health economic evaluation

4.3.1 Patients, setting, and study design

The data used in this project stem from the B_PROUD study, a prospective, non-randomised, controlled intervention study carried out in Berlin (Germany), from 1st February 2017 to 30th October 2019[45].

The B_PROUD study was approved by the ethics committee of the Charité – Universitätsmedizin Berlin on 2nd September 2015. In accordance with German data protection law and with the approval of the Berlin data protection representatives, patients were notified one month in advance about the planned follow-up assessment, which took place three months after the index event. They were also warned that they could opt out at any time before or during the telephone interview or from participation in the alternative questionnaire-based assessment.

Briefly, individuals were included in the B_PROUD primary population if they had a diagnosis of acute cerebral ischaemia (ischaemic stroke, according to the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, ICD-10: I63 or transient ischaemic attack, ICD-10: G45, excluding G45.4), and were potentially suitable candidates for either thrombectomy or thrombolysis. Further inclusion criteria for the B_PROUD study included being over 17 years old; the emergency call needed to lead to an MSU mobilisation code throughout operating hours of the MSU; stroke onset within four hours since the emergency call; occurrence of the index event within the boundaries of the MSUs' catchment areas; ability to have been ambulatory prior to the stroke event (a proxy of mRS ≤ 3), and without symptom resolution before the arrival of the emergency medical services. The primary research population did not include any other subtypes of stroke. Since no changes in short-term outcomes were detected in previous research, those with a definite non-stroke diagnosis or stroke patients who were not qualified for recanalising therapy were excluded[45]. Hence, it was hypothesised that the mobilisation of MSUs did not affect the outcomes of patients belonging to these groups.

Exposure status was established using the type of ambulance mobilisation (with MSU mobilisation or without MSU mobilisation), similar to the intention-to-treat concept in a RCT. Allocation to one of the groups was dictated by the availability of an MSU at the time of the index stroke, thus creating a natural experiment setting. Specifically, three MSUs were introduced throughout Berlin over a 17-month period. The B_PROUD study started with a single MSU, and after a five-month roll-in phase, a second MSU was introduced on 1st September 2017, followed by a third MSU on 1st September 2018. On 1st September 2018, Global Positioning System Tracking was added. Subsequently, the geographically nearest available MSU was sent to the

event location. Patients with ischaemic stroke and transient ischaemic attack, for whom an MSU and a regular ambulance were mobilised, made up the intervention group (n = 749). The comparator group included patients for whom exclusively a standard ambulance was mobilised (n = 794).[5] Additional information about B_PROUD's design, participant selection, and complete characteristics of the study population have been presented elsewhere[45].

4.3.2 Outcomes

In our cost-utility analyses, the HRQoL outcome was assessed with the three-level version of EQ-5D developed by the EuroQol Group (EQ-5D-3L), using the German Visual Analogue Scale (VAS) weights[46]. This instrument consists of five dimensions, as described in section 4.1.2. It was administered via telephone interviews or paper questionnaires by qualified research nurses. The EQ-5D-3L was administered to patients at the three-month follow-up. Previous research analysing shifts in quality of life according to the severity of stroke symptoms showed that the differences in survival probabilities and EQ-5D-3L values after the initial months remained mostly unaltered over a five-year timeframe[47]. Therefore, we have defined five years as the time frame in our study. Our time range enabled us to consider care costs, which remain a financial burden after the initial 3-month follow-up (see section 4.3.3). Although HRQoL was not collected in the Luengo-Fernandez et al. study at a three-month follow-up like in B_PROUD, the B_PROUD information falls within the one- and six-month follow-up intervals used by Luengo-Fernandez et al. [47]. We calculated Quality-Adjusted Life Years (QALYs) for the aforementioned five-year period by multiplying EQ-5D-3L scores in month three by five. We acknowledge that this projection assumes that each patient's EQ-5D-3L scores stay stable throughout a five-year period. Nevertheless, since the incremental cost-effectiveness ratio (ICER) denominator pertains to the QALYs difference between exposure groups after adjustment for confounding, this procedure was appropriate for the current study. As a result, the key hypothesis in our methodology was that the difference in the number of QALYs owing to mortality or HRQoL changes after adjustment for confounding was equal during the five-year period in both groups.

The mRS at three months was the outcome measure of interest in our cost-effectiveness analyses. The mRS questionnaire measures how impaired or dependent stroke victims are while doing everyday activities. It ranges from 0 ('no neurological symptoms') to 6 ('death'). Patients in the MSU mobilisation group showed significantly lower global disability, as determined by the mRS score, compared to those in the control group, according to a previous B_PROUD study (OR 0.71, 95% CI, 0.58 to 0.86)[45]. This score was dichotomised for the cost-effectiveness analyses: 0-1 ('excellent' or 'survival without symptoms/disability') and 2-6 ('not excellent' or 'survival with symptoms/disability' or 'dead')[48]. When possible, the three-month mRS scores were computed as the median rating of three different neurologists in an independent and blinded

fashion[45]. Unblinded scores conducted by certified nurses were only employed when the audio quality of the recording was too poor for assessment or when patients refused the recording. We hypothesised that the difference in the absolute number of survivals without symptoms/disability between exposure groups, after adjustment for confounding, stayed invariable during the course of five years.

4.3.3 Costs

The relevant cost categories were prospectively recorded by the Berlin Fire Department and the study members within the study period. These comprised prehospital healthcare expenses with medication, prehospital imaging charges (computed tomography and computed tomography angiography), and MSU-related investment and operating expenditures.

We provide both a societal and a statutory health insurance perspective in our analyses. From a societal perspective, we included any incremental MSU-related expenses generated by the Berlin Fire Department (which included value-added tax). In terms of the statutory health insurance perspective, we considered the Berlin Fire Department's fee per MSU mobilisation for reimbursement by statutory health insurance. Here, we assumed that 97% of the mobilisations which involved patient care were certainly chargeable and refunded.

Furthermore, we considered expenses related to long-term care in the five years after the stroke. Using unpublished data from a different sizable stroke study in Germany[49], we translated mRS scores into various care dependency levels named ‘Pflegestufen’ to represent the extent of the need for care. The degrees of care dependency in Germany are defined by the Medical Service of the Health Insurance and are employed to compute long-term care insurance payments for the delivery of care. Due to a change in Germany's classification system starting 2018, we translated the previous ‘Pflegestufe’ degrees to the modern ‘Pflegegrad’ (care grade) in accordance with established tables[50]. We estimated total care service costs by combining the said information about the care grade of each individual with information about each individual’s living status (‘living at home and cared for by a relative; ‘living at home and cared for by a professional’; ‘living in a nursing home’; ‘in the hospital’[18]) collected at three-month follow-up by the B PROUD study team. The supplementary Appendix 2 of Gonçalves et al. 2023 [18] describes the calculation of these costs and the assumptions made.

As suggested by the German Institute for Quality and Efficiency in Health Care (IQWIG), we employed a 3% discount rate for long-term care expenditures in the years after the index event[51]. All expenditures are reported in Euros, with 2019 serving as the reference year.

We assumed three unique scenarios: base-case, best-case, and a worst-case scenario. Briefly, under the base-case scenario, we analysed outcomes for the primary study population[18,45]. As

a result, we considered additional costs related to the MSU mobilisation for all patients with code stroke. Moreover, we hypothesised that the costs of tPA treatments would be identical whether they were administered before or after a hospital stay, and that nursing care costs were determined by the patient's mRS score. The best-case and worst-case scenarios employed the base-case scenario's assumptions with multiple changes. Under the best-case scenario, we calculated long-term care costs taking a less conservative approach when converting 'Pflegestufe' degrees to the modern 'Pflegegrad' ('care grade'). In this scenario's calculation of the care costs, we accounted for 'non-physical' deficits such as mental and communication deficits. See Supplementary Appendix 2 in Gonçalves et al. 2023 [18] for further details. Furthermore, we hypothesised that the regularity of MSU mobilisations could be raised by a factor of 1.8, which corresponded to a higher tPA frequency of treatment per MSU functioning week, as it was observed in the PHANTOM-S trial[52]. Under the worst-case scenario, we conducted a complete case analysis, thus assuming that the effects identified in the B_PROUD study exclusively pertained to subjects with complete three-month follow-up data. We further hypothesised that one-half of the imaging examinations required re-testing in a hospital setting. Finally, we considered that the consequences on the care grade dependency degree and mRS continued to be steady over a period of 18 months (based on the IST-3 observation period), instead of five years[53]. Complete details on the assumptions made under each scenario can be found in Supplementary Appendix 1 in Gonçalves et al. 2023[18].

4.3.4 Statistical methods

This economic evaluation followed the statistical analysis strategy that was registered in OSF[54].

For all analyses, the incremental costs and incremental outcomes attributable to MSU mobilisation were estimated using the parametric G-formula[55]. We first ran a linear regression model with numeric costs as the dependent variable and exposure group, and a set of a priori selected confounding variables (sex, age, diabetes mellitus, atrial fibrillation, arterial hypertension, symptoms of neurological nature at emergency medical services or MSU arrival, and living status before the index event[18]) as explanatory variables. These covariates were also used for adjustment in the assessment of clinical effectiveness in the primary B_PROUD population[45]. The predicted costs for every individual in our two hypothetical scenarios — deploying the MSU together with a regular ambulance to all patients with a code-stroke and deploying a standard ambulance alone to all patients with a code-stroke — were then determined using this model. In the interest of calculating the total costs, we added the predicted costs — obtained in the previous step — across all persons in each hypothetical scenario and calculated the difference between the total costs to get the incremental costs associated with the mobilisation

of MSUs. The incremental QALYs owing to MSU mobilisation together with a standard ambulance were calculated using the same methodology, given that the variable QALYs is also numeric.

Since the dichotomised mRS variable is binary, a slightly different approach was used. To begin with, we fitted a logistic regression to estimate each person's likelihood of surviving without disability (mRS 0-1) under each hypothetical scenario. The total number of subjects surviving without symptoms/disability was then estimated for each hypothetical scenario. Lastly, we calculated the incremental absolute number of survivals without symptoms/disability owing to the mobilisation of an MSU as the difference[18].

We used these incremental total effects (QALYs in the cost-utility analyses; dichotomised mRS in the cost-effectiveness analyses) and incremental total costs to compute ICERs.

Generally, incremental costs and incremental effects are shown as mean incremental costs and mean incremental effects. However, this would not make sense for the dichotomised mRS outcome, which corresponds to the lives saved without symptoms/disability. As mentioned above, in the G-formula, we used a logistic regression model to predict each person’s probability of surviving without disability (mRS 0-1) under each hypothetical scenario. The total number of individuals who survived without having/developing a disability was then calculated. It would not have made sense to compute the mean number of incremental survivals without symptoms/disability. Given that ICERs are ratios, we had to use the same procedure for costs, obtaining incremental overall costs instead of incremental mean costs. For consistency purposes, we followed the same approach for the cost-utility analyses.

In the cost-benefit analysis, we quantified net costs as the difference between the additional costs of MSU mobilisation and the saved care costs. We used the same statistical methods as described above to perform imputation, bootstrapping, and obtain incremental costs and effects. The main result from this analysis was not an ICER but a net monetary benefit. Normally, this is computed by multiplying a country’s willingness-to-pay threshold by the incremental effects and subtracting the incremental costs. However, as there is no official threshold in Germany, we have only considered the difference between the additional costs of MSU mobilisation and the saved care costs.

Our dataset had missing values in the EQ-5D-3L, mRS, patient’s care grade level and living status. Mean costs were computed based on information collected by the Berlin Fire Department, the B_PROUD team of the Charité – Universitätsmedizin Berlin, and by using publicly accessible sources. Thus, there was no missing data. However, there were missing data in the care cost segment since they were calculated using each patient’s details, some of which had missing values. We employed multiple imputation by chained equations with five datasets in order to impute missing data for both the base-case and best-case scenario, assuming a missing at

random data mechanism. The means across the five datasets from the multiple imputation were used for calculating the point estimates for incremental QALYs, incremental survivals without symptoms/disability, and incremental costs. Every variable that had been considered in the regression models in the initial effectiveness publication was also included in the imputation models[45], in accordance with the guidelines for handling missing data in economic evaluations[56].

In order to calculate 95% confidence intervals, we conducted (nonparametric) bootstrapping with a total of 5,000 iterations following the BootMI approach[57]. After multiple imputation on every bootstrapped dataset, we estimated the upper and lower bounds of the confidence interval from, respectively, the 2.5% and 97.5 percentile of the generated distribution of the average metric (across the five datasets that were imputed). We also generated cost-utility and cost-effectiveness planes to display the bootstrapped cost-utility and cost-effectiveness pairings that originated from the bootstrapping iteration runs, illustrating the joint uncertainty around outcomes and costs[58]. Data points falling in the ‘northeast’ quadrant of these planes suggest that the exposure produces greater health gains while being more expensive, whereas points in the ‘southwest’ quadrant indicate that the intervention produces fewer health gains but is less expensive. An exposure yielding both higher health gains and lower costs (‘southeast’ quadrant) is regarded as an economically ‘dominant’ option. On the opposite, northwest points are deemed ‘dominated’, because they correspond to lower health gains and higher costs. In our tables, we further reported the proportion of data points falling into each of the four quadrants, to give a clearer picture of the number of incremental cost-effect pairs per quadrant.

In addition to the different scenarios considered, we carried out a sensitivity analysis to apply a different approach to adjust for confounding, where the models were exclusively adjusted for the geographic coverage of the MSUs. This variable was created based on the absolute number of MSUs which were covering a geographic location (zip code) during an exact time period (one-fourth of the calendar year) at the time of the patient’s index event (because of the overlap of MSUs’ catchment locations). We hypothesised that no changes occurred in the MSU catchment zones after the GPS system was installed. We employed information from all variables used in the main analysis, as well as the MSU coverage variable, to generate the imputed datasets for these analyses.

Analyses were conducted with the software R version 4.0.3[41].

5. Results

5.1 Research Project 1 – Mapping algorithm development

5.1.1 Missing data and descriptive statistics

Casewise deletion was conducted considering the variables related to HIT-6, EQ-5D-5L, chronification stage, sex, and age. The variable chronification stage (episodic versus chronic), reported by the physicians in the study, had no missing values. Furthermore, there were also no values missing in terms of the age of participants.

Twenty-two out of 1,032 observations were removed, which corresponded to 16 patients with missing data. Seven patients out of 16 were not considered in the analyses since they lacked complete data on other time periods. The dataset finally used in our analyses contained 1010 observations, corresponding to 410 patients.

The vast majority of participants were female (87.3%) and overall, participants were on average 41.1 years old. See Table 1 in Oliveira Goncalves et al. 2021 for a detailed overview of patients' socio-demographic characteristics[16].

Figure 1 and Figure 2 depict the distributions of the two instruments.

EQ-5D-5L scores ranged from -0.57 to 1. Nobody had the worst health state as measured by this questionnaire (−0.661). The EQ-5D-5L data show a mass point at the upper perfect-health bound: 194 out of 410 participants reported having perfect/full health (a score of 1) in at least one time period. Figure 1 shows a left skew amounting to -2.33, with a corresponding kurtosis amounting to 9.45. Those with episodic migraine have significantly more skewed data than patients with chronic migraine. The average EQ-5D-5L index value for all participants amounted to 0.82 (SD 0.23), 0.86 (SD 0.18) n the case of episodic migraine subjects, and 0.72 (SD 0.30) for subjects with chronic migraine. We did not find a significant mass of data points at 0.974, which corresponds to the truncation point.

In terms of the HIT-6, values varied from 44 to 78. No participants reported the best headache-specific health state (36). This phenomenon makes sense, especially at baseline, given the trial's inclusion criteria (patients had to report at least five migraine days over the 28 days preceding the screening visit to be included in the trial). Opposite to the EQ-5D-5L, HIT-6 data did not show large ceiling effects. While data was also skewed to the left, the absolute skewness level was much lower than in the EQ-5D-5L case (−0.64 vs −2.33).

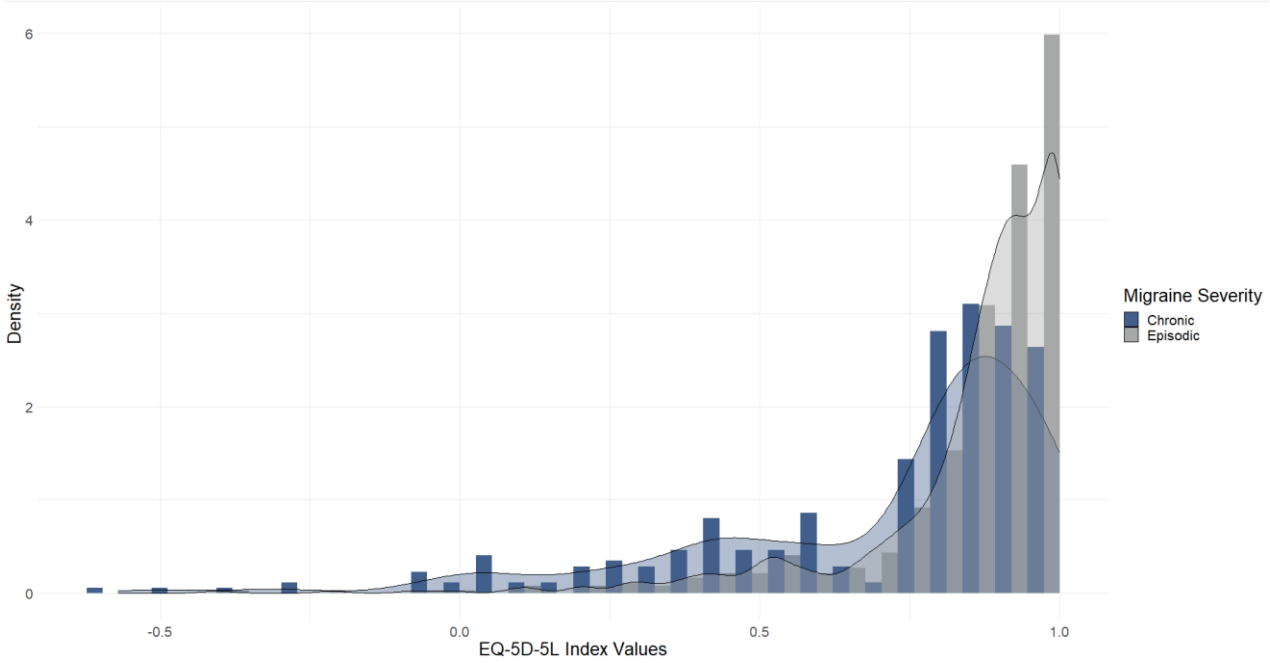


Figure 1 EQ-5D-5L number of responses histogram and kernel density plot by migraine severity (modified from Oliveira Goncalves et al., 2021 [16]*)*

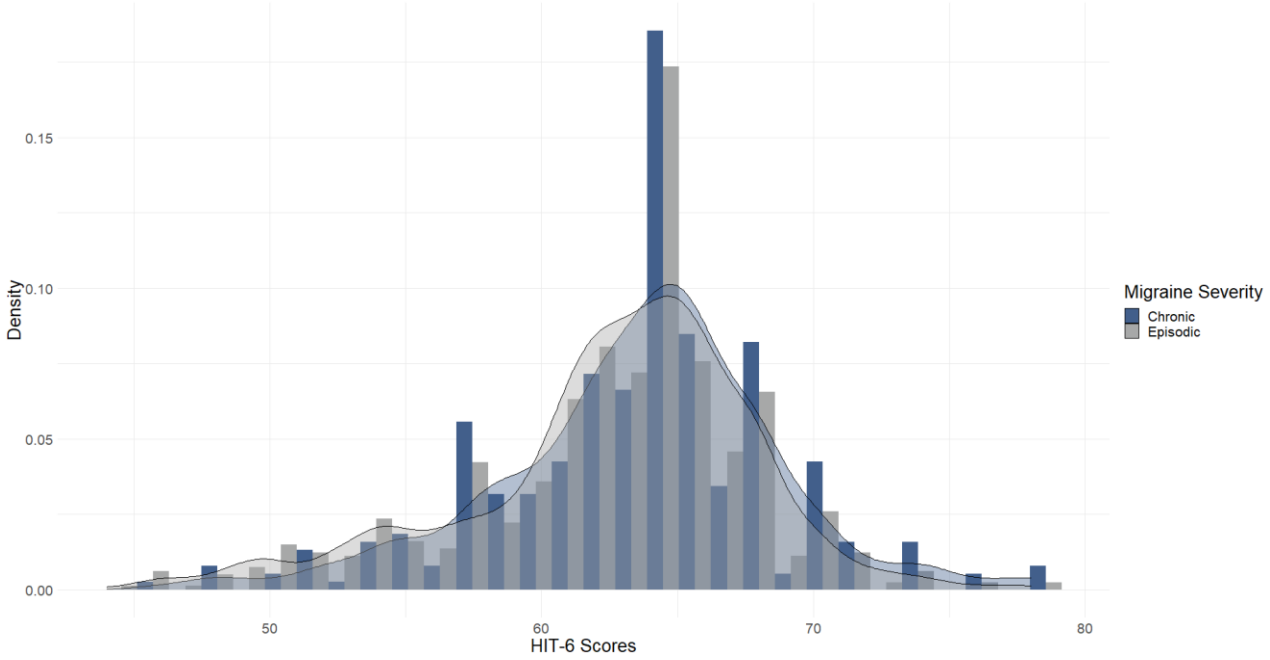


Figure 2 HIT-6 number of responses histogram and kernel density plot by migraine severity[16]*(modified from Oliveira Goncalves et al., 2021* [16]*)*

Headache Impact Test-6 (HIT-6)

5.1.2 Conceptual overlap

We had assumed that both daily activities and occupation could be captured by questions 2, 3 and 4 from the HIT-6 and by the EQ-5D-5L dimensions 'Usual activities' (see Table 1 below). Physical health could be measured with question 5 from the HIT-6 and with the dimensions 'Pain/discomfort' and 'Mobility' from the EQ-5D-5L. Self-care would be only captured by the EQ-5D-5L dimension 'Self-care'.

Table 1 Theoretical conceptual overlap between EQ-5D-5L and HIT-6 (own representation: Oliveira Gonçalves)

	Occupation/ Daily Activities	**Physical Health**	**Mental Health**	**Self-Care**
HIT-6	Question 2, Question3, Question 4	Question 1	Question 5	-
EQ-5D	'Usual activities'	'Pain/discomfort', 'Mobility'	'Anxiety/depression'	'Self-care'

Headache Impact Test-6 (HIT-6)

Besides the expected 'theoretical' overlap, we computed correlation coefficients. The correlation coefficient between the HIT-6 and the EQ-5D-5L overall scores was −0.30. When looking at the EQ-5D-5L score and the multiple HIT-6 questions, coefficients ranged from −0.153 to −0.234. The correlation coefficient between each EQ-5D-5L dimension and the HIT-6 overall score ranged between 0.077 and 0.300. Supplementary Tables A.1, A.2, and A.3 in Oliveira Goncalves et al. 2021[16] show correlation tables, including correlation coefficients stratified by migraine severity level.

In terms of responsiveness, the HIT-6 score and each question showed little to moderate responsiveness, while both the EQ-5D-5L score and each dimension were associated with low SRMs. While for the HIT-6 questions SRMs ranged from 0.211 to 0.669, these values were lower for EQ-5D dimensions (0.088 to 0.280) (see Supplementary Table A.4 in Oliveira Goncalves et al. 2021[16]). Even though the low level of responsiveness could be partially due to the inclusion of patients in the control group in the dataset, it is not clear why the HIT-6 is more sensitive to this measure than the EQ-5D.

We used three factors for the EFA. In what concerns Factor 1, it showed meaningful loadings (i.e. higher than 0.3) for all EQ-5D-5L dimensions, but not for HIT-6 questions. Factors 2 and 3 solely loaded HIT-6 questions: questions 2 to 6 for Factor 2, and questions 1 and 2 for Factor 3. Question 2 from the HIT-6 loaded in both Factor 2 and 3, with a higher load in Factor 2. Question 3 from the HIT-6 also loaded in both Factor 2 and 3 but showed a higher value in Factor 3. Overall, Factor 2 had meaningful loadings in all but one HIT-6 question (question 1 only had a meaningful load in Factor 3) (Table 2 in Oliveira Goncalves et al. 2021[16]). We obtained similar results when using a different rotation: all EQ-5D-5L dimensions loaded on the same factor, whereas HIT-6 questions loaded on both Factor 2 and Factor 3 (Supplementary Table A.5 in Oliveira Goncalves

et al. 2021[16]). These analyses did not take into account the fact data had multiple observations per subject over time. Thus, we conducted a sensitivity analysis where we exclusively used baseline observations. Our results did not considerably change concerning meaningful loadings and the number of factors.

Overall, the EFA results showed a lack of overlap between EQ-5D-5L and the HIT-6, suggesting that the two instruments may not measure the same latent constructs[16].

5.1.3 Mapping algorithms with different models

Detailed information on each model's coefficients and predictive performance can be found in the Supplementary Material of Oliveira Goncalves et al. 2021[16]. Overall, using an identical statistical approach, models that incorporated the HIT-6 overall score outperformed those that incorporated each HIT-6 item as explanatory variables. The addition of interaction variables (between sex and age, chronification stage and age, and chronification stage and sex) was shown not to markedly enhance EQ-5D score prediction in any of the developed models. On the other hand, the inclusion of quadratic terms for both the HIT-6 total score and multiple HIT-6 dimensions improved the goodness-of-fit of multiple models. The first stage of the two-part model contained just the overall HIT-6 score, the chronification stage, sex, and age; the second stage contained the same variables and the HIT-6 overall score quadratic term.

We could not carry out response mapping as it was planned. Our dataset contained few replies at the worst response levels, thus rendering it inadequate for response mapping, which requires a considerable number of observations in each response grouping[32].

Figure 3 depicts the observed and the predicted EQ-5D-5L scores for each mapping model. As is usual in mapping functions, our estimated models underestimated utility values for people with worse health states and inflated them for people with better health states[42]. While linear regression models can generate estimates greater than one (since it does not consider an upper limit), Model A (mixed-effects) linear regression model with the overall HIT-6 score as an explanatory variable) did not yield predictions that were greater than one. The highest predicted value for Model A was 0.98, while the maximum predicted value for Model B (mixed-effects linear regression model with individual HIT-6 items as explanatory variables) was 1.07.

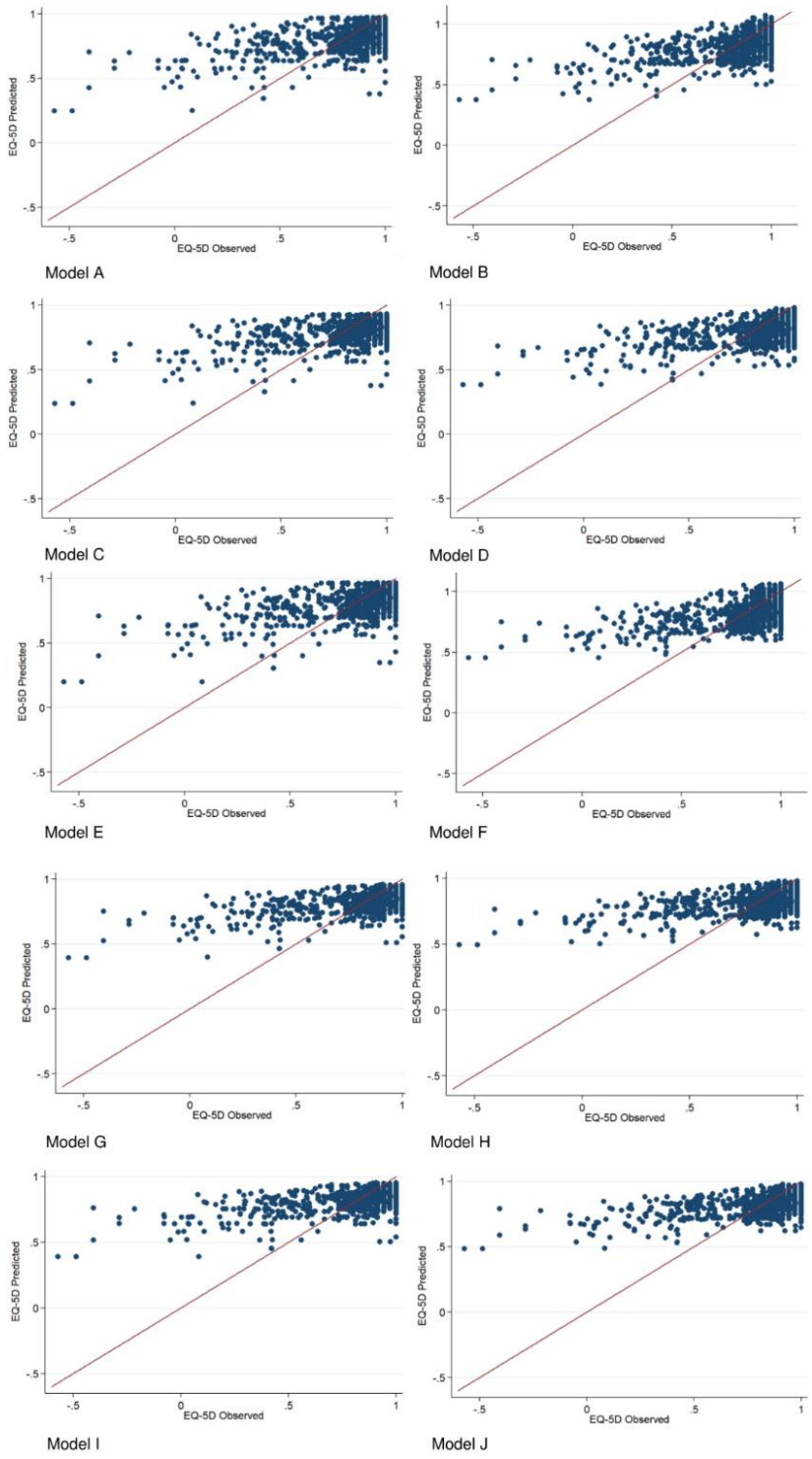


Figure 3 Observed and predicted EQ-5D-5L index values' scatterplots (from Oliveira Goncalves et al., 2021 [16]*)*

Model A – Mixed-effects linear regression, overall HIT-6 score; Model B – Mixed-effects linear regression, individual HIT-6 domains; Model C – Mixed-effects Tobit, overall HIT-6 score. Model D – Mixed-effects Tobit, individual HIT-6 domains; Model E – Two-part model, overall HIT-6 score; Model F – Two-part model, individual HIT-6 domains. Model G – Adjusted limited dependent variable mixture, overall HIT-6 score. Model H – Adjusted limited dependent variable mixture, individual HIT-6 domains. Model I – Beta Mixture Model (with inflation), overall HIT-6 score. Model J – Beta Mixture Model (with inflation), individual HIT-6 domains

There was no single model that outperformed all other models in terms of all goodness-of-fit indicators. Model E (a two-part model using the overall HIT-6 score as an independent variable) fared the best in what concerns the RMSE. Even if Model G has a higher R^2 value, it predicts those in better health and those in worse health less well than Model E. Model I has a higher R^2 value than Model E, although Model E predicts lower health conditions better. The break between perfect health and the following viable health state was included in the ALDVMMs and beta-mixture models. The models' not so strong performance may, however, be explained by the small number of data points (four) with the health state that follows perfect health (0.974).

As a result, if research teams choose to translate index values from the HIT-6 questionnaire for use in cost-utility studies, Model E can be used. We provide the variance-covariance matrix for this model in Table A.6 of the Electronic Supplementary Material of the Oliveira Goncalves et al. 2021[16]. This matrix allows researchers to conduct probabilistic sensitivity analysis, thus accounting for uncertainty. We would want to point out, however, that this mapping procedure should only be utilised as a last option.

5.2 Research Project 2 – Systematic review of mapping algorithms

5.2.1 Search results

Our systematic search yielded a total of 1,344 records from multiple predefined data sources (see section 4.2.2.). After the duplicates' removal, a total of 788 records were selected for title and abstract screening by two reviewers. In this step, 461 records were excluded: 435 did not develop a mapping algorithm; ten had no available abstract; nine were repeated; five were only an abstract or poster; and two were not in English or German. Forty-nine records were subsequently excluded based on full-text screening: 33 did not develop a mapping algorithm; seven did not used regression techniques to calculate mapping algorithms; three were repeated; one did not map from a PROM; one was not clear if a previous algorithm was used or if a new algorithm was developed; one had no full text available. Thus, our systematic review of mapping studies comprised 278 publications. Figure 4 shows the PRISMA flow diagram with the study selection process. This figure is similar to Figure 1 in Gonçalves et al. 2022[17].

Several of the studies that were included employed numerous initial and target instruments, various country weights/value sets, as well as multiple datasets. As a result, instead of providing percentages in the following synthesis of data, we provide the absolute number of studies for each retrieved attribute.

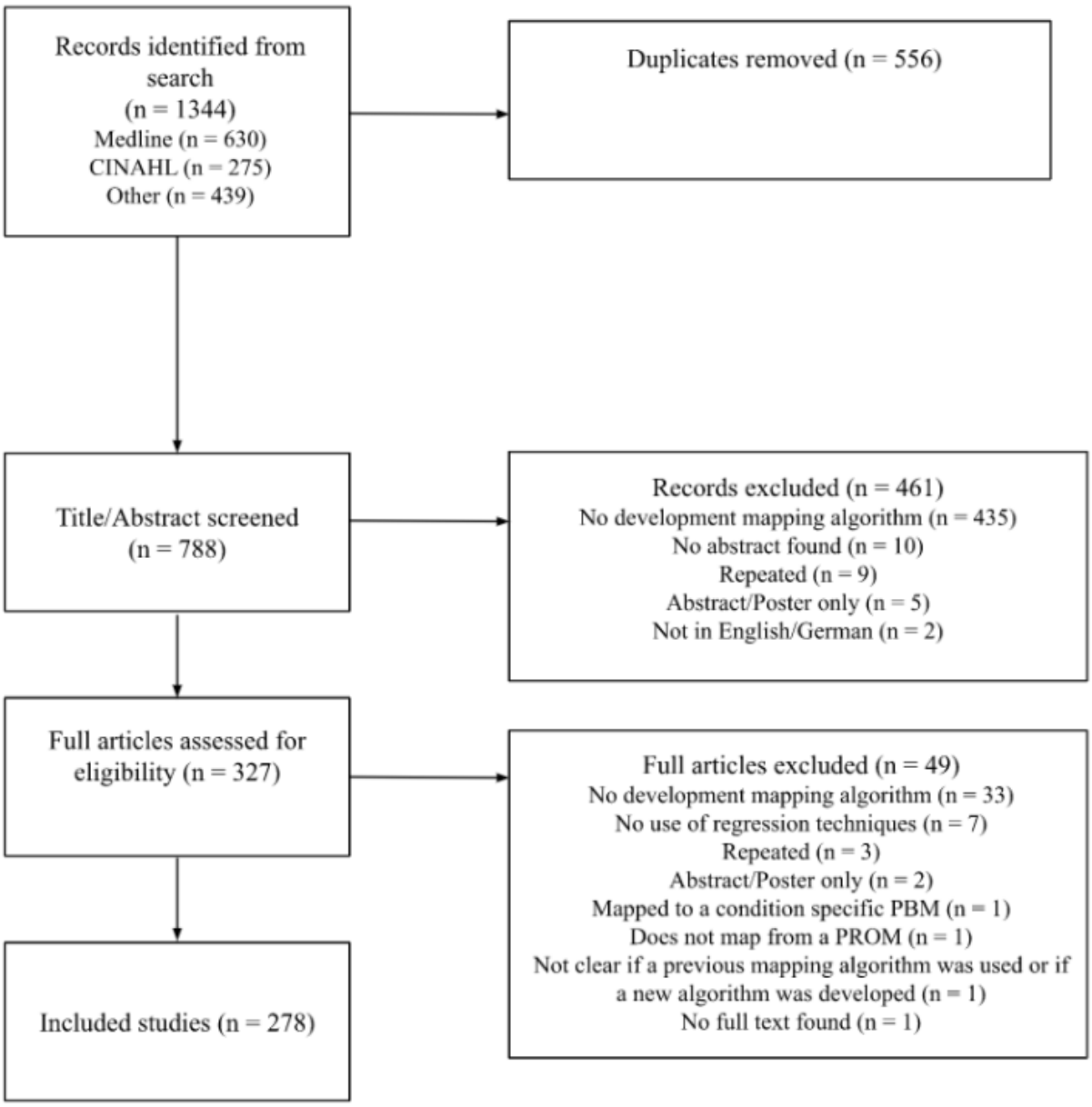


Figure 4 Study selection flow diagram (modified from Gonçalves et al., 2022 [17]*)*

CINAHL - Cumulative Index to Nursing and Allied Health Literature; PBM - Preference Based Measure; PROM - Patient Reported Outcome Measure

5.2.2 General studies' characteristics

The majority of included studies used mapping to translate non-PBMs to PBMs. Numerous studies employed the generic European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ30) (n = 22), whereas some mapped from condition-specific EORTC QLQ modules: colorectal cancer (n = 2), multiple myeloma (n = 1), breast cancer (n = 1), and head and neck cancer (n = 1). Other commonly used start instruments included different versions of the Functional Assessment of Cancer Therapy (FACT) (n = 17), and the Health Assessment Questionnaire (HAQ) (n = 20). Several studies also mapped between PBMs, with the following start PBMs: the EQ-5D-3L (n = 3), the EQ-5D-5L (n = 2), the SF-6D (n = 2), the Assessment of Quality of Life (AQoL) (n = 2), the Quality of Well-Being index (QWB) (n = 2), 15D (n = 1), and the Health Utilities Index-Mark III (HUI3) (n = 1).

Several general quality of life questionnaires without HSUVs were also employed as a start instrument: the Short Form-12 (SF-12) was employed in 12 studies, the Short Form-36 (SF-36) in ten, and the multiple versions of the Patient-Reported Outcomes Measurement Information System (PROMIS) in six. The most used target instrument was the EQ-5D-3L (n = 165). It was followed by the EQ-5D-5L (n = 62), the SF-6D (n = 56), the HUI3 (n = 21), different versions of the AQoL (n = 16), Child Health Utility instrument (CHU9D) (n = 9), the 15D (n = 10), the QWB (n = 5), and the EQ-5D-Y (n = 2).

The most prevalent condition in the study's sample population was cancer (n = 55), as anticipated given the aforementioned widespread use of cancer instruments as start measures, followed by various kinds of arthritis (n = 31). A large number of algorithms (n = 33) relied on responses from the general public.

The Multi-Instrument Comparison (MIC) dataset was often employed (n = 13) to collect responses from mixed-condition patient groups and the general population (n = 3), and patient populations suffering from specific conditions, such as asthma (n = 2), cancer (n = 2), depression (n = 2), diabetes (n = 2), and heart disease (n = 2). The MIC project is an extensive comparative research project of multiple QoL instruments administered in multiple countries. Respondents included a representative healthy cohort as well as patients suffering from conditions in eight clinical areas[59].

Databases from a wide range of nations were employed in the algorithm development studies. Furthermore, datasets occasionally included patients from different countries. The majority of studies used datasets with data from participants in the United Kingdom (UK) (n = 75), the United States (US) (n = 68), and Australia (n = 42).

In what concerns the value set utilised in PBMs, there was less country diversity than in the nations from where the datasets came. The UK value set was employed in the vast majority of studies (n = 161), followed by the US value set (n = 39) and the Canadian value set (n = 34). These categories also comprise studies that employed sets from several countries (n = 40). This happened when researchers mapped to multiple PBMs (n = 20), and when authors mapped to the same PBM (n = 20). The latter only happened for the EQ-5D-3L and the EQ-5D-5L. For some studies, it was not clear which tariff was used for at least one of the mapping functions (n = 7).

5.2.3 Data sources and time points used by the studies

There were 278 studies total, 120 of which only utilised datasets with multiple observations stemming from the same subjects throughout time, and 153 of which only employed datasets with a sole time point. In four studies it was not evident whether there were multiple observations in the used dataset. One manuscript employed multiple datasets, whereby one contained multiple observations per subject, one contained no multiple observations per subject, and for one this information was unclear. Even though we could tell whether multiple observations were utilised for most articles, this information was not always clearly stated in the manuscripts.

In various cases, the descriptions of the dataset(s) in the study were insufficient to identify whether multiple observations were present[60–62]. The objective of this systematic review did not involve examining the level of quality of the publications. Our table in the Supplementary Material 3 in Gonçalves et al. 2022, provides researchers with a complete summary of what each study reported[17].

Various methodologies were utilised in the total of 121 studies which employed a dataset with multiple observations (and in some instances, several strategies in one study). The method used most frequently involved creating mapping algorithms containing all available time points (n = 92) or by utilising baseline values as the single time point (n = 32). Twelve studies employed more than one approach (see Table 1 in Gonçalves et al. 2022[17]).

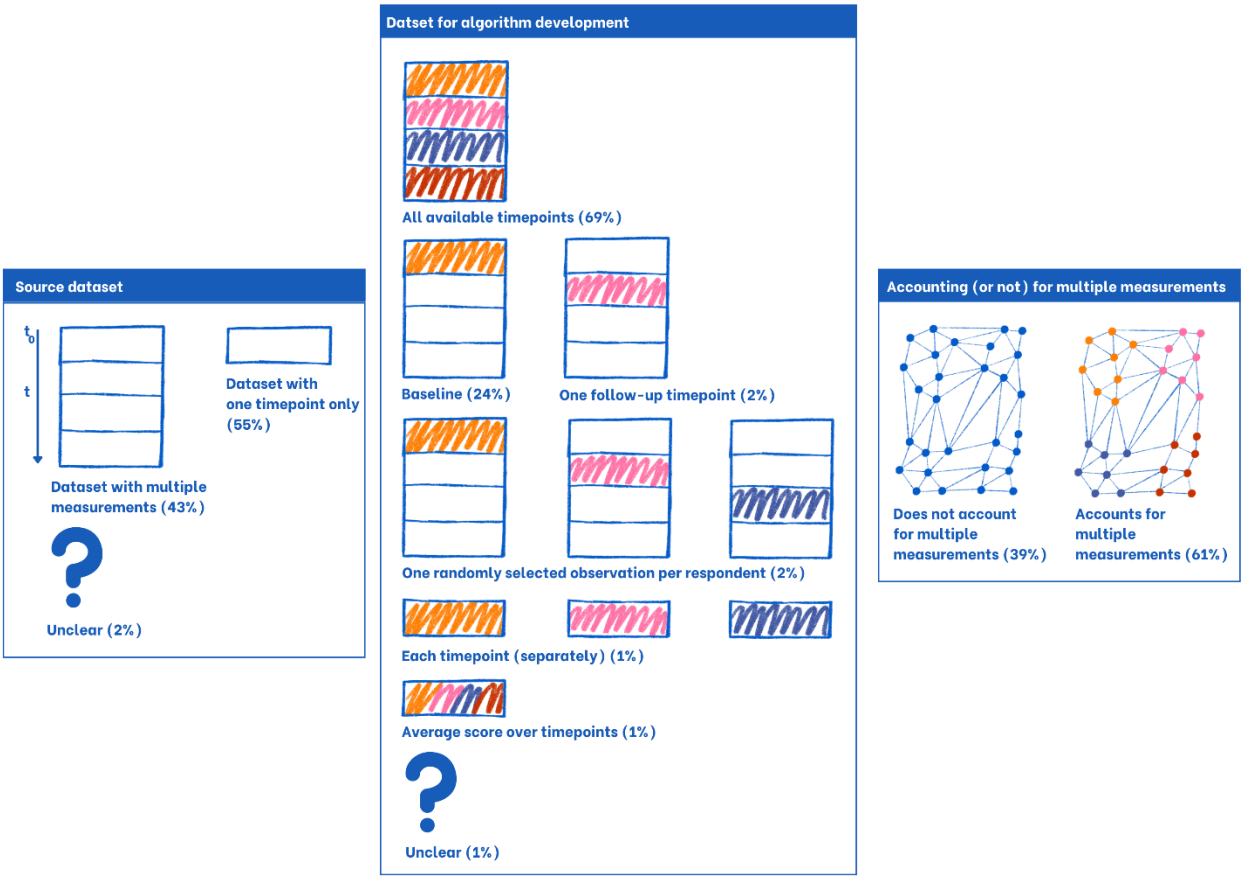


Figure 5 Overview of how mapping studies dealt with multiple observations per subject over time (own representation: Oliveira Gonçalves)

5.2.4 Methods for estimation

The earliest mapping research studies were published in the 1997-2005 period and nearly entirely depended on OLS to compute mapping algorithms. Authors defend employing OLS citing its prevalence and standard practice[29,63–105], user-friendliness[73,74,106–112], and the establishment of a benchmark against which more sophisticated models may be contrasted[113–121]. Today most studies also use more complex regression methods.

Among the 92 studies where it is evident that datasets with multiple observations were employed for estimating mapping algorithms, 36 did not take into consideration that these observations are typically more similar to one another than observations from distinct subjects, whereas 56 did. These 56 studies that accounted for multiple observations employed various techniques, with eight using more than one strategy. The strategies employed by researchers consisted of utilising random-effects or mixed-effects models (n = 30), cluster-robust standard errors (n = 21), generalised estimating equations (n = 7), pooled OLS (n = 1), and longitudinal fixed effects (n = 1). Several studies modified their data so that just one data point per subject was included in the estimation dataset. Coon et al. averaged their start questionnaire score over two time periods to

get a single observation per participant[122], whereas Madan et al. employed within-individual changes in both their start (Roland–Morris Disability Questionnaire - RMQ) and target questionnaire (EQ-5D-3L)[123]. Dixon et al.[113] and Hurst et al.[124] carried out regressions for each time point independently. Hoyle et al. acknowledged multiple observations as a challenge that must be addressed, but the particular methodological measures adopted were not explicitly specified[125].

Several studies that employed datasets comprising multiple observations from the same subjects over time described the reason behind not accounting for the correlation caused by multiple observations. Boland et al. discovered that utilising the overall dataset (i.e. with multiple observations per subject throughout time) increased model performance (as evaluated by the MAE and RMSE) when compared to a dataset containing only baseline measurements[116]. As a result, they used the complete (longitudinal) dataset while assuming observations from the same subjects throughout time as independent. Nair et al. primarily used multiple measurements from the same subjects over time[126]. However, they cross-checked their analysis by carrying out generalised estimating equations, which account for the multiple observations' characteristic of their data. They included an independent variable on time to evaluate if there was dependency between observations from the same individual. They concluded that time did not (statistically significantly) impact the EQ-5D index value, nor associations between HAQ/DAS28 (components). Thus, they decided to disregard generalised estimating equations and only relied on linear regressions. Versteegh et al. developed a mapping function from EORTC QLQ-C30 to EQ-5D-3L using a dataset with multiple observations[127]. Although the models did not account for multiple observations from the same subjects over time, the mapping function's predictive performance was evaluated separately for each time period.

Researchers in two studies stated that they were unable to adjust for multiple observations when using certain regression models in Stata's since the packages (e.g. ALDVMM, Betamix) did not include this option[16,128]. However, these software packages give the possibility to account for multiple observations with cluster-robust standard errors, which the aforementioned studies employed. There were also authors that used standard packages (ALDVMM) who did not think the cluster-robust estimator method was required since it does not alter the coefficients but just the standard errors employed in probabilistic sensitivity analyses. Thus, they chose not to adjust for multiple observations[129].

5.2.5 Mapping algorithm validation

Fourteen studies estimated their mapping functions with one time period from the original dataset (typically baseline data) and validated them with another (usually, follow-up data)[120,130–141].

Cheung et al. estimated the mapping algorithm using data from subjects who had just one measurement[142]. One measurement was chosen at random for algorithm estimation and another for algorithm validation for individuals who had measurements collected at two time points. Lee et al. computed the mapping function using baseline data from a dataset and subsequently evaluated its validity employing baseline and follow-up data[143]. Multiple studies regarded this approach to be external validation (e.g. Frew et al. 2015[137]), whereas it is actually split-sample validation.

5.2.6 Estimation and validation dataset splitting

Fourteen out 26 studies made use of a dataset with multiple observations per subject over time. Among the 14 studies that used a dataset with multiple observations, three only employed the dataset's baseline data for estimation, whereas the other 12 used all available time points. Both techniques were employed in one study (using both all observations and baseline data only).

5.2.7 Additional aspects

Only 26 of the 278 manuscripts included in this review provided the correlation matrix of their mapping functions.

We found 11 papers that determined that mapping should not be performed. This included both mapping functions created by the authors (e.g. Oliveira Gonçalves et al.[16]), and previously published algorithms. The target instrument in all but one study was a variant of the EQ-5D (i.e. the EQ-5D-3L or the EQ-5D-5L). Eight out of 11 manuscripts employed multiple observations from subjects over time as the data source for determining the mapping technique. All time points appear to have been utilised in all of them. Half of these studies took into account multiple observations.

5.3 Research Project 3 – Health economic evaluation

The MSU mobilisation group consisted of 749 patients, while the comparator group consisted of 794 patients. Baseline parameters and short-term outcomes are shown in Table 2. This table is similar to Table 1 in the manuscript by Ebinger et al.[46]. Patients included in this economic evaluation had a mean age of 73 years (SD: 13) in the MSU mobilisation group. In the usual care group, patients were 74 years old on average (SD: 13). The majority of patients were male: 53.8% in the MSU group and 52.5% in the standard care group. Prior to the stroke event, most patients lived at home without assistance (80.2% vs. 78.6%). Further information on the study design, participant selection, and full participants' characteristics have been reported elsewhere[46].

There was a small percentage of missing data in the dataset and there was complete information for all baseline variables.

Table 2 Baseline Parameters in Patients Included in the Analysis (modified from Gonçalves et al., 2023 [18]*)*

	Patients with MSU mobilisation (n = 749)	Patients without MSU mobilisation (n = 794)	Mean difference/odds ratio (95% CI)
Demographics			
Age, years, mean (SD)	73 (13)	74 (13)	−0.11 (−0.21; −0.01)[a]
Age, median [IQR]	75 [65-82]	77 [67-83]	-
Sex, female, n (%)	346 (46.2)	377 (47.5)	0.95 (0.78; 1.16)[b]
Sex, male, n (%)	403 (53.8)	417 (52.5)	1.05 (0.86; 1.29)[b]
Comorbidities			
Arterial hypertension, n (%)	589 (78.6)	649 (81.7)	0.82 (0.64; 1.06)[b]
Atrial fibrillation, n (%)	216 (28.8)	209 (26.3)	1.13 (0.91; 1.42)[b]
Diabetes mellitus, n (%)	191 (25.5)	201 (25.3)	1.01 (0.80; 1.27)[b]
Functional status pre-stroke			
Living at home without assistance, n (%)	601 (80.2)	624 (78.6)	1.11 (0.86; 1.42)[b]
Living in nursing institution, n (%)	84 (11.2)	96 (12.1)	0.92 (0.67; 1.25)[b]
Clinical information			
Documentation of presence or absence of neurological deficits at EMS arrival available	668 (89.2)	638 (80.4)	2.02 (1.51 to 2.69)[b]
First assessed National Institutes of Health Stroke Scale score, median (IQR)[c,d]	4 (2-9) (n = 746)	4 (2-9) (n = 789)	0.03 (−0.07; 0.13)[a]
National Institutes of Health Stroke Scale score assessed at hospital admission, median (IQR)[c,d]	3 (1-7) (n = 736)	4 (2-9) (n = 789)	−0.10 (−0.20; −0.003)[a]
Transient ischaemic attack, n (%)[e]	124 (16.6)	131 (16.5)	1.00 (0.77; 1.31)[b]
Ischaemic stroke, n (%)	625 (83.4)	663 (83.5)	1.00 (0.76; 1.30[b]
Large vessel occlusion documented in	163 (21.8)	177 (22.3)	0.97 (0.76; 1.23)[b]

acute vessel imaging, n (%)	

Emergency Medical Services (EMS); Interquartile Range (IQR); Standard deviation (SD)

[a] Mean difference of standardised z values (95% confidence interval)

[b] Unadjusted odds ratio (95% confidence interval)

[c] The National Institutes of Health Stroke Scale score ranges 0-42, where higher scores indicate higher neurological deficits

[d] Assessment in MSU in patients cared for by MSU or in emergency department in patients not cared for by MSU

[e] Transient ischaemic attack was classified as a transient neurological dysfunction triggered by brain loss of blood flow (following ICD 10: G45.x, except G45.4)

5.3.1 Costs

Table 1 in Gonçalves et al. 2023[18] shows the cost allocation per participant and exposure group considering a societal perspective and under the base-case scenario.

Considering a societal perspective, and under the base-case scenario, the largest cost contributors in the group who had an MSU mobilised were medical costs and expenses reimbursement, including costs for staff employed at the hospital and yielded EUR 4,383.08 per patient. Another major cost driver was related to MSU investment costs (amounting to EUR 1,286.02 per patient), with the personnel costs of the Berlin Fire Department (EUR 974.35 per patient) third. Hence, the total costs (not including long-term care costs) amounted to an average of EUR 8,491.58 during the 56-month study timeframe per patient in the MSU mobilisation group. The total costs in the standard care group amounted to EUR 1,274.54 per patient. Since at the moment of MSU mobilisation, the stroke subtype was unknown, we did not deduct costs related to non-eligible code stroke patients (e.g. for intracranial haemorrhage stroke patients or stroke mimics). However, we deducted costs related to mobilisations of MSU to non-code stroke emergency calls (ultimately, 11.6% of all MSU mobilisations were in response to non-stroke emergencies), meaning that under the base-case scenario, we only considered 88.4% of MSU-related costs. Supplementary Appendix 3, Table S3.3 in Gonçalves et al. 2023 lists complete costs considering all patients for whom MSU was mobilised (including non-code stroke patients) as well as costs for code stroke alarms only[18].

Under the base-case scenario, from the statutory health insurance perspective (excluding long-term care costs), the mean costs per patient amounted to EUR 6,246.70 in the MSU mobilisation group and EUR 1,274.55 in the standard care group.

Details on the resource use and the unit costs assigned to the resource use items to produce the reported cost figures are shown in Supplementary Appendix 3, Table S3.3 in Gonçalves et al. 2023[18].

5.3.2 Outcomes: descriptive overview

The MSU group had a higher average EQ-5D score (0.63 versus 0.59) than the group receiving conventional treatment. Additionally, more patients in the MSU mobilisation group than in the standard care group had a ‘good outcome’ as determined by the mRS score (50.92% versus 42.31%) (see Table 2 in Gonçalves et al. 2023[18]).

5.3.3 Cost-utility analyses

The base-case scenario results, considering a societal perspective are shown in Table 3 in Thesis Gonçalves et al. 2023[18]. Under this scenario, MSU mobilisation was linked to higher costs and higher QALYs. The incremental total costs as a result of MSU mobilisation yielded EUR 10,759,089.49 (9,912,284.42; 11,997,571.30). The incremental QALYs as a result of MSU mobilisation amounted to 262.52 (−41.06; 479.92). Thus, the ICER per QALY yielded EUR 40,983.82. It was found that a large proportion (95.16%) of the bootstrapped iterations, pointed to MSU mobilisation being linked to both higher QALYs and higher costs (Figure 1 in Gonçalves et al. 2023[18]).

Taking the statutory health insurance perspective, the incremental overall costs as a result of MSU mobilisation yielded EUR 7,406,034.65 (6,396,442.45; 8,539,734.95), and the incremental overall QALYs 264.22 (−40.98; 484.59). Hence, the ICER per QALY was EUR 28,029.51, which is lower than the ICER estimated when considering a societal perspective (see Supplementary Appendix 3 in Gonçalves et al. 2023[18]).

As we had anticipated, given the assumptions we made for each scenario (see Supplementary Appendix 1 in Gonçalves et al. 2023[18]), the lowest ICER was found under the best-case scenario (EUR 24,470.76 per QALY considering a societal perspective and EUR 14,843.78 considering a statutory health insurance perspective), mostly thanks to the lower incremental overall costs in comparison to the alternative scenarios. On the other side, the highest ICER value was found under the worst-case scenario, for both the societal (EUR 61,690.88 per QALY), as well as the statutory health insurance perspectives (EUR 41,539.98 per QALY). These figures were higher than in the base-case scenario, which was attributable to higher incremental overall costs and to lower incremental overall QALYs.

The bootstrap iterations' trends were analogous across all scenarios and perspectives, with circa 95% of the bootstrapped samples falling into the planes' northeast quadrant — thus pointing to higher costs and higher QALYs. The remaining ~5% of the samples fell into the northwest quadrant, meaning that the MSU intervention was dominated by standard care — with bootstrap iterations being linked to higher costs and lower QALYs.

5.3.4 Cost-effectiveness analyses

Under the base-case scenario, considering the societal perspective, the cost-effectiveness analyses revealed that incremental overall costs resulting from MSU mobilisation yielded EUR 10,793,823.78 (9,809,757.36; 12,020,619.78) and incremental survivals 132.45 (48.30; 199.85), resulting in an incremental survival without symptoms/disability of EUR 81,491.49 (see Table 3 in Gonçalves et al. 2023[18]).

The proportion of bootstrapped samples that fell into the northeast quadrant of the cost-effectiveness plane reached almost 100% and was higher than in the cost-utility analyses (see Figure 2 in Gonçalves et al. 2023[18]).

The ICER under the best-case scenario amounted to EUR 44,455.30 per survival without symptoms/disability.

The worst-case scenario yielded an ICER of EUR 116,491.15 per survival without symptoms/disability.

Under the base-case scenario, considering the statutory health insurance perspective, the incremental overall costs attributable to MSU mobilisation amounted to EUR 7,312,193.98 (6,277,094.81; 8,445,685.39) and the incremental number of survivals without symptoms/disability to 140.73 (61.09; 213.23). Accordingly, the ICER per incremental survival without symptoms/disability was EUR 51,959.46, which is a lower figure than the ICER computed when considering the societal perspective (see Supplementary Appendix 3 in Gonçalves et al. 2023[18]). All bootstrap replications fell into the northeast quadrant of the cost-effectiveness plane. This result indicates that MSU mobilisation yields both higher incremental costs and higher incremental survivals without symptoms/disability.

5.3.5 Cost-benefit analyses

The cost-benefit analysis yielded a net monetary benefit of EUR 68,364.82 per survival without symptoms/disability, which corresponds to the incremental overall costs attributable to the MSU mobilisation subtracting the incremental overall saved care costs (see Table S3.5 in Supplementary Appendix 3 in Gonçalves et al. 2023[18]).

5.3.6 Sensitivity analyses

The sensitivity analyses produced results that were consistent with those of the main analyses (see Supplementary Appendix 3 in Gonçalves et al. 2023[18]).

6. Discussion

The aim of this thesis was to explore different challenges and applications of PROMs, which were addressed by the following contributions:

- Development of a mapping algorithm from a non-preference-based measure to a preference-based measure in Thesis Article 1 (Oliveira Goncalves et al. 2021);
- Review of challenges related to the development of mapping algorithms when multiple measurements from the same subjects over time are present in Thesis Article 2 (Gonçalves et al. 2022);
- A health economic evaluation where both a PROM and a clinical outcome were assessed as outcomes in Thesis Article 3 (Gonçalves et al. 2023).

6.1 Research Project 1 – Mapping algorithm development

We intended to determine if there was a conceptual overlap between the EQ-5D and the HIT-6, and we aimed to provide a mapping function for estimating the EQ-5D score (using the value set for Germany) from the HIT-6 survey, which is a disease-specific questionnaire commonly employed in migraine studies. Our findings indicate noteworthy differences in the underlying components of the HIT-6 and the EQ-5D, as explored using exploratory factor analysis. Furthermore, the EQ-5D displayed a major ceiling effect and small SRMs across time, while the HIT-6 did not display a ceiling effect and showed a higher level of responsiveness. This study also suggested a mapping algorithm for function HIT-6 scores to EQ-5D utility values.

As both instruments have been validated in patients with migraine, we hypothesised that there would be a substantial overlap between them. However, we found that the strength of association between the HIT-6 and the EQ-5D — measured with correlation coefficients, was just low to moderate. This was true for both overall scores and for the individual questions on each instrument. The factor analytical results also pointed to a lack of conceptual overlap between these instruments, hinting that they possibly measure different underlying constructs. The absence of overlap can have different causes. First, the questions included in the two instruments have different recall periods. The HIT-6 contains three questions that refer to the previous four weeks. On the other hand, all EQ-5D questions are related to the day when the participant fills out the survey. Second, the HIT-6 contains frequency answer options (spanning from ‘never’ to ‘always’), whereas the EQ-5D has severity answer options (spanning from ‘no problems’ to ‘unable to/extreme problems’). Third, the EQ-5D describes migraine patients’ utilities during the moment when the questionnaire is being administered. Therefore, it does not distinguish if patients had an attack during the survey administration[144].

The ceiling effects we found in our study can be a consequence of the poor discrimination of the EQ-5D-5L for patients who suffer from migraine. To the best of our knowledge, only two studies have validated the administration of the HIT-6 to German patients suffering from chronic migraine. Rendas-Baum et al. carried out a study which included, among other nationalities, German patients, but the authors were unable to conduct country-specific analyses due to the small sample size of the four European nations covered[145]. Thus, no analyses specific for German patients were conducted. A study conducted by Martin et al. analysed if the United States version of the HIT-6 is analogous to the German version[146]. Unfortunately, it is unknown if the patients who were recruited had episodic or chronic migraines. Furthermore, the manuscript from Martin et al. does not report whether the enrolled individuals suffered from episodic or chronic migraine. Therefore, we believe that future research regarding the validation of the HIT-6 in German

patients with episodic and chronic migraine would be helpful to explore and potentially enlighten the reasons behind the missing conceptual overlap between this instrument and the EQ-5D-5L.

Considering the EQ-5D's low level of responsiveness, the considerable ceiling effect for migraine patients, and the lack of conceptual overlap, economic evaluations focused on these patients should investigate different ways to measure HRQoL and should not rely exclusively on QALYs collected from generic utility-based instruments. In this context, the International Headache Society describes in its guidelines that as utility instruments can be insensitive, QALYs might not take some patient preferences into consideration[147]. Hence, even if utility scores were gathered in the study and no mapping function was required, the use of QALYs may still be unsuitable. Thus, it may be more appropriate to use clinical effectiveness endpoints (e.g. monthly migraine days) when carrying out economic evaluations in migraine field. Such disease-specific outcomes, on the other hand, present a challenge to decisionmakers when comparing allocation of resources across different illnesses.

Our study has multiple strengths. First, our study participants were diagnosed by trained migraine neurologists. A further strength is the low percentage of missing data in our dataset. Moreover, we had responses from participants to the questionnaires during multiple periods, and we accounted for this by using methods appropriate for multiple observations' data.

Nevertheless, our study has several limitations. First, it was not possible to carry out an external validation of our mapping function. Second, we generated a mapping algorithm using data stemming from a RCT. Although these trials are seen as the 'gold standard' for evidence-based medicine[148], they perform less well in terms of generalisability to other settings than study designs. Indeed, the guidelines from ISPOR on mapping state that RCTs often include a less diverse pool of participants than observational studies, thanks to both their inclusion and exclusion criteria, and their reduced follow-up[11].

As a result, we contrasted a number of socio-demographic features of our study sample with those of patients with migraine who were involved in three studies mentioned in a research project conducted by the German Migraine and Headache Society (the Dortmund Health study, the KORA Augsburg study, and SHIP Pomerania study[149], see Supplementary Table A.7 in Oliveira Goncalves et al. 2021[16]). The average age recorded for episodic migraine amounted to 47.5 in the Dortmund Health study, 50.0 in the KORA Augsburg study, and 50.1 in the SHIP study. The average age for episodic migraine participants (excluding headaches caused by medication, which was an exclusion criterion in our analysis) was 60.8 in the KORA Augsburg study and 61.0 in the SHIP study. No values were reported for the Dortmund Health Study. The mean age in our study was somewhat lower: 40.1 for chronic migraine and 41.5 for episodic migraine. The lower mean age in our study could be attributed to the fact that participants had to

be comfortable using apps, and that Berlin is the German federal state with the second lowest mean age[150]. Regarding the distribution of sex, the proportion of women in the Dortmund Health study was 78.7%, 84.2% in the KORA study, and 85.6% in the SHIP study. It is important to emphasise that many people suffering from migraine choose not to seek medical assistance, which means that the literature in general may not have properly described their characteristics. In fact, only around two-thirds of individuals who suffer from migraine in Germany seek treatment with a physician[151]. It was not possible to conduct response mapping with our dataset, as response mapping demands a high number of observations in each response level, and our dataset included few responses at the worst response categories[32]. We carried out our EFA without taking into consideration multiple observations; however, we obtained the same findings in the sensitivity analysis using only baseline data in what concerns the number of factors and the meaningful loadings. We developed our mapping algorithms with mixed-effects models with random intercepts, which meant different intercepts for each cluster. Thus, we assumed that the link between the explanatory and dependent variables is analogous across different clusters. In the ALDVMMs, it was not possible to include random-effects.

6.2 Research Project 2 – Systematic review of mapping algorithms

We discovered that authors often use datasets with multiple measurements per individual over time to develop their mapping algorithms. However, our systematic review suggests that many authors do not account for interdependence of data or do not clearly describe the methods used. This analysis found 278 papers that developed mapping functions, adding 69 to the sample described in Mukuria et al.'s previous systematic review of mapping functions[43].

When typical statistical techniques (e.g. OLS without including cluster-robust standard errors) are employed to analyse hierarchical data, such as data with multiple observations per subject over time, the assumption of independent errors is compromised. When intraindividual correlations in datasets with multiple observations per subject over time are ignored, standard errors are underestimated[152].

Although many publications employed datasets with multiple observations, we discovered that not all of them used all observations for developing a mapping algorithm. We found that several authors decided to divide the datasets per time point and use one part for estimation purposes and the remaining part for validation. Such a strategy might be problematic, as it not only leads to a smaller sample size for computing mapping functions, but can also have more serious implications, notably when utilising response modelling to estimate mapping functions. Therefore, the estimation dataset needs to include observations for each response level of the several dimensions that the target questionnaire measures[32]. If researchers divide their estimation and validation sets according to the time period at which the data was gathered, the distribution of severity levels may be substantially different. For instance, if data comes from a given study where quality of life is impacted by an intervention, the answers at baseline could reveal greater severity levels and, as a result, there will be less answers in the lower severity levels in that time point. On the other hand, if the algorithm validation is performed using follow-up data only, this dataset may not contain numerous observations in the poorest response group of the various domains. On this subject, Davison et al. stated that because they employed a longitudinal dataset in which some individuals got interventions that lowered the severity of the illness under consideration while others did not, the mapping function they have developed ought to be valid for all degrees of severity of the illness under consideration[128].

Hernandez Alava et al., also highlight the importance of using longitudinal data to examine if two instruments are monotonic[153]. According to these authors, if a health state as measured by instrument X unmistakably improved from one point in time to another, then we would assume that instrument Y would also present the same improvement.

Models should undergo sensitivity analyses, where model input parameters are altered, since models are constructed using assumptions about input values. In the case of mapping algorithms used in cost-effectiveness models, guidelines advocates taking into account the uncertainty concerning the predicted PROMs' values[153,154]. One way to address this issue is by parametrising the uncertainty in the PROMs values using variance-covariance matrices obtained from the estimated regression coefficients. Nevertheless, if the standard errors were not properly calculated (for example, by disregarding intraindividual correlations among measurements from the same subjects), the variance-covariance matrices will be incorrect.

It is also worth mentioning that frequently used Stata mapping packages, such as ALDVMM and Betamix, do not offer the capability to account for multiple observations from the same subjects employing random-effects. Even though there is a cluster-robust standard errors option, this may cause some researchers to forgo taking multiple observations into account.

One drawback of this systematic review is that we did not contact the authors of the included studies to confirm findings or get further information where information gaps persisted. As a result, we may not have accurately categorised all extraction fields.

6.3 Research Project 3 – Health economic evaluation

The incremental costs and the incremental QALYs for the B_PROUD primary population resulted in an ICER of approximately EUR 41,000 per QALY. Our findings are consistent with a previous study carried out in Germany between 2011 and 2013, which found an incremental cost-effectiveness ratio amounting to EUR 32,456 per QALY[52]. Overall, our results align with the hypothesis that the mobilisation of MSUs may be deemed cost-effective when considering thresholds larger than EUR 40,000 per QALY. The meaning of this figure relies on the jurisdiction-specific thresholds. The range of potential thresholds varies, for example, EUR 20,000-EUR 80,000 in the Netherlands or USD 100,000-USD 150,000 in the US[155,156]. Although Germany lacks a recognised ICER per QALY threshold, the MSU mobilisation employment could be supported in the Netherlands, which has historically had a very analogous healthcare system (defined by a combination of mandatory social health insurance and private voluntary health insurance) where the ICER estimate could fall within the official threshold bounds.

The generalisability of economic evaluations and how their results can be extrapolated to other settings based on observational studies poses a challenge. We aimed to increase the generalisability of our study by following the reporting recommendations from Drummond et al[157]. On what concerns costs, if feasible, we provided cost data that clearly showed unit expenses as well as the levels of resource utilisation. This would enable policymakers in various health-care systems to attach their particular pricing to the identified resource consumption units. Moreover, we have provided results based on a societal and on a statutory-health insurance perspective. The former is deemed to be less country dependent than the statutory-health insurance perspective. The sites included in this study are representative of Berlin (Germany), as all the hospitals that have a Stroke Unit in the city were included in the study. Concerning the catchment areas, the three MSUs' operation areas cover approximately 94% of the inhabitants of Berlin (according to calculations from the Berlin Fire Department). The study includes a high proportion of the stroke caseload in Berlin since the vast majority of the Berlin stroke patients are transported to one of the mentioned sites because the Berlin Emergency Service legislation requires that all stroke suspects have to be brought to a hospital with a Stroke Unit. The health state valuations were assessed with commonly used instruments in the field of stroke. In what concerns the EQ-5D-3L, we have used appropriate utility weights for the German population. The instrument mRS is appropriate to the population under study (patients with transient ischaemic attack or stroke), and since it is not a preference-based patient-reported outcome, it has no associated weights. We have also provided details on the degree of incomplete observations, and we carried out analyses with and without imputing missing observations.

We had to adapt our calculation method for the cost-benefit analysis for Germany. Normally, the summary measure for cost-benefit analysis is the net monetary benefit, which is calculated by multiplying the incremental benefit of the intervention with a willingness to pay threshold for a unit of benefit minus the incremental cost of the intervention[158]. As previously mentioned, in the case of Germany, there is no official willingness to pay threshold for units of benefit. Thus, we used a different approach, as explained above.

Due to constraints on the availability of data, we had to make simplifying assumptions that should be considered when interpreting our results.

First, it was not possible to take into consideration potential savings occurring in-hospital resulting from MSU implementation, such as those related to shorter stays in expensive high-care facilities of acute care and rehabilitation establishments because of enhanced functional outcomes (see Supplementary Appendix 3 in Gonçalves et al. 2023[18]).

Second, despite the patients' average age of 73.5 years, 385 of them were under the age of 65, which is Germany's current retirement age. Our calculations are therefore conservative since they do not consider possible savings in what concerns gains of productivity when considering a societal perspective.

Third, in the base-case scenario, we employed the most prudent strategy to translating expenditures from the old ‘level-of-care’ to the modern ‘grade-of-care’ (see Supplementary Appendix 1 in Gonçalves et al. 2023[18]). Mental and communicative deficits result in higher care benefits in the present ‘grade-of-care’ system, but not in the old ‘level-of-care’ system used for the conversion in this scenario, hence a significant number of patients would have obtained higher care benefits. Deficits in cognition and communication translate into greater care benefits in the present ‘grade-of-care’ scheme, as opposed to the former ‘level-of-care’ scheme utilised for the translation in this scenario. Given that MSU patients showed lower levels of incapacity, taking a less conservative approach would have led to higher expenses in the standard care group. Fourth, in this scenario, we did also not take into consideration potential savings with transportation, which are expected to occur given that stroke work-up on the MSUs’ mobilisation improves delivery of patients to the most suitable hospital[159]. Thus, we did not include costs related to the secondary transport of patients for certain procedures such as thrombectomy and the mobilisation of other costly physician-staffed ambulances, which is standard practice in Germany. Fifth, we did not consider the time that pre-hospital stroke work-up and medical care saved in emergency rooms.

Sixth, as a result of the study’s inclusion criteria, we only took into account effects seen in this primary population of patients who had no absolute contraindications to reanalysing therapies, which excludes effects in remaining populations for whom a MSU was mobilised, such as patients who have experienced intracranial haemorrhage or patients who have received definite

diagnoses of non-stroke conditions. However, we have taken into consideration the expenses that these patients incurred.

Finally, the EQ-5D-3L was only administered three months after the index event. As a result, our study's five-year time frame was extrapolated using information other than what we collected, assuming that the quality of life would remain constant and there would be no new deaths during that period. However, what we were seeking to compare was the average difference in QALYs between the two groups after adjustment for confounding. Hence, in order for the estimate to be valid, it was enough that potential violations of this assumption cause a shift of the same amount of QALY in the MSU and non-MSU groups across the five-year period. According to Luengo-Fernandez et al., survival rates and quality of life (as determined by the EQ-5D-3L) across different categories of stroke severity were quite stable over the course of the five years following the stroke event[47]. Furthermore, Luengo-Fernandez et al. did not collect EQ-5D-3L values at month three like in our study. Since data for three months was missing, we thus hypothesised that utilities obtained at one month were comparable to those at three months. This decision was conservative since, up until three months after a stroke, physical rehabilitation typically improves considerably and promptly[160].

We provide evidence that the MSU mobilisation to acute ischaemic stroke individuals without contraindications to recanalising treatments is cost-effective compared with non-MSU pre-hospital structural settings, taking into consideration thresholds greater than an incremental EUR 40,000 per QALY. Our findings may serve as guidance for policymakers who are considering pre-hospital stroke care in urban settings in the future.

A systematic review method in economic evaluations in the field of stroke, is underway as an already accepted follow-up project funded by the Center for Stroke Research Berlin at the Charité – Universitätsmedizin Berlin. This systematic review was registered in OSF[161] and is currently being conducted.

7. Conclusion

This PhD dissertation showcases different methods within health data sciences and healthcare services research, including the development of a mapping algorithm, a systematic review and an economic evaluation. Briefly, I showed that there is little conceptual overlap between a condition-specific PROM (HIT-6) and a preference-based measure (EQ-5D-5L). Thus, mapping may not always be an appropriate approach for obtaining utilities. I demonstrated the importance of taking into account multiple observations per subject over time when developing and validating mapping algorithms. Finally, I presented the results of an economic evaluation which used an outcome measure scored by clinicians based on an assessment of the patient's level of disability, as well as a preference-based measure.

Furthermore, this thesis not only applied different methods, but also focused on different (neurological) conditions, thus having different clinical and economic implications. In terms of clinical considerations, clinicians working in the field of migraine should administer both preference-based measures and condition-specific PROMs. Preference-based measures can be later used to conduct cost-utility analyses, without the need to resort to mapping, which is only the second-best approach. Condition-specific PROMs, such as the HIT-6 should also be administered, as they typically capture the impact of a particular health condition on an individual's HRQoL better. In acute conditions such as stroke, treatments have a substantial impact after they are first administered. Thus, researchers need to capture the effects of HRQoL relatively regularly, especially if HRQoL may improve or decrease very quickly. In the case of stroke, PROMs should be administered one month after the event and again three months later, but once the condition stabilises, HRQoL can be measured less frequently, e.g. at six months, one year, and five years. If the HRQoL is not measured shortly enough after the intervention, researchers miss the highest improvement. On the other hand, if the HRQoL is only measured shortly after intervention but then only measured again after a long time has passed (e.g. after two years), researchers might not obtain a good picture of what happened given the high attrition within these two years.

In the last few years, the importance of PROMs has been recognised by researchers, clinicians, patients, and decisionmakers; and their importance will increase as we move towards a more patient-centred approach to care. All these groups can benefit from the correct use of PROMs. Patients can express their views on the healthcare process. Policymakers can use PROMs to assist them in evidence-based decision-making, by including them as outcomes in health economic evaluations. Furthermore, PROMs can be used as performance measurements. Finally, healthcare providers can better understand how to provide care that targets what affects patients' perceived HRQoL. Hence, researchers should carefully consider the strengths and limitations when selecting outcome measures in trials and choose the most appropriate tool for

their specific research question and jurisdiction. This will ensure that valuable data can be collected and analysed, which in turn can be used to allocate resources appropriately, ultimately benefiting the entire population.

Multiple ongoing projects are reinforcing the current interest in PROMs at high decision-making levels, such as the Patient-Reported Indicator Surveys (PaRIS) initiative. The PaRIS initiative was launched by the Organisation for Economic Co-operation and Development (OECD) and aims to promote the global use of patient-reported indicators in a way that facilitates international comparisons, collaborative learning and research[162].

Finally, it is crucial to emphasise that the utilisation of PROMs alone does not guarantee that patients will be able empowered to actively participate in decisions regarding their healthcare. Even though these tools can provide valuable information, they may not inevitably lead to the meaningful involvement of patients in the healthcare decision-making process.

8. References

1. World Health Organization. Constitution. World Health Organization; 1989.
2. Papanicolas I, Rajan D, Karanikolos M, Soucat A, Figueras J. Health system performance assessment: a framework for policy analysis. World Health Organization; 2022.
3. Eriksen J, Bygholm A, Bertelsen P. The Purpose of Patient-Reported Outcome (PRO) Post Its Digitalization and Integration into Clinical Practice: An Interdisciplinary Redefinition Resembling PROs Theoretical and Practical Evolvement. Appl Sci. 2020 Oct 26;10(21):7507.
4. Black N, Jenkinson C. Measuring patients' experiences and outcomes. BMJ. 2009 Jul 2;339:b2495.
5. Wu AW, Snyder C. Getting ready for patient-reported outcomes measures (PROMs) in clinical practice. Healthc Pap. 2011;11(4):48–53; discussion 55-8.
6. Weldring T, Smith SMS. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). Health Serv Insights. 2013 Aug 4;6:61–8.
7. Bottomley A, Jones D, Claassens L. Patient-reported outcomes: assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency. Eur J Cancer. 2009 Feb;45(3):347–53.
8. Leong KP, Yeak SC, Saurajen AS, Mok PK, Earnest A, Siow JK, Chee NW, Yeo SB, Khoo ML, Lee JC, Seshadri R, Chan SP, Tang CY, Chng HH. Why generic and disease-specific quality-of-life instruments should be used together for the evaluation of patients with persistent allergic rhinitis. Clin Exp Allergy. 2005;35(3):288–98.
9. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013 Jan 28;346:f167.
10. Lamu AN, Gamst-Klaussen T, Olsen JA. Preference Weighting of Health State Values: What Difference Does It Make, and Why? Value Health. 2017 Mar;20(3):451–7.
11. Wailoo AJ, Hernandez-Alava M, Manca A, Mejia A, Ray J, Crawford B, Botteman M, Busschbach J. Mapping to Estimate Health-State Utility from Non-Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. Value Health. 2017 Jan;20(1):18–27.
12. Health care expenditure by financing scheme [Internet]. Eurostat. [cited 2023 Feb 21]. Available from: https://ec.europa.eu/eurostat/databrowser/view/HLTH_SHA11_HF__custom_5013685/default/table?lang=en

13. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. Int J Stroke. 2009 Jun;4(3):200–5.

14. Dromerick AW, Edwards DF, Diringer MN. Sensitivity to changes in disability after stroke: A comparison of four scales useful in clinical trials. J Rehabil Res Dev. 2003;40(1–4):1–8.

15. Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale: a systematic review. Stroke. 2009 Oct;40(10):3393–5.

16. Oliveira Goncalves AS, Panteli D, Neeb L, Kurth T, Aigner A. HIT-6 and EQ-5D-5L in patients with migraine: assessment of common latent constructs and development of a mapping algorithm. Eur J Health Econ. 2021;23(1):47–57.

17. Gonçalves ASO, Werdin S, Kurth T, Panteli D. Mapping Studies to Estimate Health-State Utilities From Non–Preference-Based Outcome Measures: A Systematic Review on How Repeated Measurements are Taken Into Account. Value Health. 2022;26(4):589–97.

18. Gonçalves ASO, Rohmann JL, Piccininni M, Kurth T, Ebinger M, Endres M, Freitag E, Harmel P Dr med, Lorenz-Meyer I, Rohrpasser-Napierkowski I Dr rer nat, Busse R, Audebert HJ. Economic Evaluation of a Mobile Stroke Unit Service in Germany. Ann Neurol. 2023 Jan 13;93(5):942–51.

19. Gonçalves ASO, Laumeier I, Hofacker MD, Raffaelli B, Burow P, Dahlem MA, Heintz S, Jürgens TP, Naegel S, Rimmele F, Scholler S, Kurth T, Reuter U, Neeb L. Study Design and Protocol of a Randomized Controlled Trial of the Efficacy of a Smartphone-Based Therapy of Migraine (SMARTGEM). Front Neurol. 2022 Jun 16;13:912288.

20. van Reenen M, Janssen B. EQ-5D-5L user guide: basic information on how to use the EQ-5D-5L instrument. Rotterdam: EuroQol Research Foundation. 2015;9.

21. Ludwig K, Graf von der Schulenburg JM, Greiner W. German Value Set for the EQ-5D-5L. Pharmacoeconomics. 2018 Jun;36(6):663–74.

22. Yang M, Rendas-Baum R, Varon SF, Kosinski M. Validation of the Headache Impact Test (HIT-6^{TM}) across episodic and chronic migraine. Cephalalgia. 2011 Feb;31(3):357–67.

23. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Academic Press; 2013. 490 p.

24. Tabachnick BG, Fidell LS. Using Multivariate Statistics, Allyn and Bacon, Boston, MA. Using Multivariate Statistics, 4th ed Allyn and Bacon, Boston, MA. 2001;

25. Holgado-Tello F, Moscoso S, Barbero–García I, Vila E. Polychoric versus Pearson correlations in Exploratory and Confirmatory Factor Analysis with ordinal variables. Quality and Quantity. 2010;44:153–66.

26. Yang Y, Xia Y. On the number of factors to retain in exploratory factor analysis for ordered categorical data. Behav Res Methods. 2015 Sep 1;47(3):756–72.

27. Glorfeld LW. An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. Educ Psychol Meas. 1995 Jun 1;55(3):377–93.

28. Li CH. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behav Res Methods. 2016 Sep 1;48(3):936–49.

29. Young TA, Mukuria C, Rowen D, Brazier JE, Longworth L. Mapping Functions in Health-Related Quality of Life: Mapping from Two Cancer-Specific Health-Related Quality-of-Life Instruments to EQ-5D-3L. Med Decis Making. 2015 Oct;35(7):912–26.

30. Alava MH, Wailoo A. Fitting adjusted limited dependent variable mixture models to EQ-5D. Stata J. 2015;15(3):737–50.

31. Gray LA, Alava MH. A command for fitting mixture regression models for bounded dependent variables using the beta distribution. Stata J. 2018;18(1):51–75.

32. Gray LA, Hernandez Alava M, Wailoo AJ. Development of Methods for the Mapping of Utilities Using Mixture Models: Mapping the AQLQ-S to the EQ-5D-5L and the HUI3 in Patients with Asthma. Value Health. 2018;21(6):748–57.

33. Hernandez Alava M, Wailoo A, Wolfe F, Michaud K. A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. Med Decis Making. 2014;34(7):919–30.

34. Le QA, Doctor JN. Probabilistic Mapping of Descriptive Health Status Responses Onto Health State Utilities Using Bayesian Networks: An Empirical Analysis Converting SF-12 Into EQ-5D Utility Index in a National US Sample. Med Care. 2011;49(5):451–60.

35. Kelman L. Migraine changes with age: IMPACT on migraine classification. Headache. 2006;46(7):1161–71.

36. Peterlin BL, Gupta S, Ward TN, Macgregor A. Sex matters: evaluating sex and gender in migraine and headache research. Headache. 2011;51(6):839–42.

37. Pistoia F, Sacco S. Migraine and Use of Combined Hormonal Contraception. In: Maassen van den Brink A, MacGregor EA, editors. Gender and Migraine. Cham: Springer International Publishing; 2019. p. 69–79.

38. Andreou AP, Edvinsson L. Mechanisms of migraine as a chronic evolutive condition. J Headache Pain. 2019 Dec 23;20(1):117.

39. Gluzmann P, Panigo D. GSREG: Stata module to perform Global Search Regression. 2013 Nov 24 [cited 2020 Aug 3]; Available from: https://EconPapers.repec.org/RePEc:boc:bocode:s457737

40. Kuhn M, Johnson K. Over-Fitting and Model Tuning. In: Applied Predictive Modeling. New York, NY: Springer New York; 2013. p. 61–92.

41. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput Vienna, Austria. 2017;

42. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2010 Apr 1;11(2):215–25.

43. Mukuria C, Rowen D, Harnan S, Rawdin A, Wong R, Ara R, Brazier J. An Updated Systematic Review of Studies Mapping (or Cross-Walking) Measures of Health-Related Quality of Life to Generic Preference-Based Measures to Generate Utility Values. Appl Health Econ Health Policy. 2019;17(3):295–313.

44. Petrou S, Rivero-Arias O, Dakin H, Longworth L, Oppe M, Froud R, Gray A. The MAPS Reporting Statement for Studies Mapping onto Generic Preference-Based Outcome Measures: Explanation and Elaboration. Pharmacoeconomics. 2015;33(10):993–1011.

45. Ebinger M, Siegerink B, Kunz A, Wendt M, Weber JE, Schwabauer E, Geisler F, Freitag E, Lange J, Behrens J, Erdur H, Ganeshan R, Liman T, Scheitz JF, Schlemm L, Harmel P, Zieschang K, Lorenz-Meyer I, Napierkowski I, Waldschmidt C, Nolte CH, Grittner U, Wiener E, Bohner G, Nabavi DG, Schmehl I, Ekkernkamp A, Jungehulsing GJ, Mackert BM, Hartmann A, Rohmann JL, Endres M, Audebert HJ. Association Between Dispatch of Mobile Stroke Units and Functional Outcomes Among Patients With Acute Ischemic Stroke in Berlin. JAMA. 2021 Feb 2;325(5):454–66.

46. Claes C, Greiner W, Uber A, Graf von der Schulenburg JM. An interview-based comparison of the TTO and VAS values given to EuroQol states of health by the general German population. In: Proceedings of the 15th Plenary Meeting of the EuroQol Group Hannover, Germany: Centre for Health Economics and Health Systems Research, University of Hannover. 1999. p. 13–38.

47. Luengo-Fernandez R, Gray AM, Bull L, Welch S, Cuthbertson F, Rothwell PM, Oxford Vascular Study. Quality of life after TIA and stroke: ten-year results of the Oxford Vascular Study. Neurology. 2013 Oct 29;81(18):1588–95.

48. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. Stroke. 2007 Mar;38(3):1091–6.

49. Audebert HJ, Schultes K, Tietz V, Heuschmann PU, Bogdahn U, Haberl RL, Schenkel J, Telemedical Project for Integrative Stroke Care (TEMPiS). Long-term effects of specialized stroke care with telemedicine support in community hospitals on behalf of the Telemedical Project for Integrative Stroke Care (TEMPiS). Stroke. 2009 Mar;40(3):902–8.

50. Verband der Ersatzkassen. Pflegegrade [Internet]. [cited 2021 Mar 27]. Available from: https://www.vdek.com/presse/glossar_gesundheitswesen/pflegestufen.html

51. IQWIG, Institute for Quality and Efficiency in Health Care. General Methods Version 6.0. 2020.

52. Gyrd-Hansen D, Olsen KR, Bollweg K, Kronborg C, Ebinger M, Audebert HJ. Cost-effectiveness estimate of prehospital thrombolysis: results of the PHANTOM-S study. Neurology. 2015 Mar 17;84(11):1090–7.

53. IST-3 collaborative group. Effect of thrombolysis with alteplase within 6 h of acute ischaemic stroke on long-term outcomes (the third International Stroke Trial [IST-3]): 18-month follow-up of a randomised controlled trial. Lancet Neurol. 2013 Aug;12(8):768–76.

54. Gonçalves ASO, Rohmann JL. B_PROUD Economic Evaluation: Statstical Analysis Plan. OSF. 2022;

55. Hernán MA, Robins JM. Causal inference: what if. Boca Raton: Chapman & Hall/CRC; 2020.

56. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. Pharmacoeconomics. 2014 Dec;32(12):1157–70.

57. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018 Jun 30;37(14):2252–66.

58. York Health Economics Consortium. Cost-Effectiveness Plane [Internet]. 2016 [cited 2022 Apr 1]. Available from: https://yhec.co.uk/glossary/cost-effectiveness-plane/

59. Richardson J, Khan MA, Iezzi A, Maxwell A. Cross-national comparison of twelve quality of life instruments, MIC Paper 7: Germany. Centre for Health Economics, Monash University, Melbourne; 2013. Report No.: Research Paper 85.

60. Kharroubi SA, Edlin R, Meads D, McCabe C. Bayesian statistical models to estimate EQ-5D utility scores from EORTC QLQ data in myeloma. Pharm Stat. 2018 Jul 1;17(4):358–71.

61. Park SY, Park EJ, Suh HS, Ha D, Lee EK. Development of a transformation model to derive general population-based utility: Mapping the pruritus-visual analog scale (VAS) to the EQ-5D utility. J Eval Clin Pract. 2017;23(4):755–61.

62. Carreon LY, Bratcher KR, Das N, Nienhuis JB, Glassman SD. Estimating EQ-5D values from the Neck Disability Index and numeric rating scales for neck and arm pain. J Neurosurg Spine. 2014;21(3):394–9.

63. Khairnar R, Pugh SL, Sandler HM, Lee WR, Villalonga Olives E, Mullins CD, Palumbo FB, Bruner DW, Shaya FT, Bentzen SM, Shah AB, Malone SC, Michalski JM, Dayes IS, Seaward SA, Albert M, Currey AD, Pisansky TM, Chen Y, Horwitz EM, DeNittis AS, Feng FY, Mishra MV. Mapping expanded prostate cancer index composite to EQ5D utilities to inform economic evaluations in prostate cancer: Secondary analysis of NRG/RTOG 0415. PLoS One. 2021;16(4):e0249123.

64. Nahvijou A, Safari H, Yousefi M, Rajabi M, Arab-Zozani M, Ameri H. Mapping the cancer-specific FACT-B onto the generic SF-6Dv2. Breast Cancer. 2021;28(1):130–6.

65. Ahadi MS, Vahidpour N, Togha M, Daroudi R, Nadjafi-Semnani F, Mohammadshirazi Z, Akbari-Sari A, Ghorbani Z. Assessment of Utility in Migraine: Mapping the Migraine-Specific Questionnaire to the EQ-5D-5L. Value in health regional issues. 2021;25:57–63.

66. Franken MD, de Hond A, Degeling K, Punt CJA, Koopman M, Uyl-de Groot CA, Versteegh MM, van Oijen MGH. Evaluation of the performance of algorithms mapping EORTC QLQ-C30 onto the EQ-5D index in a metastatic colorectal cancer cost-effectiveness model. Health Qual Life Outcomes. 2020;18(1):240.

67. Kularatna S, Senanayake S, Chen G, Parsonage W. Mapping the Minnesota living with heart failure questionnaire (MLHFQ) to EQ-5D-5L in patients with heart failure. Health Qual Life Outcomes. 2020;18(1):115.

68. Lamu AN. Does linear equating improve prediction in mapping? Crosswalking MacNew onto EQ-5D-5L value sets. Eur J Health Econ. 2020;21(6):903–15.

69. Liu T, Li S, Wang M, Sun Q, Chen G. Mapping the Chinese Version of the EORTC QLQ-BR53 Onto the EQ-5D-5L and SF-6D Utility Scores. The Patient: Patient-Centered Outcomes Research. 2020;13(5):537–55.

70. Su J, Liu T, Li S, Zhao Y, Kuang Y. A mapping study in mainland China: predicting EQ-5D-5L utility scores from the psoriasis disability index. J Med Econ. 2020;23(7):737–43.

71. Sweeney R, Chen G, Gold L, Mensah F, Wake M. Mapping PedsQLTM scores onto CHU9D utility scores: estimation, validation and a comparison of alternative instrument versions. Qual Life Res. 2020;29(3):639–52.

72. Xu RH, Wong ELY, Jin J, Dou Y, Dong D. Mapping of the EORTC QLQ-C30 to EQ-5D-5L index in patients with lymphomas. Eur J Health Econ. 2020;21(9):1363–73.

73. Abdin E, Chong SA, Seow E, Verma S, Tan KB, Subramaniam M. Mapping the Positive and Negative Syndrome Scale scores to EQ-5D-5L and SF-6D utility scores in patients with schizophrenia. Qual Life Res. 2019;28(1):177–86.

74. Vilsboll AW, Kragh N, Hahn-Pedersen J, Jensen CE. Mapping Dermatology Life Quality Index (DLQI) scores to EQ-5D utility scores using data of patients with atopic dermatitis from the National Health and Wellness Study. Qual Life Res. 2020;29(9):2529–39.

75. Ameri H, Yousefi M, Yaseri M, Nahvijou A, Arab M, Akbari Sari A. Mapping the cancer-specific QLQ-C30 onto the generic EQ-5D-5L and SF-6D in colorectal cancer patients. Expert Rev Pharmacoecon Outcomes Res. 2019;19(1):89–96.

76. Robinson T, Oluboyede Y. Estimating CHU-9D Utility Scores from the WAItE: A Mapping Algorithm for Economic Evaluation. Value Health. 2019;22(2):239–46.

77. Sharma R, Gu Y, Sinha K, Aghdaee M, Parkinson B. Mapping the Strengths and Difficulties Questionnaire onto the Child Health Utility 9D in a large study of children. Qual Life Res. 2019;28(9):2429–41.

78. Yang Q, Yu XX, Zhang W, Li H. Mapping function from FACT-B to EQ-5D-5 L using multiple modelling approaches: data from breast cancer patients in China. Health Qual Life Outcomes. 2019;17(1):N.PAG-N.PAG.

79. Gamst-Klaussen T, Lamu AN, Chen G, Olsen JA. Assessment of outcome measures for cost-utility analysis in depression: mapping depression scales onto the EQ-5D-5L. BJPsych Open. 2018;4(4):160–6.

80. Kaambwa B, Ratcliffe J. Predicting EuroQoL 5 Dimensions 5 Levels (EQ-5D-5L) Utilities from Older People's Quality of Life Brief Questionnaire (OPQoL-Brief) Scores. The Patient: Patient-Centered Outcomes Research. 2018;11(1):39–54.

81. Kaambwa B, Smith C, de Lacey S, Ratcliffe J. Does Selecting Covariates Using Factor Analysis in Mapping Algorithms Improve Predictive Accuracy? A Case of Predicting EQ-5D-5L and SF-6D Utilities from the Women's Health Questionnaire. Value Health. 2018;21(10):1205–17.

82. Lamu AN, Chen G, Gamst-Klaussen T, Olsen JA. Do country-specific preference weights matter in the choice of mapping algorithms? The case of mapping the Diabetes-39 onto eight country-specific EQ-5D-5L value sets. Qual Life Res. 2018 Jul;27(7):1801–14.

83. Lamu AN, Olsen JA. Testing alternative regression models to predict utilities: mapping the QLQ-C30 onto the EQ-5D-5L and the SF-6D. Qual Life Res. 2018;27(11):2823–39.

84. Moore A, Young CA, Hughes DA. Mapping ALSFRS-R and ALSUI to EQ-5D in Patients with Motor Neuron Disease. Value Health. 2018;21(11):1322–9.

85. Peak J, Goranitis I, Day E, Copello A, Freemantle N, Frew E. Predicting health-related quality of life (EQ-5D-5 L) and capability wellbeing (ICECAP-A) in the context of opiate dependence using routine clinical outcome measures: CORE-OM, LDQ and TOP. Health Qual Life Outcomes. 2018;16(1):106.

86. Wee HL, Yeo KK, Chong KJ, Khoo EYH, Cheung YB. Mean Rank, Equipercentile, and Regression Mapping of World Health Organization Quality of Life Brief (WHOQOL-BREF) to EuroQoL 5 Dimensions 5 Levels (EQ-5D-5L) Utilities. Med Decis Making. 2018;38(3):319–33.

87. Wijnen BFM, Mosweu I, Majoie M, Ridsdale L, de Kinderen RJA, Evers S, McCrone P. A comparison of the responsiveness of EQ-5D-5L and the QOLIE-31P and mapping of QOLIE-31P to EQ-5D-5L in epilepsy. Eur J Health Econ. 2018;19(6):861–70.

88. Collado-Mateo D, Chen G, Garcia-Gordillo MA, Iezzi A, Adsuar JC, Olivares PR, Gusi N. ‘Fibromyalgia and quality of life: mapping the revised fibromyalgia impact questionnaire to the preference-based instruments’. Health Qual Life Outcomes. 2017;15(1):114.

89. Crump RT, Lai E, Liu G, Janjua A, Sutherland JM. Establishing utility values for the 22-item Sino-Nasal Outcome Test (SNOT-22) using a crosswalk to the EuroQol-five-dimensional questionnaire-three-level version (EQ-5D-3L). Int Forum Allergy Rhinol. 2017;7(5):480–7.

90. Dzingina MD, McCrone P, Higginson IJ. Does the EQ-5D capture the concerns measured by the Palliative care Outcome Scale? Mapping the Palliative care Outcome Scale onto the EQ-5D using statistical methods. Palliat Med. 2017;31(8):716–25.

91. Joyce VR, Sun H, Barnett PG, Bansback N, Griffin SC, Bayoumi AM, Anis AH, Sculpher M, Cameron W, Brown ST, Holodniy M, Owens DK. Mapping MOS-HIV to HUI3 and EQ-5D-3L in Patients With HIV. MDM Policy & Practice. 2017;2(2):2381468317716440.

92. Kaambwa B, Chen G, Ratcliffe J, Iezzi A, Maxwell A, Richardson J. Mapping Between the Sydney Asthma Quality of Life Questionnaire (AQLQ-S) and Five Multi-Attribute Utility Instruments (MAUIs). Pharmacoeconomics. 2017;35(1):111–24.

93. Wong CKH, Cheung PWH, Samartzis D, Luk KD, Cheung KMC, Lam CLK, Cheung JPY. Mapping the SRS-22r questionnaire onto the EQ-5D-5L utility score in patients with adolescent idiopathic scoliosis. PLoS ONE [Electronic Resource]. 2017;12(4):e0175847.

94. Chen G, Finger RP, Holloway EE, Iezzi A, Richardson J. Estimating Utility Weights for the Vision Related Quality of Life Index. Optom Vis Sci. 2016;93(12):1495–501.

95. Acaster S, Pinder B, Mukuria C, Copans A. Mapping the EQ-5D index from the cystic fibrosis questionnaire-revised using multiple modelling approaches. Health Qual Life Outcomes. 2015;13:33.

96. Chen G, Iezzi A, McKie J, Khan MA, Richardson J. Diabetes and quality of life: Comparing results from utility instruments and Diabetes-39. Diabetes Research & Clinical Practice. 2015;109(2):326–33.

97. Furber G, Segal L, Leach M, Cocks J. Mapping scores from the Strengths and Difficulties Questionnaire (SDQ) to preference-based utility values. Qual Life Res. 2014;23(2):403–11.

98. Kim SH, Kim SO, Lee SI, Jo MW. Deriving a mapping algorithm for converting SF-36 scores to EQ-5D utility score in a Korean population. Health Qual Life Outcomes. 2014;12(1):145–145.

99. Longworth L, Yang Y, Young T, Mulhern B, Hernandez Alava M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P, Devianee Keetharuth A, Brazier J. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. Health Technol Assess. 2014;18(9):1–224.

100. Mihalopoulos C, Chen G, Iezzi A, Khan MA, Richardson J. Assessing outcomes for cost-utility analysis in depression: comparison of five multi-attribute utility instruments with two depression-specific outcome measures. British Journal of Psychiatry. 2014;205(5):390–7.

101. Voko Z, Nemeth R, Nagyjanosi L, Jermendy G, Winkler G, Hidvegi T, Kalotai Z, Kalo Z. Mapping the Nottingham Health Profile onto the Preference-Based EuroQol-5D Instrument for Patients with Diabetes. Value in Health Regional Issues. 2014;4:31–6.

102. Yang Y, Wong MY, Lam CL, Wong CK. Improving the mapping of condition-specific health-related quality of life onto SF-6D score. Qual Life Res. 2014;23(8):2343–53.

103. Kay S, Tolley K, Colayco D, Khalaf K, Anderson P, Globe, D. Mapping EQ-5D utility scores from the Incontinence Quality of Life Questionnaire among patients with neurogenic and idiopathic overactive bladder. Value Health. 2013;16(2):394–402.

104. Teckle P, McTaggart-Cowan H, Van der Hoek K, Chia S, Melosky B, Gelmon K, Peacock S. Mapping the FACT-G cancer-specific quality of life instrument to the EQ-5D and SF-6D. Health Qual Life Outcomes. 2013;11:203.

105. Xie F, Pullenayegum EM, Li SC, Hopkins R, Thumboo J, Lo NN. Use of a disease-specific instrument in economic evaluations: mapping WOMAC onto the EQ-5D utility index. Value Health. 2010;13(8):873–8.

106. Gartner FR, Marinus J, van den Hout WB, Vleggeert-Lankamp C, Stiggelbout AM. The Cervical Radiculopathy Impact Scale: development and evaluation of a new functional outcome measure for cervical radicular syndrome. Disability & Rehabilitation. 2020;42(13):1894–905.

107. Mlcoch T, Sedova L, Stolfa J, Urbanova M, Suchy D, Smrzova A, Jircikova J, Pavelka K, Dolezal T. Mapping the relationship between clinical and quality-of-life outcomes in patients with ankylosing spondylitis. Expert Rev Pharmacoecon Outcomes Res. 2017;17(2):203–11.

108. Grochtdreis T, Brettschneider C, Hajek A, Schierz K, Hoyer J, Koenig HH. Mapping the Beck Depression Inventory to the EQ-5D-3L in Patients with Depressive Disorders. J Ment Health Policy Econ. 2016;19(2):79–89.

109. Lindkvist M, Feldman I. Assessing outcomes for cost-utility analysis in mental health interventions: mapping mental health specific outcome measure GHQ-12 onto EQ-5D-3L. Health Qual Life Outcomes. 2016;14(1):134.

110. Rundell SD, Bresnahan BW, Heagerty PJ, Comstock BA, Friedly JL, Jarvik JG, Sullivan SD. Mapping a patient-reported functional outcome measure to a utility measure for comparative effectiveness and economic evaluations in older adults with low back pain. Med Decis Making. 2014;34(7):873–83.

111. Oppe M, Devlin N, Black N. Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. Value Health. 2011 Sep;14(6):884–91.

112. Wijeysundera HC, Tomlinson G, Norris CM, Ghali WA, Ko DT, Krahn MD. Predicting EQ-5D utility scores from the Seattle Angina Questionnaire in coronary artery disease: a mapping algorithm using a Bayesian framework. Med Decis Making. 2011;31(3):481–93.

113. Dixon P, Hollingworth W, Sparrow J. Mapping to Quality of Life and Capability Measures in Cataract Surgery Patients: From Cat-PROM5 to EQ-5D-3L, EQ-5D-5L, and ICECAP-O Using Mixture Modelling. MDM Policy & Practice. 2020;5(1):2381468320915447.

114. Gray LA, Hernandez Alava M, Wailoo AJ. Mapping the EORTC QLQ-C30 to EQ-5D-3L in patients with breast cancer. BMC Cancer. 2021;21(1):1237.

115. Yousefi M, Behzadi Sheikhrobat Y, Najafi S, Ghaffari S, Ghaderi H, Memarzadeh SE, Mahboub-Ahari A, Barouni M, Biglu MH. Mapping catquest scores onto EQ-5D utility values in patients with cataract disease. Iran Red Crescent Med J [Internet]. 2016 Mar 6;19(5). Available from: https://sites.kowsarpub.com/ircmj/articles/14991.html

116. Boland MR, van Boven JF, Kocks JW, van der Molen T, Goossens LM, Chavannes NH, Rutten-van Molken MP. Mapping the clinical chronic obstructive pulmonary disease questionnaire onto generic preference-based EQ-5D values. Value Health. 2015;18(2):299–307.

117. Le QA. Probabilistic mapping of the health status measure SF-12 onto the health utility measure EQ-5D using the US-population-based scoring models. Qual Life Res. 2014;23(2):459–66.

118. Hernández Alava M, Wailoo A, Wolfe F, Michaud K. The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis. Rheumatology . 2013;52(5):944–50.

119. Askew RL, Swartz RJ, Xing Y, Cantor SB, Ross MI, Gershenwald JE, Palmer JL, Lee JE, Cormier JN. Mapping FACT-melanoma quality-of-life scores to EQ-5D health utility weights. Value Health. 2011;14(6):900–6.

120. Cheung YB, Thumboo J, Gao F, Ng GY, Pang G, Koo WH, Sethi VK, Wee J, Goh C. Mapping the English and Chinese versions of the Functional Assessment of Cancer Therapy-General to the EQ-5D utility index. Value Health. 2009;12(2):371–6.

121. Sullivan PW, Ghushchyan V. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. Med Decis Making. 2006;26(4):401–9.

122. Coon C, Bushmakin A, Tatlock S, Williamson N, Moffatt M, Arbuckle R, Abraham L. Evaluation of a crosswalk between the European Quality of Life Five Dimension Five Level and the Menopause-Specific Quality of Life questionnaire. Climacteric. 2018;21(6):566–73.

123. Madan J, Khan KA, Petrou S, Lamb SE. Can Mapping Algorithms Based on Raw Scores Overestimate QALYs Gained by Treatment? A Comparison of Mappings Between the Roland-Morris Disability Questionnaire and the EQ-5D-3L Based on Raw and Differenced Score Data. Pharmacoeconomics. 2017;35(5):549–59.

124. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol. 1997;36(5):551–9.

125. Hoyle CK, Tabberer M, Brooks J. Mapping the COPD Assessment Test onto EQ-5D. Value Health. 2016;19(4):469–77.

126. Nair SC, Welsing PM, Marijnissen AK, Sijtsma P, Bijlsma JW, van Laar JM, Lafeber FP, de Wit GA. Does disease activity add to functional disability in estimation of utility for rheumatoid arthritis patients on biologic treatment? Rheumatology. 2016;55(1):94–102.

127. Versteegh MM, Leunis A, Luime JJ, Boggild M, Uyl-de Groot CA, Stolk EA. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. Med Decis Making. 2012;32(4):554–68.

128. Davison NJ, Thompson AJ, Turner AJ, Longworth L, McElhone K, Griffiths CEM, Payne K, Group, Badbir Study. Generating EQ-5D-3L Utility Scores from the Dermatology Life Quality Index: A Mapping Study in Patients with Psoriasis. Value Health. 2018;21(8):1010–8.

129. Pennington BM, Hernandez-Alava M, Hykin P, Sivaprasad S, Flight L, Alshreef A, Brazier J. Mapping From Visual Acuity to EQ-5D, EQ-5D With Vision Bolt-On, and VFQ-UI in Patients With Macular Edema in the LEAVO Trial. Value Health. 2020;23(7):928–35.

130. Kaambwa B, Billingham L, Bryan S. Mapping utility scores from the Barthel index. Eur J Health Econ. 2013;14(2):231–41.

131. Erim DO, Bennett AV, Gaynes BN, Basak RS, Usinger D, Chen RC. Mapping the Memorial Anxiety Scale for Prostate Cancer to the SF-6D. Qual Life Res. 2021;30(10):2919–28.

132. Ayala A, Forjaz MJ, Ramallo-Fariña Y, Martín-Fernández J, García-Pérez L, Bilbao A. Response Mapping Methods to Estimate the EQ-5D-5L From the Western Ontario McMaster Universities Osteoarthritis in Patients With Hip or Knee Osteoarthritis. Value Health. 2021;24(6):874–83.

133. Bilbao A, Martin-Fernandez J, Garcia-Perez L, Arenaza JC, Ariza-Cardiel G, Ramallo-Farina Y, Ansola L. Mapping WOMAC Onto the EQ-5D-5L Utility Index in Patients With Hip or Knee Osteoarthritis. Value Health. 2020;23(3):379–87.

134. Martín-Fernández J, Morey-Montalvo M, Tomás-García N, Martín-Ramos E, Muñoz-García JC, Polentinos-Castro E, Rodríguez-Martínez G, Arenaza JC, García-Pérez L, Magdalena-Armas L, Bilbao A. Mapping analysis to predict EQ-5D-5 L utility values based on the Oxford Hip Score (OHS) and Oxford Knee Score (OKS) questionnaires in the Spanish population suffering from lower limb osteoarthritis. Health Qual Life Outcomes. 2020;18(1):1–15.

135. Shi Y, Thompson J, Walker AS, Paton NI, Cheung YB, Team, Earnest Trial. Mapping the medical outcomes study HIV health survey (MOS-HIV) to the EuroQoL 5 Dimension (EQ-5D-3 L) utility index. Health Qual Life Outcomes. 2019;17(1):83.

136. Ward Fuller G, Hernandez M, Pallot D, Lecky F, Stevenson M, Gabbe B. Health State Preference Weights for the Glasgow Outcome Scale Following Traumatic Brain Injury: A Systematic Review and Mapping Study. Value Health. 2017;20(1):141–51.

137. Frew EJ, Harrison M, Rossello Roig M, Martin TP. Providing an extended use of an otological-specific outcome instrument to derive cost-effectiveness estimates of treatment. Clin Otolaryngol. 2015;40(6):593–9.

138. Cheung YB, Luo N, Ng R, Lee CF. Mapping the functional assessment of cancer therapy-breast (FACT-B) to the 5-level EuroQoL Group's 5-dimension questionnaire (EQ-5D-5L) utility index in a multi-ethnic Asian population. Health Qual Life Outcomes. 2014;12:180.

139. Lee L, Kaneva P, Latimer E, Feldman LS. Mapping the Gastrointestinal Quality of Life Index to short-form 6D utility scores. J Surg Res. 2014;186(1):135–41.

140. Hawton A, Green C, Telford C, Zajicek J, Wright D. Using the Multiple Sclerosis Impact Scale to estimate health state utility values: mapping from the MSIS-29, version 2, to the EQ-5D and the SF-6D. Value Health. 2012;15(8):1084–91.

141. Barton GR, Sach TH, Jenkinson C, Avery AJ, Doherty M, Muir KR. Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores? Health Qual Life Outcomes. 2008;6:51.

142. Cheung YB, Tan HX, Luo N, Wee HL, Koh GCH. Mapping the Shah-modified Barthel Index to the Health Utility Index Mark III by the Mean Rank Method. Qual Life Res. 2019;28(12):3177–85.

143. Lee CF, Ng R, Luo N, Cheung YB. Development of Conversion Functions Mapping the FACT-B Total Score to the EQ-5D-5L Utility Value by Three Linking Methods and Comparison with the Ordinary Least Square Method. Appl Health Econ Health Policy. 2018;16(5):685–95.

144. Xu R, Insinga RP, Golden W, Hu XH. EuroQol (EQ-5D) health utility scores for patients with migraine. Qual Life Res. 2011 May 1;20(4):601–8.

145. Rendas-Baum R, Yang M, Varon SF, Bloudek LM, DeGryse RE, Kosinski M. Validation of the Headache Impact Test (HIT-6) in patients with chronic migraine. Health Qual Life Outcomes. 2014 Aug 1;12(1):117.

146. Martin M, Blaisdell B, Kwong JW, Bjorner JB. The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. J Clin Epidemiol. 2004 Dec 1;57(12):1271–8.

147. Diener HC, Ashina M, Durand-Zaleski I, Kurth T, Lantéri-Minet M, Lipton RB, Ollendorf DA, Pozo-Rosich P, Tassorelli C, Terwindt G. Health technology assessment for the acute and preventive treatment of migraine: A position statement of the International Headache Society. Cephalalgia. 2021 Mar;41(3):279–93.

148. Jones DS, Podolsky SH. The history and fate of the gold standard. Lancet. 2015;385(9977):1502–3.

149. Straube A, Pfaffenrath V, Ladwig KH, Meisinger C, Hoffmann W, Fendrich K, Vennemann M, Berger K. Prevalence of chronic migraine and medication overuse headache in Germany—the German DMKG headache study. Cephalalgia. 2010;30(2):207–13.

150. Statista. Durchschnittsalter der Bevölkerung in Deutschland nach Bundesländern* im Jahr 2018 [Internet]. [cited 2020 Mar 7]. Available from: https://de.statista.com/statistik/daten/studie/1093993/umfrage/durchschnittsalter-der-bevoelkerung-in-deutschland-nach-bundeslaendern/

151. Kiel S. Wer leidet? [Internet]. 2009 [cited 2020 Jul 13]. Available from: https://schmerzklinik.de/service-fuer-patienten/migraene-wissen/wer-leidet/

152. Holmes Finch W, Bolin JE, Kelley K. Multilevel Modeling Using R. CRC Press; 2019. 252 p.

153. Hernandez Alava M, Wailoo A, Pudney S, Gray L, Manca A. Mapping clinical outcomes to generic preference-based outcome measures: development and comparison of methods. Health Technol Assess. 2020;24(34):1–68.

154. Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value Health. 2013;16(1):202–10.

155. Zwaap J, Knies S, van der Meijden C, Staal P, van der Heiden L. Cost-effectiveness in practice. Zorginstituut Nederland; 2015.

156. Neumann PJ, Cohen JT, Weinstein MC. Updating Cost-Effectiveness — The Curious Resilience of the $50,000-per-QALY Threshold. Vol. 371, New England Journal of Medicine. 2014. p. 796–7.

157. Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: recommendations for the design, analysis, and reporting of studies. Int J Technol Assess Health Care. 2005 Spring;21(2):165–71.

158. York Health Economics Consortium. Net Monetary Benefit [Internet]. 2016 [cited 2023 Apr 28]. Available from: https://yhec.co.uk/glossary/net-monetary-benefit/

159. Wendt M, Ebinger M, Kunz A, Rozanski M, Waldschmidt C, Weber JE, Winter B, Koch PM, Freitag E, Reich J, Schremmer D, Audebert HJ, STEMO Consortium. Improved prehospital triage of patients with stroke in a specialized stroke ambulance: results of the pre-hospital acute neurological therapy and optimization of medical care in stroke study. Stroke. 2015 Mar;46(3):740–5.

160. Barral M, Armoiry X, Boudour S, Aulagner G, Schott AM, Turjman F, Gory B, Viprey M. Cost-effectiveness of stent-retriever thrombectomy in large vessel occlusion strokes of the anterior circulation: Analysis from the French societal perspective. Rev Neurol . 2020 Mar;176(3):180–8.

161. Gonçalves ASO, Huerta-Gutierrez R, Cuartero ET, Piccininni M, Kurth T, Jones L, Rohmann JL. Protocol for Assessing “The benefits and costs of good methods: A Systematic Methods Overview of Economic evaluations in the field of stroke”. OSF [Internet]. 2022; Available from: http://dx.doi.org/10.17605/OSF.IO/H58E2

162. Kendir C, Naik R, Bloemeke J, de Bienassis K, Larrain N, Klazinga N, Guanais F, van den Berg M. All hands on deck: Co-developing the first international survey of people living with chronic conditions: Stakeholder engagement in the design, development, and field trial implementation of the PaRIS survey. OECD Health Working Papers No. 149; 2023.

9. Statutory Declaration

“I, Ana Sofia Oliveira Gonçalves, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic *“Patient-reported outcome measures: applications and challenges in the field of neurology”* (*“Patient-reported outcome measures: Anwendungen und Herausforderungen im Bereich der Neurologie”*), independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding statistical processing) and results (in particular regarding figures and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; http://www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

Date Signature

10. Declaration of your own contribution to the publications

Ana Sofia Oliveira Gonçalves contributed the following to the below listed publications:

Publication 1: **Ana Sofia Oliveira Gonçalves**, Dimitra Panteli, Lars Neeb, Tobias Kurth & Annette Aigner, HIT-6 and EQ-5D-5L in patients with migraine: assessment of common latent constructs and development of a mapping algorithm, The European Journal of Health Economics, 2021

Contribution:

In Thesis Article 1, I was responsible for the study conception and design. I downloaded the data from the electronic case report form REDCap, where SMARTGEM data was stored. I wrote the statistical code for the statistical analyses by myself, in both R 3.6.3 and Stata 15. I independently created all original figures and tables, which were later edited by the journal graphic design team. I substantially contributed to the interpretation of the results, and I drafted the first version of the manuscript that then was revised by my supervisors. Furthermore, I independently coordinated the journal submission and the revision process and compiled input from all co-authors.

Publication 2: **Ana Sofia Oliveira Gonçalves**, Sophia Werdin, Tobias Kurth, Dimitra Panteli, Mapping Studies to Estimate Health-State Utilities From Non–Preference-Based Outcome Measures: A Systematic Review on How Multiple observations are Taken Into Account, Value in Health, 2022

Contribution:

In Thesis Article 2, me and one of my supervisors were responsible for the study conception and design. Since this manuscript is a systematic literature review, the steps “Title/Abstract screening” and “Full Text screening” were independently carried out by me and then by my co-author Sophia Werdin. I was the sole responsible for the data extraction and statistical analysis. I created all the original tables and figures included in the manuscript, which were later edited by the journal graphic design team. Additionally, I developed the first interpretation of the results and drafted the first version of the manuscript. Furthermore, I coordinated the journal submission and the entire revision process.

Publication 3: **Ana Sofia Oliveira Gonçalves**, Jessica Lee Rohmann, Marco Piccininni, Tobias Kurth, Martin Ebinger, Matthias Endres, Erik Freitag, Peter Harmel, Irina Lorenz-Meyer, Ira

Rohrpasser-Napierkowski, Reinhard Busse, Heinrich J. Audebert. Economic Evaluation of a Mobile Stroke Unit Service in Germany. Annals of Neurology, 2023

Together with my co-author Jessica Lee Rohmann, I was responsible for the study conception and design, statistical analysis and interpretation, and drafting of the manuscript. Furthermore, I was responsible, together with my co-author Marco Piccininni, for writing the R code, conducting the statistical analysis and interpreting the results from the analyses. I created all the original tables and figures included in the manuscript, which were later edited by the journal graphic design team. I drafted the first version of the manuscript. Furthermore, I coordinated the journal submission and the revision process and compiled input from all co-authors.

_____

Signature, date and stamp of first supervising university professor / lecturer

_____

Signature of doctoral candidate

11. Printing copy(s) of the publication(s)

11.1 Research Project 1 – Mapping algorithm development

The European Journal of Health Economics (2022) 23:47–57
https://doi.org/10.1007/s10198-021-01342-9

ORIGINAL PAPER

HIT-6 and EQ-5D-5L in patients with migraine: assessment of common latent constructs and development of a mapping algorithm

Ana Sofia Oliveira Gonçalves[1] · Dimitra Panteli[2] · Lars Neeb[3] · Tobias Kurth[1] · Annette Aigner[4,5]

Received: 25 November 2020 / Accepted: 29 June 2021 / Published online: 10 July 2021

Abstract

Objective The aims of this study were to assess whether there is a conceptual overlap between the questionnaires HIT-6 and EQ-5D and to develop a mapping algorithm allowing the conversion of HIT-6 to EQ-5D utility scores for Germany.

Methods This study used data from an ongoing randomised controlled trial for patients suffering from migraine. We assessed the conceptual overlap between the two instruments with correlation matrices and exploratory factor analysis. Linear regression, tobit, mixture, and two-part models were used for mapping, accounting for repeated measurements, tenfold cross-validation was conducted to validate the models.

Results We included 1010 observations from 410 patients. The EQ-5D showed a substantial ceiling effect (47.3% had the highest score) but no floor effect, while the HIT-6 showed a very small ceiling effect (0.5%). The correlation between the instruments' total scores was moderate (−0.30), and low to moderate among each domain (0.021–0.227). The exploratory factor analysis showed insufficient conceptual overlap between the instruments, as they load on different factors. Thus, there is reason to believe that the instruments' domains do not capture the same latent constructs. To facilitate future mapping, we provide coefficients and a variance–covariance matrix for the preferred model, a two-part model with the total HIT-6 score as the explanatory variable.

Conclusion This study showed that the German EQ-5D and the HIT-6 lack the conceptual overlap needed for appropriate mapping. Thus, the estimated mapping algorithms should only be used as a last resort for estimating utilities to be employed in economic evaluations.

Keywords Mapping · EQ-5D · QALY · Utilities · HIT-6 · Migraine

JEL Classification I1 · C3

Introduction

Migraine is a common neurological condition affecting 10.6% of the German population (one-year prevalence) [1]. It is associated with comorbidities such as psychiatric

Tobias Kurth and Annette Aigner contributed equally to this work.

✉ Ana Sofia Oliveira Gonçalves
ana.goncalves@charite.de

Dimitra Panteli
dimitra.panteli@tu-berlin.de

Lars Neeb
lars.neeb@charite.de

Tobias Kurth
tobias.kurth@charite.de

Annette Aigner
annette.aigner@charite.de

[1] Institute of Public Health, Charité - Universitätsmedizin, Berlin, Charitépl. 1, 10117 Berlin, Germany

[2] Department of Health Care Management, Technische Universität Berlin, Berlin, Germany

[3] Department of Neurology, Charité - Universitätsmedizin Berlin, Berlin, Germany

[4] Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Berlin, Germany

[5] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany

disorders (depression and anxiety, among others), respiratory disorders, and chronic pain, and it leads to a significant reduction in quality of life [2, 3].

This condition also imposes an economic burden on health care systems due to increased demand for goods and services and work-related productivity losses [4, 5]. As healthcare systems face the challenge of limited resources, economic evaluations provide tools for decision-makers to analyse competing alternatives—in terms of both costs and consequences [6]. Cost-utility analyses, a form of economic evaluation, measure consequences with generic measures of health gain, commonly expressed in quality-adjusted life years (QALYs) [6]. The EuroQol five-dimensional questionnaire (EQ-5D) is a generic utility-based instrument which allows the estimation of utility scores, and thus, the calculation of country-specific QALYs [7]. It analyses five different dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The initial version of EQ-5D had only three levels within each dimension, while the improved EQ-5D-5L, henceforth EQ-5D-5L will be referred to as EQ-5D, has five levels, while maintaining the same five dimensions. The levels indicate no problems (1), slight problems (2), moderate problems (3), severe problems (4), and unable to/extreme problems (5). Health states are defined by combining digits for the five dimensions, enabling 3125 possible health states. Health states can be represented with five-digit codes or converted using country-specific single index values.

Clinical trials in migraine often use monthly migraine days as a primary endpoint and the International Headache Society actually recommends the use of monthly migraine days as the primary endpoint for HTAs involving preventive treatments [8]. Where generic preference-based measures are not deemed ideal, other approaches include using condition-specific instruments or condition-specific preference-based measures (to our knowledge there is none for migraine). However, analyses with disease-specific instruments do not allow decision-makers to compare resource allocation across different conditions.

Nevertheless, several trials in the migraine field (e.g. [9, 10]) only collect migraine-specific health-related quality-of-life (HRQOL) instruments, but do not collect generic preference-based ones, which can be used to conduct cost-utility analyses. There are several migraine-specific HRQOL questionnaires such as the Headache Impact Test-6 (HIT-6), the Migraine Disability Assessment (MIDAS), and the Migraine-Specific Quality-of-Life Questionnaire (MSQ). The HIT-6 is a headache-specific questionnaire, which evaluates how headaches affect someone's ability to function on the job, at school, at home, and in social situations [11]. This instrument does not have a preference-based scoring system, thus it does not permit the calculation of QALYs. Mapping overcomes this issue by providing an algorithm which allows the estimation of QALYs even if a preference-based HRQOL instrument was not included in the study. However, to perform mapping between two instruments, there should be a conceptual overlap between them. ISPOR guidelines on mapping state that these algorithms can only be successful if there is sufficient overlap between the analysed instruments [12]. Although the selected instruments do not have to measure the same symptoms or functional (dis)abilities, they do need to address the same underlying concepts.

One study by Gillard et al. has already mapped the EQ-5D to the HIT-6, but used quality weights for England in a Brazilian population [13]. Several studies have shown the impact of using different country-specific value sets of EQ-5D on the interpretation of results [14, 15]. Furthermore, the authors used variables in their algorithm which are not always collected in trials, such as ethnicity. A large number of trials which do not involve drugs but e.g. behavioural interventions often do not collect ethnic information (e.g. [16–19]). Applying the validation method of splitting the available data set into two has been criticised because of its limited ability to depict the uncertainty in the results and increased bias in the performance estimates in proportionally large test sets [20].

Based on these considerations, we will address the issue of whether there is enough conceptual overlap between the two instruments using not only correlation tables, but also exploratory factor analysis (EFA). Based on regression-type approaches we will develop a mapping function to predict EQ-5D utility values for Germany from HIT-6 values, including variables widely used in migraine trials, and validate them with tenfold cross-validation.

Materials and methods

Data

This study is based on data from the SMARTGEM project. SMARTGEM is an ongoing national randomised controlled clinical trial, which seeks to assess if a digital intervention via the use of a headache app and online consultations leads to a decrease in migraine frequency. The intervention consists of a certified medical app where patients document trigger factors, attacks, and medication in an electronic calendar; the app analyses the diary and evaluates trigger factors and proposes individually tailored treatment plans; a web-based tool where patients communicate both with other patients and specialists. HIT-6 and EQ-5D were completed by all users at baseline and after 3, 6, 9, and 12 months. Registration ID in the German clinical trials register is DRKS00016328.

Statistical analyses

Conceptual overlap

To analyse the strength of the relationship between HIT-6 scores and EQ-5D domains, correlation coefficients accounting for repeated measurements were computed. We also examined the capacity of each instrument to detect changes in HRQOL over time, referred to as responsiveness, by computing standardised response mean(s) (SRM). SRM is defined as a ratio of the difference in the mean baseline and mean follow-up values divided by their mean standard deviations' (sd) difference. We considered SMR values of less than 0.2 as small, from 0.2 to 0.5 as moderate, and values above 0.8 as large, following Cohen's criteria [21]. EFA was conducted to explore the overlap in the underlying constructs of the two instruments. If factors have meaningful loadings from the two different instruments (EQ-5D and HIT-6), these instruments are assumed to capture the same underlying latent structure. We considered factor loadings above 0.3 as 'meaningful' [22]. For ordered data, the preferred method to determine the number of factors is to conduct parallel analysis with polychoric correlations instead of Pearson correlations [23]. It is believed that Pearson correlations underestimate the relationship between ordered categorical data because of the categorisation [24]. Furthermore, Glorfeld (1995) showed that parallel analysis performs well with non-normally distributed data [25]. The chosen factoring mode was weighted least squares, which makes no distributional assumption, thus being appropriate for ordinal data [26]. Varimax and promax rotations were used to interpret factor loadings.

Mapping model development

Since there is no specific model recommended by guidelines on best practices for mapping, we applied several models [12]. As our data contains repeated measurements per individual over time, we accounted for dependencies between observations by including random effects in our models, and estimated mixed-effects linear regression models (fit by maximum likelihood), mixed-effects tobit censored at the upper bound at 1, adjusted limited dependent variable mixture models, mixture beta regression models, and two-part models. For mixed-effects linear models and mixed-effects tobit, we compared models where the overall HIT-6 score versus the HIT-6 several dimensions were used as independent variables. Interaction terms and quadratic terms were considered. Models with the lowest BIC (Bayesian information criterion) were chosen. With regard to the two-part model, in a first stage, a mixed-effects logistic regression is fit to predict the probability of a respondent having full health. In a second stage a mixed-effects linear regression only based on those without full health was estimated. The overall expected EQ-5D index score was calculated using an expected value approach [27].

$$E(\mathrm{EQ}-5D) = \Pr(\text{Full Health}) * (\mathrm{EQ}-5 \text{ in Full Health}) + (1 - \Pr(\text{Full Health})) * (\mathrm{EQ}-5 \text{ Not in Full Health}).$$

We also fitted adjusted limited dependent variable mixture models with one to four components, with the Stata command aldvmm, which was specifically developed to deal with health utility data [28]. These models allow to limit the dependent variable to the EQ-5D country-specific range, while taking into account the gap between 1 and the next feasible value (0.974 in the case of Germany). We conducted both adjusted limited dependent variable mixture models and mixture beta regression models with and without the inclusion of this truncation point, as well as with and without the inclusion of a probability mass at full health and at this truncation point (for beta mixture models only). We used the estimated parameters from a constant-only model in our mixture models, to find the global maximum, since mixture models are known to have multiple optima [28]. Unlike in the other models, we could not account for repeated measures by including random effects. We did, however, compute robust cluster-corrected standard errors.

HIT-6 related variables, sex, age, and migraine type were pre-defined as having to be part of the model. Given that mapping algorithms are intended to be used by other researchers, we only considered age and sex[1] as possible socio-demographic explanatory variables, since they are almost always collected in studies. Studies have shown that age has an impact on the symptoms of people with migraine, e.g. decrease in frequency of photophobia and phonophobia [29]. Migraine also affects three times more women than men, and it is known that fluctuations in female hormones play an important role in this relationship [30, 31]. Since especially in women, the impact of migraines varies with age, we tested whether there is an interaction between age and sex [29]. We also included the information whether patients suffered from episodic or chronic migraines and its interaction with age. Migraine characteristics evolve across time (e.g. the conversion of episodic to chronic migraine), thus the importance of testing the inclusion of an interaction term between migraine type and age [32].

We conducted complete-case analysis based on the following variables: EQ-5D domains, HIT-6 domains, migraine type, age, and sex.

[1] In German, there are no different terms to define sex versus gender. The term "Geschlecht" can be both understood as sex or gender. In this project, participants filled in their own "Geschlecht".

Validation

We plotted the observed and predicted EQ-5D values to visualise the models' performance. Given the lack of external data to conduct external validation, a tenfold cross-validation was carried out to compare the predictions of each model with the actual EQ-5D scores. This method is recommended for small samples [20]. Models' predictive performance was assessed with root mean squared error (RMSE), mean absolute error (MAE), and R^2, reporting the mean of all 10 cycles.

Statistical analyses were performed using *R* 3.6.3 and Stata 15 [33, 34]. We used additional *R* packages for data handling [35] and plotting [36], repeated measures correlation [37], and factor analysis [38, 39], a Stata package for variable selection [40] (a preliminary version of gsreg 2.0 provided by the authors was used, which allowed its use with a mixed-effects linear regression estimated by maximum-likelihood), the package aldvmm to fit adjusted limited dependent variable mixture models [28], as well as the betamix package for conducting beta mixture regressions [40].

Results

The dataset used for the analysis contains 1010 observations, based on 410 patients, as 16 patients had missing data, such that 22 out of the 1032 (2.13%) observations had to be removed. thus the dataset used for the analysis contains 1010 observations, based on 410 patients. 7 out of 16 were excluded from the analysis because they did not have full data on other time points.

87.3% of all participants were female, with an average age of 41 years (Table 1).

Health utility values derived from EQ-5D ranged from −0.57 to 1. We observed a ceiling effect in EQ-5D scores, with a skewness of −2.33 and a kurtosis of 9.45, pointing to a left skew with few negative observations (Fig. 1). Data are considerably more skewed for patients with episodic migraine than for patients with chronic migraine. The mean EQ-5D utility value was 0.82 (sd 0.23) for all patients, 0.86 (sd 0.18) for patients with episodic migraine, and 0.72 (sd 0.30) for patients with chronic migraine.

HIT-6 scores ranged from 44 to 78 (possible score range 36–78). The skewness of −0.64 indicated that the HIT-6 scores are only slightly skewed to the left (Fig. 2). There was no floor effect, no patient had the lowest score possible, and the ceiling effect was small (5 out of 1010 observations; 0.5%).

In EQ-5D, there was no floor effect (proportion of respondents reporting the worst level for all five dimensions), i.e. no patient had the lowest utility score possible (−0.661 in the German value set). However, the ceiling effect (proportion of participants reporting the best level for all dimensions) amounted to 47.3% (194/410).

Conceptual overlap

We consider that occupation and daily activities can be measured by the EQ-5D dimension "usual activities" and by questions 2, 3, and 4 from the HIT-6. Physical health is captured by "pain/discomfort" and "mobility" in the EQ-5D, and by question 5 from the HIT-6. Self-care is only measured by the EQ-5D.

The correlation coefficient between EQ-5D score value and the HIT-6 total score amounted to −0.30. In terms of EQ-5D value and the different HIT-6 dimensions, the coefficients ranged between −0.153 and −0.234. The correlation coefficients between each EQ-5D domain and the overall HIT-6 score ranged from 0.077 to 0.300 (Table A.1). Lastly, the correlation coefficients among each domain from the two instruments ranged from 0.021 to 0.227. The highest correlation (0.227) was found between EQ-5D pain/discomfort and HIT-6 q4. See Supplementary Tables A.1, A.2, and A.3 for correlation tables, additionally stratified by migraine severity level.

The EQ-5D total score and the different dimensions show small SRMs, while the HIT-6 total score and its different questions show small to moderate responsiveness. For EQ-5D dimensions, SRM values range from 0.088 to 0.280 and for the HIT-6 from 0.211 to 0.669 (see Supplementary Table A.4). Although the lack of responsiveness may be in part because we are also analysing patients in the control group, this still does not explain why the responsiveness of the HIT-6 is higher than that of the EQ-5D.

We considered three factors in the EFA. Factor 1 had meaningful loadings (i.e. higher than 0.3) on all EQ-5D domains, but not on HIT-6 domains. Factors 2 and 3 loaded only on HIT-6 domains, specifically questions 2–6 for Factor 2, questions 1 and 2 for Factor 3. Considering that this question had a higher loading in Factor 2, thus belonging to this factor, Factor 2 had meaningful loadings in five out of six HIT-6 domains (Table 2). Similarly, using an orthogonal rotation, all EQ-5D items loaded on the same factor, while HIT-6 items loaded on both Factors 2 and 3 (Supplementary Table A.5).

As the EFA does not correctly take the repeated measurement nature of the data into account, we performed a sensitivity analysis based on baseline data only. The results did not relevantly differ in terms of number of factors and meaningful loadings.

The lack of overlap in all three factors, using the two different types of rotations, suggests that the EQ-5D and the HIT-6 potentially do not capture the same latent constructs.

Table 1 Patient characteristics and measurements of EQ-5D and HIT-6 at baseline

	Chronic migraine (132)	Episodic migraine (278)	Total (410)
Age, mean (SD)	40.1 (11.4)	41.5 (12.0)	41.1 (11.8)
Sex (%)[a]			
Female	119 (90.2%)	239 (86.0%)	358 (87.3%)
Male	13 (9.8%)	39 (14.0%)	52 (12.7%)
BMI			
Mean (SD)	25.0 (4.89)	24.6 (4.66)	24.7 (4.74)
Missing	2 (1.5%)	1 (0.4%)	3 (0.7%)
Comorbidities (%)			
Yes	80 (60.6%)	154 (55.4%)	234 (57.1%)
No	49 (37.1%)	121 (43.5%)	170 (41.5%)
Missing	3 (2.3%)	3 (1.1%)	6 (1.5%)
Marital status (%)			
Married	55 (41.7%)	134 (48.2%)	189 (46.1%)
Single	65 (49.2%)	119 (42.8%)	184 (44.9%)
Widowed	1 (0.8%)	4 (1.4%)	5 (1.2%)
Divorced	8 (6.1%)	18 (6.5%)	26 (6.3%)
Missing	3 (2.3%)	3 (1.1%)	6 (1.5%)
Professional qualification (%)			
Other	11 (8.3%)	11 (4.0%)	22 (5.4%)
University[b]	44 (33.3%%)	120 (43.2%)	164 (40.0%)
Without a degree	12 (9.1%)	24 (8.6%)	36 (8.8%)
Apprenticeship	62 (47.0%)	120 (43.2%)	182 (44.4%)
Missing	3 (2.3%)	3 (1.1%)	6 (1.5%)
Officially recognised disability (%)			
Yes	22 (16.7%)	45 (16.2%)	67 (16.3%)
No	107 (81.1%)	231 (83.1%)	338 (82.4%)
Missing	3 (2.3%)	2 (0.7%)	5 (1.2%)
EQ-5D-5L			
Mean utility from − 0.661 to 1 (SD)	0.689 (0.296)	0.842 (0.198)	0.792 (0.244)
VAS mean from 0 to 100 (SD)	58.0 (23.4)	70.7 (20.2)	66.6 (22.1)
HIT-6			
Mean (SD)	65.2 (4.21)	64.4 (4.38)	64.7 (4.33)
Severity level (%)			
Severe impact	2 (1.5%)	10 (3.6%)	12 (2.9%)
Substantial impact	7 (5.3%)	16 (5.8%)	23 (5.6%)
Some impact	122 (92.4%)	251 (90.3%)	373 (91.0%)
Little or no impact	1 (0.8%)	1 (0.4%)	2 (0.5%)

SD standard deviation, *VAS* Visual Analog Scale.

[a]In German there are no different terms to define sex versus gender. The term "Geschlecht" can be both understood as sex or gender. In this project, participants filled in their own "Geschlecht".

[b]Including university of applied sciences

Mapping models

Table 3 and the Excel file in the Electronic Supplementary Material present information on the models' coefficients and their predictive ability. Overall, for the same statistical method, models which included the HIT-6 total score performed better than those which included all HIT-6 questions as independent variables. The inclusion of interaction terms (between age and sex, migraine type and age, and migraine type and age) did not relevantly improve the prediction of EQ-5D scores within any of the six models. On the contrary, the addition of quadratic terms both for HIT-6

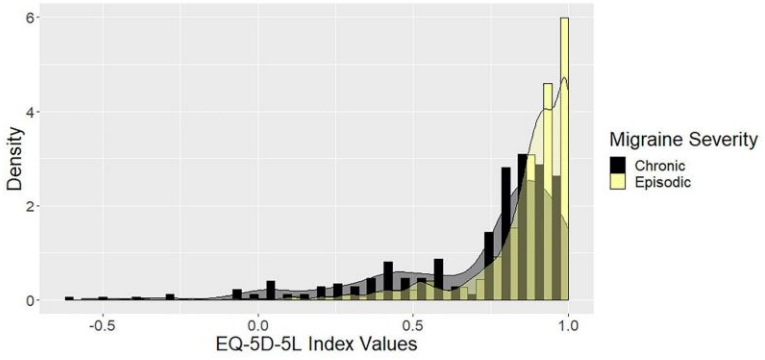


Fig. 1 EQ-5D-5L histogram of number of responses histogram and kernel density plot (for episodic vs chronic migraine)

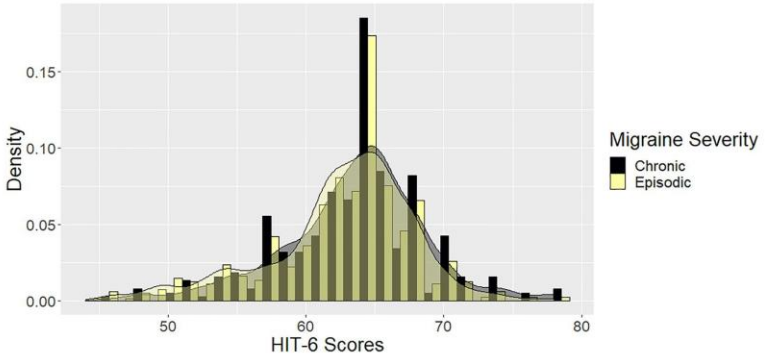


Fig. 2 HIT-6 histogram of number of responses histogram and kernel density plot (for episodic vs chronic migraine)

Table 2 Summary of the Exploratory Factor Analysis (EFA) results for 3 loadings and their cumulative variance (varimax rotation)

	Factor 1' loadings	Factor 2' loadings	Factor 3' loadings
Mobility	0.749	0.125	0.156
Self-care	0.556		0.110
Daily activities	0.840	0.237	
Pain/discomfort	0.761	0.170	0.115
Anxiety/depression	0.433	0.257	
HIT-6 Q1	0.119	0.225	0.629
HIT-6 Q2	0.190	0.539	0.404
HIT-6 Q3		0.342	0.355
HIT-6 Q4	0.196	0.769	0.239
HIT-6 Q5	0.220	0.684	0.167
HIT-6 Q6	0.208	0.858	0.176
Cumulative variance	0.229	0.449	0.528

Meaningful loadings are underlined (i.e. higher than 0.3)

overall score and for several HIT-6 dimensions proved to enhance some of the models with regard to their goodness-of-fit. In the two-part model, the first model only included the total HIT-6 score, the type of migraine, age, and sex, the second included the same variables plus the quadratic term of the HIT-6 score.

Figure 3 shows the observed and the predicted EQ-5D values for the different models. Our models underestimated utilities for those with poorer health states and overestimated them for those with better health states, as is common in mapping studies [41]. Although linear regression models can yield estimates above 1 (given that there is no upper bound), Model A (mixed-effects linear regression with the total HIT-6 score as an independent variable) did not generate estimates out of the bound. For Model A, the maximum predicted value was 0.98 and for Model B (mixed-effects linear regression with the individual HIT-6 questions as independent variables) 1.07.

Table 3 Performance measurements and validation results of 10 evaluated mapping models

Model	Specification	Predicted mean	Predicted minimum	Predicted maximum	Cross-validation		
					RMSE	MAE	Pseudo R^2
Actual EQ-5D-5L value	n.a.	0.817	−0.57	1	n.a.	n.a.	n.a.
Model A	ME Linear	0.8173	0.2488	0.9740	0.1970	0.1380	0.2778
Model B	ME Linear	0.8152	0.3748	1.0704	0.2002	0.1411	0.2558
Model C	ME Tobit	0.8156	0.2376	0.9305	0.1991	0.1366	0.2754
Model D	ME Tobit	0.8004	0.3837	0.9809	0.2046	0.1431	0.2394
Model E	TPM	0.7999	0.1813	0.9469	0.1212	0.1355	0.2843
Model F	TPM	0.7873	0.4389	1.1453	0.1244	0.1424	0.2338
Model G	ALDVMM	0.7882	0.3946	0.9589	0.1992	0.1345	0.2882
Model H	ALDVMM	0.8278	0.4947	0.9797	0.2023	0.1368	0.2593
Model I	BETAMIX with PM at full health	0.8273	0.3910	0.9551	0.1991	0.1347	0.2939
Model J	BETAMIX with PM at full health	0.8259	0.4839	0.9782	0.2018	0.1362	0.2705

ALDVMM adjusted limited dependent variable mixture, *BETAMIX* Beta Mixture Model (with inflation), *MAE* mean absolute error, *ME* mixed-effects, *n.a.* not applicable, *PM* probability mass, *RMSE* root mean square error, *SD* standard deviation, *TPM* two-part model

No model performed best across all goodness-of-fit measures. Model E (two-part model with the total HIT-6 score as the explanatory variable) performed the best in terms of RMSE (Table 3). Although the R^2 value is higher for Model G, this model predicts less well both individuals at full health and those with poorer health states than Model E. The R^2 value is also higher for Model I than E, but the latter predicts poorer health states better. The adjusted limited dependent variable mixture models and beta-mixture models took into account the gap between full health and the next feasible health state. However, the low number of observations (4) with the state directly after full health (0.974) may explain why these models did not perform better.

Hence, if researchers wish to estimate utilities from the HIT-6 to be employed in cost-utility analyses, Model E should be the preferred model. The corresponding variance–covariance matrix is available in the Electronic Supplementary Material, in Table A.6, to allow probabilistic sensitivity analysis to be carried out and account for uncertainty. However, we would like to remark that this mapping algorithm should only be used as a last resort.

Discussion

We aimed to assess whether there is a conceptual overlap between the HIT-6 and the EQ-5D and to present a mapping algorithm for the estimation of the EQ-5D score (with German weights) from the HIT-6 questionnaire, a disease-specific survey widely used in clinical trials with migraine patients. Our study points to major differences in the underlying constructs of the HIT-6 and the EQ-5D. The EQ-5D showed a high ceiling effect and small SRMs across time, whereas the HIT-6 did not show a ceiling effect and had a higher responsiveness. This study also provides a mapping algorithm which can be used to map HIT-6 values to EQ-5D utility values.

We expected some overlap between the two instruments since both have been validated in migraine patients. The strength of association between the instruments measured with correlation coefficients was only low to moderate—both for the total scores and for each instrument's individual questions. Furthermore, the EFA showed that the HIT-6 and EQ-5D do not have a sufficient conceptual overlap and potentially estimate different underlying constructs. There are several reasons that might explain the lack of overlap. First, the recall period in the instruments' questions is different. While all EQ-5D questions refer to the day the questionnaire is filled out, three questions in the HIT-6 refer to the previous 4 weeks. Second, the HIT-6 has frequency response categories (ranging from never to always), while the EQ-5D has response categories based on levels of severity. Third, the specificities of both the EQ-5D and the HIT-6 may also play a role. A criticism of the use of the EQ-5D to describe health utilities in patients with migraine is the fact that the survey is conducted at random points in time, thus not differentiating whether or not patients were having a migraine attack at the moment they filled out the survey [42]. The 47.3% participants with level 1 for all five dimensions (ceiling effect) may indicate that the EQ-5D poorly discriminates within patients with migraine. To our knowledge, only two studies validated the use of HIT-6 in German patients with chronic migraine. Although the study by Rendas-Baum et al. [43] included German patients, the authors could not carry out country-specific assessments because of an insufficient sample size of the four European countries included. Thus,

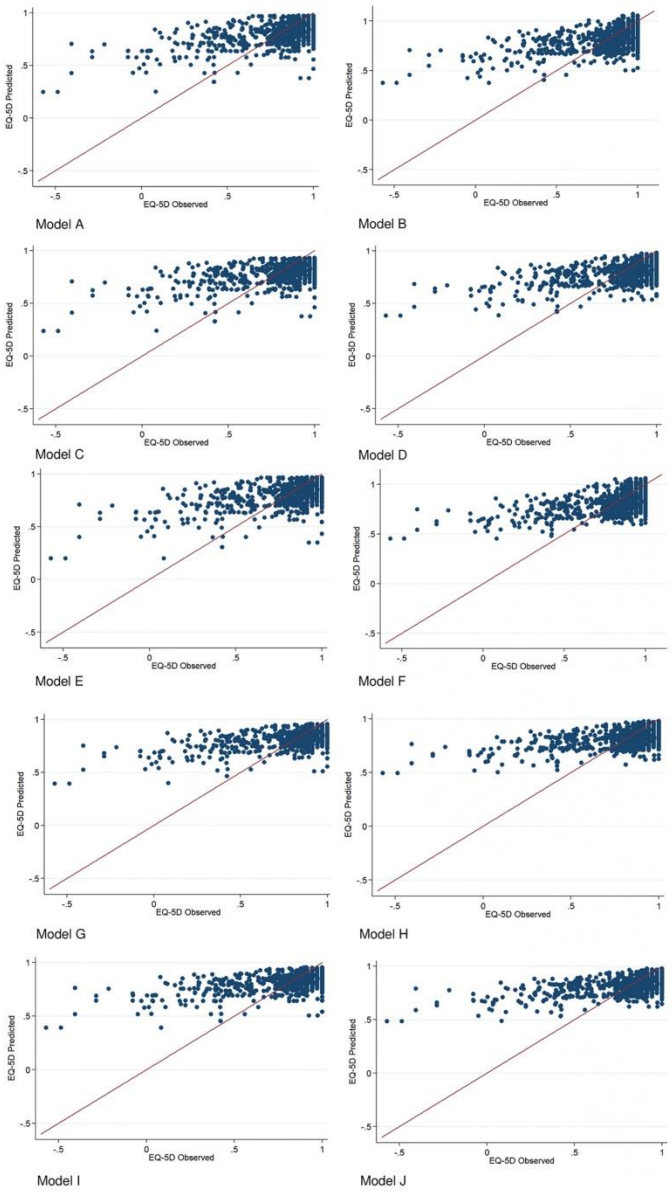


Fig. 3 Scatter plots comparing observed vs predicted EQ-5D-5L utility values. Legend—Model **A**: Mixed-effects linear regression, total HIT-6 score. Model **B**: Mixed-effects linear regression, individual HIT-6 questions. Model **C**: Mixed-effects Tobit, total HIT-6 score. Model **D**: Mixed-effects Tobit, individual HIT-6 questions. Model **E**: Two-part model, total HIT-6 score. Model **F**: Two-part model, individual HIT-6 questions. Model **G**: Adjusted limited dependent variable mixture, total HIT-6 score. Model **H**: Adjusted limited dependent variable mixture, individual HIT-6 questions. Model **I**: Beta Mixture Model (with inflation), total HIT-6 score. Model **J**: Beta Mixture Model (with inflation), individual HIT-6 questions

they treated the data as one group. Another study by Martin et al. [44] evaluated whether the German version of the HIT-6 is comparable to the United States English HIT-6. Unfortunately, there is no information whether the recruited patients suffered from episodic or chronic migraines. Thus, further research on the validation of HIT-6 in German patients who suffer from episodic and chronic migraine could help explain the lack of conceptual overlap between this questionnaire and the EQ-5D.

Given the lack of responsiveness, as well as the substantial ceiling effect of the EQ-5D for migraine patients, economic evaluations with these patients should consider other approaches to determine value, not necessarily QALYs obtained from generic utility-based instruments. In fact, the guidelines of the International Headache Society state that QALYs may fail to account for specific patient preferences due to the insensitive nature of utility instruments [8]. Thus, the use of QALYs may not be appropriate, even where utility values were collected in the study and no mapping algorithm has to be used. Using clinical effectiveness endpoints (such as monthly migraine days) to conduct cost-effectiveness analyses may thus be more suitable for economic evaluations for migraine. However, these analyses with disease-specific outcomes would pose a different problem, as they do not allow decision-makers to compare resource allocation across different conditions.

Strengths of our study include the fact that trained migraine neurologists provided the migraine diagnosis to the study participants' and the low percentage of missing data. Furthermore, we could use multiple observations per person and evaluated this data with methods suitable for repeated measurements where possible. The conceptual overlap of EQ-5D and HIT-6 was evaluated carefully prior to investigating mapping algorithms, where the latter were carried out with a broad set of multivariable modelling approaches.

A limitation of our study is that no external validation could be carried out since no dataset containing both EQ-5D answers and HIT-6 was available. Randomised controlled trials are often considered the 'gold standard' for evidence-based medicine [45], and although they have several strengths in comparison to other designs, their estimates may lack generalisability with respect to different settings [46]. The ISPOR Task Force Report on Mapping mentions that such trials frequently include less diverse patients than observational studies, due to their inclusion criteria, as well as their limited follow-up [12]. Thus, we have compared some socio-demographic characteristics of our study population to those of migraine patients from a study from the German Migraine and Headache Society, which included 7417 adults from three regions in Germany (see Supplementary Table A.7) [47]. The mean ages reported for episodic migraine were 47.5 (Dortmund Health Study), 50.0 (KORA Augsburg Study), and 50.1 (SHIP Study). For episodic migraine (excluding medication overuse headache, an exclusion criterion of our study) age values were 60.8 in the KORA Augsburg Study and 61.0 in the SHIP Study (no values were available for the Dortmund Health Study). In our study, the mean age was somewhat younger at 40.1 for chronic migraine and 41.5 for episodic migraine, which can be explained by the fact that participants need to have some affinity for using apps and because Berlin is the federal state with the second lowest average age [48]. In terms of sex distribution in the episodic migraine population, the Dortmund Health Study reported 78.7% women, the KORA study 84.2%, and the SHIP 85.6%. The proportion of women in our study was comparable with 86% of participants with episodic migraine. It should be also highlighted that many of those suffering from migraine never seek professional care, such that their characteristics may not be reported in the literature. In Germany, only about two thirds of those suffering from migraine consult a physician to receive treatment [49]. Response mapping models were not conducted, since this method requires many observations in each response category and this dataset contained few responses in the worst levels [50]. The EFA was conducted without taking repeated measurements into account. However, in the sensitivity analysis with baseline data only, we obtained the same results, in terms of number of factors and meaningful loadings. As in other mapping studies, compared to observed EQ-5D, mapped EQ-5D values underestimate scores for those with 'perfect' health and overestimate scores for those with worse health states [51]. We ran mixed-effects models with random intercepts only (i.e. different intercepts for each cluster), hence assuming that the association between the independent and dependent variables is highly similar across clusters. Unfortunately, it is not possible to introduce random effects in the adjusted limited dependent variable mixture models.

Conclusion

Our results suggest that the German versions of EQ-5D and the HIT-6 are not measuring the same underlying concepts due to conceptual differences. Therefore, mapping algorithms shall only be used as a last resort for estimating utilities to be employed in cost-utility analyses.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10198-021-01342-9.

Funding Open Access funding enabled and organized by Projekt DEAL. No direct funding was received to write this manuscript. Data for this study steams from SMARTGEM, a randomised-controlled trial financially supported by the German Innovation Committee for the promotion of new forms of care (01NVF17038).

Availability of data and material (data transparency) The datasets generated and/or analysed during the current study are not publicly available due to data protection reasons.

Code availability (software application or custom code) The code used to conduct the current study is available from the corresponding author on reasonable request.

Declarations

Conflicts of interest (include appropriate disclosures) Both Ana Sofia Oliveira Gonçalves and Lars Neeb receive financial support from SMARTGEM, but not for the specific publication of this manuscript. Outside of the submitted work: TK received honoraria from Eli Lilly, Newsenselab, TotalEnergies, Teva, and The BMJ.

Ethics approval (include appropriate approvals or waivers) The local ethics review board at the Charité – Universitätsmedizin Berlin approved the protocol for this study (approval number: EA4/110/18).

Consent to participate and for publication The manuscript does not contain individual person's data in any form.

References

1. Radtke, A., Neuhauser, H.: Prevalence and burden of headache and migraine in Germany. Headache **49**, 79–89 (2009)
2. Buse, D.C., Manack, A., Serrano, D., et al.: Sociodemographic and comorbidity profiles of chronic migraine and episodic migraine sufferers. J. Neurol. Neurosurg. Psychiatry **81**, 428–432 (2010)
3. Blumenfeld, A.M., Varon, S.F., Wilcox, T.K., et al.: Disability, HRQoL and resource use among chronic and episodic migraineurs: results from the International Burden of Migraine Study (IBMS). Cephalalgia **31**, 301–315 (2011)
4. Bloudek, L.M., Stokes, M., Buse, D.C., et al.: Cost of healthcare for patients with migraine in five European countries: results from the International Burden of Migraine Study (IBMS). J. Headache Pain **13**, 361–378 (2012)
5. Stewart, W.F., Wood, G.C., Razzaghi, H., et al.: Work impact of migraine headaches. J. Occup. Environ. Med. **50**, 736–745 (2008)
6. Drummond, M.F., Sculpher, M.J., Claxton, K., et al.: Methods for the economic evaluation of health care programmes. Oxford University Press, Oxford (2015)
7. EuroQol Research Foundation. EQ-5D-5L User Guide, 2019. Available from: https://euroqol.org/publications/user-guides.
8. Diener, H.C., Ashina, M., Durand-Zaleski, I., et al.: Health technology assessment for the acute and preventive treatment of migraine: a position statement of the International Headache Society. Cephalalgia **41**, 279–293 (2021)
9. Diener, H.-C., Bussone, G., Van Oene, J.C., et al.: Topiramate reduces headache days in chronic migraine: a randomized, double-blind, placebo-controlled study. Cephalalgia **27**, 814–823 (2007)
10. Lipton, R.B., Rosen, N.L., Ailani, J., et al.: OnabotulinumtoxinA improves quality of life and reduces impact of chronic migraine over one year of treatment: pooled results from the PREEMPT randomized clinical trial program. Cephalalgia **36**, 899–908 (2016)
11. Kosinski, M., Bayliss, M.S., Bjorner, J.B., et al.: A six-item short-form survey for measuring headache impact: the HIT-6. Qual. Life Res. **12**, 963–974 (2003)
12. Wailoo, A.J., Hernandez-Alava, M., Manca, A., et al.: Mapping to estimate health-state utility from non-preference-based outcome measures: an ISPOR good practices for outcomes research task force report. Value Health **20**, 18–27 (2017)
13. Gillard, P.J., Devine, B., Varon, S.F., et al.: Mapping from disease-specific measures to health-state utility values in individuals with migraine. Value Health **15**, 485–494 (2012)
14. Gerlinger, C., Bamber, L., Leverkus, F., et al.: Comparing the EQ-5D-5L utility index based on value sets of different countries: impact on the interpretation of clinical study results. BMC Res. Notes **12**, 18–18 (2019)
15. Lamu, A.N., Chen, G., Gamst-Klaussen, T., Olsen, J.A.: Do country-specific preference weights matter in the choice of mapping algorithms? The case of mapping the Diabetes-39 onto eight country-specific EQ-5D-5L value sets. Qual. Life Res. **27**, 1801–1814 (2018)
16. Connelly, M., Rapoff, M.A., Thompson, N., Connelly, W.: Headstrong: a pilot study of a CD-ROM intervention for recurrent pediatric headache. J. Pediatr. Psychol. **31**, 737–747 (2005)
17. Sorbi, M.J., Kleiboer, A.M., van Silfhout, H.G., et al.: Medium-term effectiveness of online behavioral training in migraine self-management: a randomized trial controlled over 10 months. Cephalalgia **35**, 608–618 (2014)
18. Hedborg, K., Muhr, C.: Multimodal behavioral treatment of migraine: an internet-administered, randomized, controlled trial. Ups. J. Med. Sci. **116**, 169–186 (2011)
19. Devineni, T., Blanchard, E.B.: A randomized controlled trial of an internet-based treatment for chronic headache. Behav. Res. Ther. **43**, 277–292 (2005)
20. Kuhn, M., Johnson, K.: Over-fitting and model tuning. In: Applied predictive modeling, pp. 61–92. Springer, New York (2013)
21. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Academic Press (2013)
22. Tabachnick, B.G., Fidell, L.S.: Using multivariate statistics, 4th edn. Allyn and Bacon, Boston (2001)
23. Holgado-Tello, F., Moscoso, S., Barbero-García, I., Vila, E.: Polychoric versus pearson correlations in exploratory and confirmatory factor analysis with ordinal variables. Qual. Quant. **44**, 153–166 (2010)
24. Yang, Y., Xia, Y.: On the number of factors to retain in exploratory factor analysis for ordered categorical data. Behav. Res. Methods **47**, 756–772 (2015)
25. Glorfeld, L.W.: An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain. Educ. Psychol. Measure. **55**, 377–393 (1995)
26. Li, C.-H.: Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behav. Res. Methods **48**, 936–949 (2016)
27. Young, T.A., Mukuria, C., Rowen, D., et al.: Mapping functions in health-related quality of life: mapping from two cancer-specific health-related quality-of-life instruments to EQ-5D-3L. Med. Decis. Making **35**, 912–926 (2015)

28. Alava, M.H., Wailoo, A.: Fitting adjusted limited dependent variable mixture models to EQ-5D. Stat. J. **15**, 737–750 (2015)
29. Kelman, L.: Migraine changes with age: IMPACT on migraine classification. Headache **46**, 1161–1171 (2006)
30. Peterlin, B.L., Gupta, S., Ward, T.N., Macgregor, A.: Sex matters: evaluating sex and gender in migraine and headache research. Headache **51**, 839–842 (2011)
31. Pistoia, F., Sacco, S.: Migraine and use of combined hormonal contraception. In: Maassen van den Brink, A., MacGregor, E.A. (eds.) Gender and migraine, pp. 69–79. Springer International Publishing, Cham (2019)
32. Andreou, A.P., Edvinsson, L.: Mechanisms of migraine as a chronic evolutive condition. J. Headache Pain **20**, 117 (2019)
33. R Core Team: R: A language and environment for statistical computing. R Found Stat Comput, Vienna (2017)
34. StataCorp, L.L.C.: Stata statistical software: release 15 (2017). StataCorp LP, College Station (2017)
35. Wickham, H., Francois, R., Henry, L., & Müller, K.: dplyr: A grammar of data manipulation. R package version 0.4, 3, p156 (2015)
36. Wickham, H.: ggplot2: Elegant graphics for data analysis. Springer, Berlin (2016)
37. Bakdash, J.Z., Marusich, L.R.: Repeated measures correlation. Front. Psychol. **8**, 456 (2017)
38. Fox, J.: polycor: Polychoric and Polyserial Correlations. R package version 0.7–10. https://CRAN.R-project.org/package=polycor (2019)
39. Revelle, W. R.: psych: Procedures for Personality and Psychological Research. Software (2017)
40. Gluzmann, P., Panigo, D.: GSREG: Stata module to perform Global Search Regression. https://EconPapers.repec.org/RePEc:boc:bocode:s457737 (2013)
41. Brazier, J.E., Yang, Y., Tsuchiya, A., et al.: A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur. J. Health Econ. **11**, 215–225 (2010)
42. Xu, R., Insinga, R.P., Golden, W., Hu, X.H.: EuroQol (EQ-5D) health utility scores for patients with migraine. Qual. Life Res. **20**, 601–608 (2011)
43. Rendas-Baum, R., Yang, M., Varon, S.F., et al.: Validation of the Headache Impact Test (HIT-6) in patients with chronic migraine. Health Qual. Life Outcomes **12**, 117 (2014)
44. Martin, M., Blaisdell, B., Kwong, J.W., Bjorner, J.B.: The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. J. Clin. Epidemiol. **57**, 1271–1278 (2004)
45. Jones, D.S., Podolsky, S.H.: The history and fate of the gold standard. Lancet **385**, 1502–1503 (2015)
46. Sculpher, M.J., Claxton, K., Drummond, M., McCabe, C.: Whither trial-based economic evaluation for health care decision making? Health Econ. **15**, 677–687 (2006)
47. Straube, A., Pfaffenrath, V., Ladwig, K.-H., et al.: Prevalence of chronic migraine and medication overuse headache in Germany—the German DMKG headache study. Cephalalgia **30**, 207–213 (2010)
48. Statista Durchschnittsalter der Bevölkerung in Deutschland nach Bundesländern* im Jahr 2018. https://de.statista.com/statistik/daten/studie/1093993/umfrage/durchschnittsalter-der-bevoelkerung-in-deutschland-nach-bundeslaendern/. Accessed 7 Mar 2020
49. Schmerzklinik Kiel (2009) Wer leidet? https://schmerzklinik.de/service-fuer-patienten/migraene-wissen/wer-leidet/. Accessed 13 Jul 2020
50. Gray, L.A., Hernández Alava, M., Wailoo, A.J.: Development of methods for the mapping of utilities using mixture models: mapping the AQLQ-S to the EQ-5D-5L and the HUI3 in patients with asthma. Value Health **21**, 748–757 (2018)
51. Oppe, M., Devlin, N., Black, N.: Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. Value Health **14**, 884–891 (2011)

11.2 Research Project 2 – Systematic review of mapping algorithms

Oliveira Gonçalves AS, Werdin S, Kurth T, Panteli D. Mapping Studies to Estimate Health-State Utilities From Nonpreference-Based Outcome Measures: A Systematic Review on How Multiple observations are Taken Into Account. Value Health. 2023 Apr 1;26(4):589–97.

https://doi.org/10.1016/j.jval.2022.09.2477

11.3 Research Project 3 – Health economic evaluation

RESEARCH ARTICLE

Economic Evaluation of a Mobile Stroke Unit Service in Germany

Ana Sofia Oliveira Gonçalves, MSc,[1] Jessica L. Rohmann, PhD,[1,2] Marco Piccininni, PhD,[1,2] Tobias Kurth, MD, ScD,[1] Martin Ebinger, MD,[2,3] Matthias Endres, MD,[2,4,5,6,7,8] Erik Freitag, MD,[4] Peter Harmel, MD,[4] Irina Lorenz-Meyer, MSc,[4] Ira Rohrpasser-Napierkowski, Dr. rer. nat.,[4] Reinhard Busse, MD,[9,10] and Heinrich J. Audebert, MD[2,4]

Background: Lower global disability and higher quality of life among ischemic stroke patients was found to be associated with the dispatch of mobile stroke units (MSUs) among patients eligible for recanalizing treatments in the Berlin_Prehospital Or Usual Delivery of stroke care (B_PROUD) study. The current study assessed the cost-utility and cost-effectiveness of additional MSU dispatch using data from this prospective, controlled, intervention study.
Methods: Outcomes considered in the economic evaluation included quality-adjusted life years (QALYs) derived from the 3-level version of EQ-5D (EQ-5D-3L) and modified Rankin Scale (mRS) scores for functional outcomes 3-months after stroke. Costs were prospectively collected during the study by the MSU provider (Berlin Fire Brigade) and the B_PROUD research team. We focus our results on the societal perspective. As we aimed to determine the economic consequences of the intervention beyond the study's follow-up period, both care costs and QALYs were extrapolated over 5 years.
Results: The additional MSU dispatch resulted in an incremental €40,984 per QALY. The best-case scenario and the worst-case scenario yielded additional costs of, respectively, €24,470.76 and €61,690.88 per QALY. In the cost-effectiveness analysis, MSU dispatch resulted in incremental costs of €81,491 per survival without disability. The best-case scenario and the worst-case scenario yielded additional costs of, respectively, €44,455.30 and €116,491.15 per survival without disability.
Interpretation: Among patients eligible for recanalizing treatments in ischemic stroke, MSU dispatch was associated with both higher QALYs and higher costs and is cost-effective when considering internationally accepted thresholds ranging from an additional €40,000 to €80,000 per QALY.

ANN NEUROL 2023;00:1–10

Stroke is the second leading cause of death, and the third cause of disability-adjusted life-years (DALYs) worldwide.[1] Intravenous thrombolysis treatment with recombinant tissue plasminogen activator (tPA) is the only approved pharmacologic treatment in acute ischemic stroke and recommended for up to 4–5 hours after

View this article online at wileyonlinelibrary.com. DOI: 10.1002/ana.26602

Received Sep 9, 2022, and in revised form Jan 6, 2023. Accepted for publication Jan 9, 2023.

Address correspondence to Ms Gonçalves, Institute of Public Health, Charité - Universitätsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany. E-mail: ana.goncalves@charite.de

Reinhard Busse and Heinrich J. Audebert Contributed equally.

From the [1]Institute of Public Health, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin und Humboldt Universität zu Berlin, Berlin, Germany; [2]Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin und Humboldt Universität zu Berlin, Berlin, Germany; [3]Klinik für Neurologie, Medical Park Berlin Humboldtmühle, Berlin, Germany; [4]Klinik und Hochschulambulanz für Neurologie, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin und Humboldt Universität zu Berlin, Berlin, Germany; [5]Berlin Institute of Health (BIH), Berlin, Germany; [6]German Centre for Cardiovascular Research (DZHK), partner site Berlin, Berlin, Germany; [7]NeuroCure Cluster of Excellence, Berlin, Germany; [8]German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany; [9]Department of Health Care Management, Berlin University of Technology, Berlin, Germany; and [10]European Observatory on Health Systems and Policies, Brussels, Belgium

Additional supporting information can be found in the online version of this article.

symptom onset when no contraindications are present.[2] Studies have shown that earlier treatment initiation results in better outcomes.[2,3] From symptom onset, time-to-treatment can be reduced by deploying "mobile stroke units" (MSUs) equipped with computed tomography scanners and trained personnel, enabling acute stroke work-up and administration of thrombolytic treatment prior to hospital arrival.[4,5]

For healthcare systems to justify reimbursement of MSU deployment, MSU dispatch must be clinically effective and demonstrate cost-effectiveness. The cost-utility analysis in the PHANTOM-S study (Berlin, Germany, 2011–2013) estimated an incremental cost-effectiveness ratio (ICER) of €32,456/quality-adjusted life-year (QALY; $32,306.70; EUR 1 = US$ 0.9954) based on shorter onset-to-treatment times and increased thrombolysis rates due to MSU deployment.[6] This estimate fell below the cost-utility threshold for the UK (£30,000/QALY),[7] the Netherlands (up to €80,000/QALY),[8] and the United States (a study proposed a threshold of either $100,000/QALY or $150,000/QALY).[9] A 2020 Australian study by Kim et al estimated that MSUs cost an additional AU$30,982 per additional DALY averted in comparison with standard care.[10] A recent Norwegian study analyzed the economic consequences of the dispatch of MSUs and reported an ICER between $38,660/QALY and $113,700/QALY, depending on the number of potentially treated patients.[11] None of these studies used prospectively collected data on quality of life (QoL) and long-term care costs. Hence, we are confident that our study adds a significant novelty to previously published MSU cost-effectiveness studies.[12]

Recent results from the Berlin-based B_PROUD study showed that the dispatch of MSUs, compared with conventional ambulances alone, was associated with significantly better functional outcomes and lower global disability 3 months after stroke among ischemic stroke patients with no contraindications to recanalizing treatments.[4] We now aimed to assess cost-utility, cost-effectiveness, and cost-benefit of the additional MSU dispatch compared with standard care for eligible acute ischemic stroke patients in Berlin, Germany.

Methods

Setting, Patients, and Study Design

We used data from B_PROUD, a prospective, non-randomized, controlled intervention study conducted in Berlin, Germany. Study inclusion commenced on February 1, 2017, until May 8, 2019, and follow-up data collection ended on October 30, 2019.

Three MSUs were consecutively rolled-out across Berlin over 17 months.[4] Exposure status was determined according to ambulance dispatch type (with/without MSU dispatch) in each MSU's catchment area, analogous to the intention-to-treat principle in a randomized controlled trial. Group allocation was determined by MSU availability at the time of dispatch; creating a natural experiment setting. In overlapping catchment areas, the geographically closest available MSU was sent to the event location. MSUs in our setting are staffed with a paramedic, a radiology technician with emergency training, and a neurologist with training in emergency medicine. Conventional ambulances are staffed with a paramedic and an ambulance technician, who are trained to stabilize and transport patients to hospitals.

Patients were included in the B_PROUD study if they had a final diagnosis of acute cerebral ischemia (ischemic stroke or transient ischemic attack, according to the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision) and were likely to be eligible for thrombolysis or thrombectomy.[4] Other criteria for enrollment in the trial included being at least 18 years old; the emergency call had to prompt an MSU dispatch code during the MSU's operating hours; stroke onset within the previous 4 hours; patients had to be within the catchment area of an MSU; patients had to have been ambulatory before their stroke (a rough proxy of modified Rankin Scale scores [mRS]). Other stroke subtypes were excluded from the primary study population. Patients with final non-stroke diagnoses or stroke patients not eligible for recanalizing treatments were not included because no differences in short term outcomes were found in the preceding study.[6] We therefore assumed that the dispatch of MSUs had no effect on outcomes for these patient groups. The intervention group consisted of ischemic stroke and TIA patients for whom an MSU and a conventional ambulance were simultaneously dispatched (n = 749). The comparator group consisted of those for whom only a conventional ambulance was dispatched (n = 794).[4] Further details on the study design, and study population characteristics were published elsewhere.[4]

Outcomes

For the cost-utility analyses, the outcome of health-related QoL was measured with the 3-level version of EQ-5D (EQ-5D-3L) 3 months after the index event, using the German Visual Analogue Scale value set.[13] This questionnaire comprises five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression—and was assessed by trained study nurses using telephone interviews or paper questionnaires. A previous study

investigating QoL changes according to stroke symptom severity found that the differences in EQ-5D-3L scores and survival probabilities after the initial months remained relatively constant between stroke severity categories over a 5-year period.[14] Thus, we considered a 5-year time horizon in our analyses. This time horizon allowed us to take into account care costs, for which the financial burden extends beyond the 3-month follow-up (see the Costs section). While QoL in the Luengo-Fernandez study was not assessed at the 3-month follow up, at which times QoL information was collected in B_PROUD, it falls within their 1- and 6-month follow-up interval.[14] QALY scores for 5 years were calculated by multiplying the EQ-5D-3L score at 3 months by 5. We recognize that this extrapolation implies that the EQ-5D-3L scores for each patient remains constant over a 5-year period. However, this approach was suitable for the present study because the ICER denominator corresponds to the difference in QALYs between groups after confounding adjustment. Therefore, the relevant assumption in our approach was that the variation in the amount of QALYs due to death or to QoL changes after confounding adjustment was the same in both groups across the 5-year period.

The primary outcome measure in the cost-effectiveness analyses was the 3-month mRS, a measure of the degree of disability or dependence in the daily activities among people who have suffered a stroke (range 0, no neurological symptoms, to 6, death). In the B_PROUD study, patients in the MSU dispatch group had significantly less global disability as measured by mRS score compared with those in the standard care dispatch group (odds ratio [OR] 0.71; 95% confidence interval [CI], 0.58 to 0.86).[4] For the cost-effectiveness analysis, this score was dichotomized in two levels: 0–1 ("excellent" or/"survival without disability") and 2–6 ("not excellent"/"survival with disability" or "dead").[15] Determination of 3-month mRS scores was performed as the median of three independent neurologists rating outcomes in a blinded manner whenever possible.[4] For the dichotomized mRS outcome, we assumed that the difference in the number of survivals without disability between groups, after confounding adjustment, remained constant over the 5-year period.

Costs

Costs were prospectively documented throughout the study period by the MSU provider (Berlin Fire Brigade) and by the B_PROUD study personnel. These included MSU-related investment and running costs, prehospital medication costs, and prehospital imaging costs. For the MSU investment costs, we obtained the annualized cost (per 12 months) of an MSU and multiplied it by the running duration of all MSUs (56 months of operation), as shown in Table S3.3.

Our analyses present cost from two different perspectives: the societal perspective and the statutory health insurance perspective. Taking the societal perspective, we considered all additional MSU-related costs incurred by the Berlin Fire Brigade (including value-added tax). Taking the statutory health insurance perspective, we used the fee calculated per MSU dispatch by the Berlin Fire Brigade for reimbursement by the statutory health insurances (assuming that 97% of all dispatches with patient care were actually billable and reimbursed). As analyses from the statutory health insurance are specific to Germany, we present them in the Supplementary Material and focus on the societal perspective in the manuscript.

We also accounted for projected costs associated with long-term (nursing) care in the 5-year period following the index event. To denote the extent of the need for care, we used unpublished information from another large German stroke outcome study[16] and converted mRS values into different levels of care dependency called "Pflegestufen," used in Germany until 2017. In Germany, the levels of care dependency are determined by the Medical Service of the Health Insurance (MDK) and are used to calculate long-term care insurance benefits for care provision. Since Germany changed the way of categorizing these levels in 2017, we converted the original "Pflegestufe" levels to the contemporary "Pflegegrad" (care grade) according to official, published conversion tables.[17] Synthesizing information from 3-month follow-up on individual's care-grade and information on each patient's living situation (at home being cared for by a relative, at home being cared for by a professional, in a nursing care home, or in hospital), we calculated total care service costs.

We applied a 3% discount rate to long-term (nursing) care costs for the years following the intervention, as recommended by the German Institute for Quality and Efficiency in Health Care.[18] Furthermore, we used the 3% rate to convert 2020 costs (which we received from the named data sources) to 2019, which is the base year for our analyses. We report all costs in Euros.

We considered a base-case, a best-case, and a worst-case scenario (detailed in Supplementary Appendix 1). Under the base-case scenario, we analyzed outcomes in the primary study population.[4] We therefore accounted for additional costs incurred by MSU operation for all code stroke patients. Furthermore, we assumed that medication costs for individual tPA treatment were the same in the pre-hospital and in-hospital settings and that costs for nursing care depended on the mRS score. A best-case scenario was created following the base-case scenario, but

calculating long-term care costs with a less conservative conversion from 3-month disability to nursing costs was used (Supplementary Appendix 2). We further assumed that the frequency of MSU dispatches can be increased by a factor of 1.8, corresponding to the higher tPA treatment frequency per MSU operation week, as seen during the PHANTOM-S study period.[6] The worst-case scenario used the base-case scenario's assumptions with several modifications. As we assumed that the effects observed in B_PROUD only applied to patients with full 3-month follow-up information, we restricted the analysis to complete cases only instead of using multiple imputation under this scenario. Moreover, we assumed that effects on mRS (and levels of care dependency) only remained stable over 18 months (according to the observation period in IST-3).[19] Finally, we assumed that half of the imaging examinations would have to be repeated in-hospital.

Statistical Methods

We used the parametric G-formula to estimate incremental costs and incremental outcomes due to MSU dispatch for all analyses.[20] We fitted a linear regression model with costs as the dependent variable, and MSU dispatch, and a set of covariates that have been selected in the clinical effectiveness study (age, sex, arterial hypertension, diabetes mellitus, atrial fibrillation, neurological symptoms at emergency medical services or MSU arrival, and living situation before stroke) as independent variables.[4] We then used this model to predict the expected costs for each individual for both hypothetical scenarios in which the MSU was dispatched in addition to a conventional ambulance to all code-stroke patients versus dispatch of a conventional ambulance alone for all code-stroke patients. In order to obtain overall costs, we summed the predicted costs across all individuals in each scenario and computed the difference between the overall costs to obtain incremental costs related to MSU dispatch. We used the same procedure to compute incremental QALYs due to MSU dispatch in addition to a conventional ambulance, since QALYs is also a numeric variable.

For the dichotomized mRS, we ran a logistic regression model to predict each individual's probability of surviving without disability (mRS 0–1) under each scenario and to estimate the absolute number of individuals who survived without disability per scenario. Finally, we computed the incremental number of survivors without disability due to MSU dispatch as the difference.

Incremental overall costs and incremental overall effects (QALYs in the cost-utility analysis and dichotomized mRS in the cost-effectiveness analysis) were used to estimate ICERs.

As care cost calculations depended on individual patient data, some of which had missing values, there were missing values in the costs and in other outcomes (QALY and dichotomized mRS). Assuming missing at random data, we used multiple imputation by chained equations to impute missing values for the base-case and best-case scenarios. Point estimates for incremental costs, incremental QALYs, and incremental survivals were obtained as averages across the five imputed datasets. In accordance with the guidelines,[21] imputation models included all covariates that were included in the regression models in the primary effectiveness publication (see above).[4]

To compute 95% CIs, we performed nonparametric bootstrapping with 5,000 replications according to the BootMI method.[22] On each bootstrapped dataset, after multiple imputation, the 2.5% and 97.5% percentiles of the resulting distribution of the average metric (across 5 imputed datasets) were considered as confidence limits. We further used cost-utility and cost-effectiveness planes to illustrate the bootstrapped cost-utility and cost-effectiveness pairs resulting from the bootstrapping replication runs, depicting the joint uncertainty surrounding costs and outcomes.[23] Data points falling into the "northeast" quadrant of these figures indicate that the intervention generates more health gains but is also more expensive, while points in the "southwest" quadrant indicate a cost-saving intervention that generates less health gains. An intervention that is associated with higher health gains and lower costs ("southeast" quadrant), is considered to be an economically "dominant" strategy. We also report the percentage of data points that fell into each quadrant.

Furthermore, we conducted a sensitivity analysis, in which the models were only adjusted for MSU coverage, operationalized as the number of MSUs covering the location (zip code) during the specific quarter of the calendar year at the moment of the patient's stroke (due to overlapping of MSU catchment areas). We assumed no relevant change in the MSU availability occurred after the introduction of the GPS system. To create the imputed datasets, we used information from all covariates in the main analyses (see above) plus the MSU coverage variable.

All analyses were conducted using R version 4.0.3 and RStudio.[24]

Standard Protocol Approvals, Registrations, and Patient Consents

The study was approved by the Charité Ethics Committee (EA4/109/15). The economic evaluation was conducted in accordance with the statistical analysis plan.

Results

Patients' baseline parameters, clinical, and process information can be found in Supplementary Appendix 3, Table S3.1.

Costs

Cost breakdown per patient and exposure group under the base-case scenario is shown for the societal perspective, in Table 1. Under this scenario, the largest cost contributors in the MSU group were medical costs/reimbursement of expenses, which included costs for hospital employed personnel (neurologists, technicians, teleradiology, and project management) and amounted to €4,383.08 per included stroke patient. The MSU investment costs were the second-largest cost contributor (€1,286.02 per patient). The total costs (excluding long-term care costs) were estimated to be on average €8,491.58 during the 56-month study period per patient allocated to the MSU group. In the standard care group, the total costs amounted to €1,274.54 per patient.

Outcomes

The average EQ-5D-3L score was higher in the MSU group than in the standard care group (0.63 vs 0.59). The percentage of patients who reported a "good outcome" as measured with the mRS was also higher in the MSU group than in the standard care group (50.92% vs 42.31%) (Table 2).

Cost-Utility Analyses

Results under the base-case scenario are shown in Table 3, which shows that MSU dispatch was associated with both higher costs and QALYs. Incremental overall costs due to MSU dispatch amounted to €10,759,089.49 (9,912,284.42; 11,997,571.30). Incremental QALYs due to MSU dispatch were 262.52 (−41.06; 479.92), and the resulting ICER per QALY was €40,983.82. A large majority (95.16%) of the bootstrapped replications, showed that MSU dispatch was associated with both higher QALYs and higher costs (Figure).

As expected, given our assumptions (Supplementary Appendix 1), the ICERs under the best-case scenario yielded the lowest values (€24,470.76 per QALY), mostly attributable to lower incremental overall costs than in other scenarios. On the opposite, under the worst-case scenario, the ICER for the societal (€61,690.88 per QALY) was higher than in the base-case scenario due to higher incremental overall costs but also because of lower incremental overall QALYs.

Bootstrap replications' patterns were similar under all scenarios and between perspectives, with approximately 95% of the bootstrapped samples showing that MSU dispatch was associated with higher QALYs and higher costs.

Taking the statutory health insurance perspective, the ICER per QALY amounted to €28,029·51, which is lower than the ICER calculated when taking a societal perspective. Detailed results for the analyses taking the statutory health insurance perspective are given in Supplementary Appendix 3, Table S3.4.

Cost-Effectiveness Analyses

Cost-effectiveness analyses, under the base-case scenario, showed that incremental overall costs due to MSU dispatch amounted to €10,793,823.78 (9,809,757.36; 12,020,619.78) and incremental survivals to 132.45 (48.30; 199.85). Thus, an incremental survival without disability was associated with additional costs amounting to €81,491.49 (Table 3).

The best-case scenario and the worst-case scenario yielded, respectively, €44,455.30 and €116,491.15 per survival without disability.

The percentage of bootstrapped samples indicating higher incremental survivals and higher costs associated with MSU dispatch reached almost 100% (Figure).

Results for the analyses taking the statutory health insurance perspective are given in Supplementary Appendix 3, Table S3.4.

Sensitivity Analyses

Sensitivity analyses' results (Supplementary Appendix 3, Tables S3.6 and S3.7) were similar to the findings of the main analyses.

Discussion

For the primary population of the B_PROUD study, incremental costs and incremental QALYs generated an ICER of circa €41,000 per QALY. These values are similar to the results of a previous study conducted in Berlin/Germany, which estimated an ICER of €32,456 per QALY.[6] Overall, our findings support that MSUs can be considered cost-effective for thresholds greater than €40,000 per QALY. The interpretation of this number depends on the acceptable and setting-specific threshold, and the range of possible thresholds varies (eg, between €20,000 and €80,000 in the Netherlands or between $100,000 and $150,000 in the US). While Germany does not have an official threshold, the MSU implementation would be, for example, justified in the Netherlands, which has a similar healthcare system (with a mixed compulsory social insurance and private voluntary insurance) and where this ICER figure could lie in the official threshold range. In determining the cost-effectiveness of a health intervention, the WHO Commissions on Macroeconomics and Health

TABLE 1. Mean Additional Costs Per Patient Included in the B_PROUD Study (€, Unadjusted, Base-Case Scenario, Societal Perspective)

Societal Perspective	Mean Cost per Patient MSU Group (n = 749)	Mean Cost per Patient Standard Care Group (n = 794)
MSU-related costs		
Investment costs	€1,286.02	
Interest	€67.15	
Maintenance and repairs	€698.90	
Fuel	€77.75	
Personnel	€974.35	
Medication	€654.40	€481.11*
Consumables	€76.93	
Other operating costs	Facility expenses: €81.99 Medical attire for MSU personnel: €33.92 Medical costs/reimbursement of expenses (neurologists, technicians, teleradiology and project management): €4,383.08 Personnel pension provisions and allowances: €397.09 Internal costs incurred by the Fire Brigade†: €511.44	
Savings through the assignment of MSU emergency physicians to conventional emergency cars when MSUs were out of service‡	−€371.31	
Savings by avoiding repeated imaging in hospitals§	−€380.13	
Costs for additional emergency physician car deployments¶		€620.35
Costs for additional rescue helicopter deployments		€145.40
Costs for additional auxiliary vehicle deployments		€27.68
Long-term care costs over 5 years	Depends on patient's mRS, living situation and time after index event (Supplementary Appendix 2)	Depends on patient's mRS, living situation and time after index event (Supplementary Appendix 2)

Note: All costs are reported per participant over a 56-month running period except for the nursing care costs, which were projected over the 5-year period after the index event.

*There were 382 patients in the standard care group who received tPA at the hospital, with a cost of €1,000 and there are 794 patients in this group: €1,000 × 382/794.

†Internal costs incurred by the Fire Brigade, including administration, insurance, dispatch, housing, trainings, and routine emergency medical services' data collection (excluding costs for command operations).

‡During B_PROUD, the three MSUs accumulated a total of 535 off-duty working days. During this period, MSU physicians worked on conventional physician staffed ambulances.

§We used internal cost calculation fees of Charité Universtätsmedizin Berlin in order to quantify the in-hospital cost savings by shifting brain imaging to the prehospital setting.

¶MSUs with an emergency physician aboard were dispatched to all patients with code stroke in the coverage area during operational times. In cases when no MSUs were available, Berlin regulations stipulate that a conventional ambulance without an emergency physician on board is to be dispatched to code stroke patients who do not present with compromised vital signs nor reduced vigilance, while a conventional ambulance plus an emergency physician car are dispatched to patients with compromised vital signs or with reduced vigilance. In the latter case, if no emergency physician cars are available, helicopters are sent to the scene.

Abbreviations: MSU = mobile stroke unit; mRS = modified Rankin Scale; tPA = recombinant tissue plasminogen activator.

TABLE 2. Quality of Life Outcomes (Available Case, Per Participant, Unadjusted)

	MSU group (Available Case, per Patient)	Standard Care Group (Available Case, per Patient)
No.	618	649
EQ-5D-3L	0.63	0.59
No.	654	683
Lives saved without disability, mRS 0–1 (%)	50.92% (n = 333)	42.31% (n = 289)

Abbreviations: mRS = modified Rankin Scale; MSU = mobile stroke unit.

suggests a threshold of one to three times the GDP per capita.[25] For Germany, this threshold would amount from $58,386 to $175,158. Our ICER result would also be considered cost-effective using these reference values.[26]

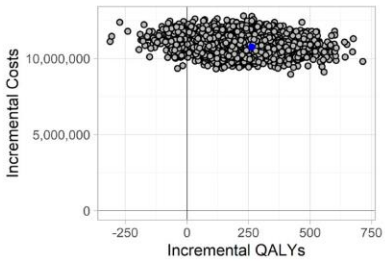
Previous studies (1) relied on extrapolations of IVT treatment effects in the rather crude 60 or 90 minutes onset to treatment time intervals, (2) projected disability scores to estimate QoL and/or (3) used probabilistic models to estimate the incremental costs per QALY. The present analysis makes use of prospectively measured outcomes and is therefore more robust compared with previous evaluations.

One challenge for economic evaluations based on trials is how they can be applied to other settings. On the cost side, we have shown cost data whenever possible, with a clear indication of unitary costs and the amount of resource consumption. This allows decision-makers in other healthcare systems to apply their own prices to the

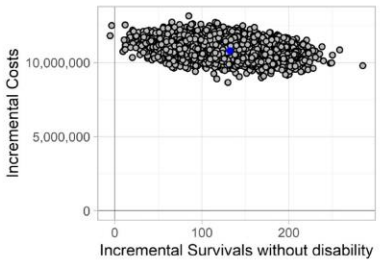
TABLE 3. Cost-Utility and Cost-Effectiveness Analyses for the Societal Perspective

	Incremental Overall Cost in € (Bootstrapped 95% CI) for 1,543 Patients*	Incremental Overall QALY or Survival Without Disability (Bootstrapped 95% CI) for 1,543 Patients*	ICER per QALY or per Incremental Survivals Without Disability in €	NE %	NW %
Cost-utility analysis					
Base-case scenario	10,759,089.49 (9,912,284.42; 11,997,571.30)	262.52 (−41.06; 479.92)	40,983.82	95.16	4.84
Best-case scenario	5,839,431.83 (4,351,335.96; 7,553,436.36)	238.63 (−35.09; 484.01)	24,470.76	95.36	4.64
Worst-case scenario	13,129,271.81 (12,072,370.83; 14,205,721.97)	212.82 (−46.27; 471.22)	61,690.88	94.7	5.3
Cost-effectiveness analysis					
Base-case scenario	10,793,823.78 (9,809,757.36; 12,020,619.78)	132.45 (48.30; 199.85)	81,491.49	99.96	0.04
Best-case scenario	5,912,181.64 (4,261,885.24; 7,592,776.25)	132.99 (57.46; 208.94)	44,455.30	99.98	0.02
Worst-case scenario	13,129,271.81 (12,072,370.83; 14,205,721.97)	112.71 (37.56; 188.15)	116,491.15	99.86	0.14

Abbreviations; CI = confidence interval; ICER = incremental cost-effectiveness ratio; QALY = quality-adjusted life-years; NE = Northeast Quadrant; NW = Northwest Quadrant.
*The figure "1,543 patients" corresponds to the pseudo-population created when using the G-formula method[18] (see the Methods section).



A Cost-utility analysis (base-case scenario, societal perspective)



B Cost-effectiveness analysis (base-case scenario, societal perspective)

FIGURE: Cost-utility and cost-effectiveness planes.

same units of resource use. Furthermore, we focused our analyses on the societal perspective, which is less country-specific than statutory-health insurance (reported in the Supplementary Appendix only). The completeness of MSU coverage in Berlin makes it generalizable for the whole city. All hospitals with a stroke unit in Berlin participated in the data collection. With regard to catchment areas, once all MSUs were in operation, they covered more than 94% of the Berlin population (according to the Berlin Fire Brigade).[4]

Several limitations to our study should be considered when interpreting our results. First, one challenge for economic evaluations is their generalizability to other settings.[27,28] In terms of costs, we reported results from a societal and a statutory health insurance perspective. In the Berlin setting, emergency physician cars are dispatched to the scene in addition to conventional ambulances in selected cases with unstable vital parameters or loss of consciousness. We recognize that in other settings, ambulances may only be paramedic-staffed. Thus, related costs in the conventional care group may be higher in our case. On the other hand, several MSUs are operated using telemedicine in order to avoid neurologist staffing of the MSU. This is likely to largely reduce the costs of MSU care, but there may be other disadvantages.

We made some simplifying assumptions due to limitations in data availability. First, we did not account for possible in-hospital savings related to MSU dispatch due to shorter lengths of stay in costly high care monitoring units or rehabilitation hospitals as a consequence of improved functional outcomes (Supplementary Appendix 3, Table S3.8). Furthermore, we did not account for reduced resource consumption regarding consumables nor for potential benefits in terms of time savings in emergency departments created by pre-hospital stroke work-up and medical treatment. Second, although the average age of patients in this study was 73.5 years, a total of 385 included patients were less than 65 years old, which is the current standard retirement age in Germany. Our computations are therefore conservative, as they do not reflect potential savings in terms of productivity gains when taking a societal perspective.

Third, under the base-case scenario, we used the most conservative approach in converting costs from level-of-care to grade-of-care (Supplementary Appendix 1). A substantial proportion of patients would likely have received a larger amount of care benefits because mental and communication deficits mean higher care benefits in the currently used "grade-of-care" system, but not in the old "level-of-care" system used for the conversion in this analysis. With less global disability in the MSU dispatch arm, this would have resulted in a higher cost load in the conventional dispatch group. Fourth, relating again to the base-case scenario, we did not account for possible transport savings, which are likely to have been incurred since stroke work-up on the dispatch of MSUs improves the direct transport of patients to the most appropriate hospital.[29] Hence, secondary transports of patients for specific treatments such as thrombectomy and the dispatch of costly emergency physician cars, which is commonplace in the German setting, were not included. Fifth, we did not account for benefits in terms of time savings in emergency departments created by pre-hospital stroke work-up and medical treatment.

Furthermore, due to the B_PROUD eligibility criteria, we only considered the primary study population without absolute contraindications to recanalizing treatments, which does not capture effects in other subgroups with MSU dispatch such as intracranial hemorrhage patients or patients with final diagnoses of non-stroke diseases. We have, however, accounted for costs incurred by these patients. As the information about EQ-5D-3L was collected 3 months after the index event, our use of a 5-year time horizon relies on extrapolation beyond the collected data with the theoretical assumption that QoL remains constant and without additional mortality over the 5-year period. Our interest lies in comparing the average difference in QALYs between the MSU and non-MSU groups after confounding adjustment. For our

estimate to be accurate, it is sufficient that any possible violations of this assumption lead to a change of the same amount of QALY in both groups across the 5-year period. Luengo-Fernandez et al have shown a relative stability in survival probability and QoL (as measured by the EQ-5D-3L) during the 5-year period after stroke in different categories of stroke severity.[14]

In summary, our findings indicate that the dispatch of MSU to patients with suspected acute stroke is cost-effective compared with conventional care only, considering internationally accepted thresholds higher than additional €40,000 per QALY. These findings can inform decision-makers planning future pre-hospital stroke care in metropolitan areas. Similar to the results of previous health economics analyses of MSUs, the present study indicates that their cost-effectiveness depends on the volume of ischemic stroke patients who can be treated with intravenous thrombolysis in these vehicles. It is therefore crucial to optimize the identification of acute stroke patients on a dispatcher level and to streamline the related processes of MSU operations.

Currently, most MSUs in service are financed on a charitable basis or with research grants. However, results from several economic evaluations of MSUs in different jurisdictions have shown that this intervention can be cost-effective. By establishing a transparent decision-making process, policy-makers in local settings may therefore take these results into consideration and act in the interest of patients.

Acknowledgments

The authors thank Simon Dyberg for his assistance in developing and revising the costs' tables and Jakob Beilstein and Fernando Urrutia Gonzalez for their assistance in constructing the MSU coverage variable. We further extend our gratitude to the Berlin Fire Brigade and the Berlin Senatsverwaltung for support in providing the cost data. Finally, we thank Jeanette Fahron, who provided important guidance during the planning phases of the B_PROUD study as the institutional data protection representative.

Data collection in the B_PROUD study was funded by the German Research Foundation, and in the B-SPATIAL registry, by the Federal Ministry of Education and Research via the Center for Stroke Research Berlin. Open Access funding enabled and organized by Projekt DEAL.

Author Contributions

A.S.O.G., J.L.R., R.B., and H.J.A. were responsible for the study conception and design. A.S.O.G., J.L.R., M.P., T.K., E.F., P.H., I.R.N., I.L.M., R.B., and H.J.A. contributed to the acquisition and analysis of data. All authors contributed to drafting the text or preparing the figures.

Potential Conflicts of Interest

H.J.A reported personal fees from Boehringer Ingelheim, the manufacturer of Alteplase, which is used for intravenous thrombolysis in ischemic stroke. He also reported personal fees from NovoNordisk, the manufacturer of NovoSeven (factor VIIa), a drug that is used for acute life-threatening bleedings and with an ongoing study in patients with spontaneous intracerebral hemorrhage. The other authors have nothing to report. Full disclosures are listed in the submitted ICMJE forms.

Data Availability Statement

Pseudonymized participant data that underlie the results reported in this article can be made available in a de-identified manner upon request to researchers who provide a methodologically sound proposal after de-identification beginning at 12 months and ending 5 years after publication. Proposals should be directed to heinrich.audebert@charite.de. The statistical analysis plan for the analyses presented in this manuscript, is available in an online repository. A sample of the code used to analyze the data is available as a Supplementary Appendix 4.

References

1. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. Lancet 2020 Oct 17;396:1204–1222.
2. Lees KR, Bluhmki E, von Kummer R, et al. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. Lancet 2010;375:1695–1703.
3. Saver JL, Goyal M, van der Lugt A, et al. Time to treatment with endovascular Thrombectomy and outcomes from ischemic stroke: A meta-analysis. JAMA 2016;316:1279–1288.
4. Ebinger M, Siegerink B, Kunz A, et al. Association between dispatch of Mobile stroke units and functional outcomes among patients with acute ischemic stroke in Berlin. JAMA 2021;325:454–466.
5. Grotta JC, Yamal JM, Parker SA, et al. Prospective, multicenter, controlled trial of Mobile stroke units. N Engl J Med 2021;385:971–981.
6. Gyrd-Hansen D, Olsen KR, Bollweg K, et al. Cost-effectiveness estimate of prehospital thrombolysis: results of the PHANTOM-S study. Neurology 2015;84:1090–1097.
7. National Institute for Health and Care Excellence. Guide to the Methods of Technology Appraisal. National Institute for Health and Care Excellence (NICE); 2013.
8. Zwaap J, Knies S, van der Meijden C, et al. *Cost-effectiveness in practice*. Zorginstituut Nederland, 2015.

9. Neumann PJ, Cohen JT, Weinstein MC. Updating cost-effectiveness — the curious resilience of the $50,000-per-QALY threshold. New Engl J Med 2014;371:796–797.
10. Kim J, Easton D, Zhao H, et al. Economic evaluation of the Melbourne Mobile stroke unit. Int J Stroke 2021;16:466–475.
11. Lund UH, Stoinska-Schneider A, Larsen K, et al. Cost-effectiveness of Mobile stroke unit Care in Norway. Stroke 2022;53:3173–3181.
12. Chen J, Lin X, Cai Y, et al. A systematic review of Mobile stroke unit among acute stroke patients: time metrics, adverse events, functional result and cost-effectiveness. Front Neurol 2022;13:803162.
13. Claes C, Greiner W, Uber A, von der Schulenburg JMG. An interview-based comparison of the TTO and VAS values given to EuroQol states of health by the general German population. *Proceedings of the 15th plenary meeting of the EuroQol group Hannover*. Germany: Centre for Health Economics and Health Systems Research, University of Hannover, 1999:13-38.
14. Luengo-Fernandez R, Gray AM, Bull L, et al. Quality of life after TIA and stroke: ten-year results of the Oxford vascular study. Neurology 2013;81:1588–1595.
15. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. Stroke 2007;38:1091–1096.
16. Audebert HJ, Schenkel J, Heuschmann PU, et al. Telemedic pilot project for integrative stroke care group. Effects of the implementation of a telemedical stroke network: the Telemedic pilot project for integrative stroke care (TEMPiS) in Bavaria, Germany. Lancet Neurol 2006;5:742–748.
17. Verband der Ersatzkassen. Pflegegrade [Internet]. [cited 2021 Mar 27]. Available from: https://www.vdek.com/presse/glossar_gesundheitswesen/pflegestufen.html
18. IQWIG, Institute for Quality and Efficiency in health care. General Methods Version 6.0. 2020.
19. IST-3 collaborative group. Effect of thrombolysis with alteplase within 6 h of acute ischaemic stroke on long-term outcomes (the third international stroke trial [IST-3]): 18-month follow-up of a randomised controlled trial. Lancet Neurol 2013;12:768–776.
20. Hernán MA, Robins JM. *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC, 2020.
21. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. Pharmacoeconomics 2014;32:1157–1170.
22. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med 2018;37:2252–2266.
23. York Health Economics Consortium. Cost-Effectiveness Plane [Internet]. 2016 [cited 2022 Apr 1]. Available from: https://yhec.co.uk/glossary/cost-effectiveness-plane/
24. R Core Team. *R: A language and environment for statistical computing*. Austria: R Found Stat Comput Vienna, 2017:2017.
25. Bertram MY, Lauer JA, De Joncheere K, et al. Cost-effectiveness thresholds: pros and cons. Bull World Health Organ 2016;94:925–930.
26. OECD. Gross domestic product (GDP): GDP per capita, USD, current prices and PPPs [Internet]. [cited 2022 Sep 9]. Available from: https://stats.oecd.org/index.aspx?queryid=61433
27. Sculpher MJ, Pang FS, Manca A, et al. Generalisability in economic evaluation studies in healthcare: a review and case studies. Health Technol Assess 2004;8:1–192.
28. Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: recommendations for the design, analysis, and reporting of studies. Int J Technol Assess Health Care 2005;21:165–171.
29. Wendt M, Ebinger M, Kunz A, et al. Improved prehospital triage of patients with stroke in a specialized stroke ambulance: results of the pre-hospital acute neurological therapy and optimization of medical care in stroke study. Stroke 2015;46:740–745.
30. Audebert HJ, Schultes K, Tietz V, et al. Long-term effects of specialized stroke care with telemedicine support in community hospitals on behalf of the Telemedical project for integrative stroke care (TEMPiS). Stroke 2009;40:902–908.

12. Curriculum Vitae

My curriculum vitae does not appear in the electronic version of my paper for reasons of data protection.

13. Publication list

Peer-reviewed publications

Number of first authorships: 4

Number of last authorships: 1

The following journal impact factors are based on the "Journal Citation Reports" from Clarivate Analytics, accessed via the Charité Library website on 5th March 2023.

Oliveira Gonçalves AS, Rohmann JL, Piccininni M, Kurth T, Ebinger M, Endres M, Freitag E, Harmel P, Lorenz-Meyer I, Rohrpasser-Napierkowski I, Busse R, Audebert HJ. Economic Evaluation of a Mobile Stroke Unit Service in Germany. Ann Neurol. 2023 Jan 13;93(5):942–51.

Impact Factor: 11.274

Oliveira Gonçalves AS, Werdin S, Kurth T, Panteli D. Mapping Studies to Estimate Health-State Utilities From Nonpreference-Based Outcome Measures: A Systematic Review on How Multiple observations are Taken Into Account. Value Health. 2023 Apr 1;26(4):589–97.

Impact Factor: 5.101

Gonçalves ASO, Laumeier I, Hofacker MD, Raffaelli B, Burow P, Dahlem MA, Heintz S, Jürgens TP, Naegel S, Rimmele F, Scholler S, Kurth T, Reuter U, Neeb L. Study Design and Protocol of a Randomized Controlled Trial of the Efficacy of a Smartphone-Based Therapy of Migraine (SMARTGEM). Front Neurol. 2022 Jun 16;13:912288.

Impact Factor: 4.086

Haas V, Nadler J, Crosby RD, Madden S, Kohn M, Le Grange D, **Gonçalves ASO**, Hebebrand J, Correll CU. Comparing randomized controlled trials of outpatient family-based or inpatient multimodal treatment followed by outpatient care in youth with anorexia nervosa: Differences in populations, metrics, and outcomes. Eur Eat Disord Rev. 2022 Nov;30(6):693–705.

Impact Factor: 5.360

Oliveira Gonçalves AS, Panteli D, Neeb L, Kurth T, Aigner A. HIT-6 and EQ-5D-5L in patients with migraine: assessment of common latent constructs and development of a mapping algorithm. Eur J Health Econ. 2022 Feb 1;23(1):47–57.

Impact Factor: 5.271

Raffaelli B, Mecklenburg J, Overeem LH, Scholler S, Dahlem MA, Kurth T, **Oliveira Gonçalves AS**, Reuter U, Neeb L. Determining the Evolution of Headache Among Regular Users of a Daily Electronic Diary via a Smartphone App: Observational Study. JMIR Mhealth Uhealth. 2021 Jul 7;9(7):e26401.

Impact Factor: 4.947

Raffaelli B, Mecklenburg J, Scholler S, Overeem LH, **Oliveira Gonçalves AS**, Reuter U, Neeb L. Primary headaches during the COVID-19 lockdown in Germany: analysis of data from 2325 patients using an electronic headache diary. J Headache Pain. 2021 Jun 22;22(1):59.

Impact Factor: 8.588

Wetzel B, Pryss R, Baumeister H, Edler JS, **Gonçalves ASO**, Cohrdes C. “How come you don’t call me?” Smartphone communication app usage as an indicator of loneliness and social well-being across the adult lifespan during the COVID-19 pandemic. Int J Environ Res Public Health. 2021;18(12):6212.

Impact Factor: 4.614

Oxelmark L, Whitty JA, Ulin K, Chaboyer W, **Oliveira Gonçalves AS**, Ringdal M. Patients prefer clinical handover at the bedside; nurses do not: Evidence from a discrete choice experiment. Int J Nurs Stud. 2020 May;105:103444.

Impact Factor: 6.612

Gc VS, Alshurafa M, Sturgess DJ, Ting J, Gregory K, **Oliveira Gonçalves AS**, Whitty JA. Cost-minimisation analysis alongside a pilot study of early Tissue Doppler Evaluation of Diastolic Dysfunction in Emergency Department Non-ST Elevation Acute Coronary Syndromes (TEDDy-NSTEACS). BMJ Open. 2019 May 30;9(5):e023920.

Impact Factor: 3.006

Retzler J, Smith AB, **Oliveira Gonçalves AS**, Whitty JA. Preferences for the administration of testosterone gel: evidence from a discrete choice experiment. Patient Prefer Adherence. 2019 May 1;13:657–64.

Impact Factor: 2.314

Sperzel J, Staudacher I, Goeing O, Stockburger M, Meyer T, **Goncalves ASO**, Sydow H, Schoenfelder T, Amelung VE. Comments on the authors’ reply to the critical appraisal concerning “Wearable cardioverter defibrillators for the prevention of sudden cardiac arrest: a health technology assessment and patient focus group study.” Medical Devices: Evidence and Research. 2018;11:377–8.

Impact Factor: not available

Sperzel J, Staudacher I, Goeing O, Stockburger M, Meyer T, **Gonçalves ASO**, Sydow H, Schoenfelder T, Amelung VE. Critical appraisal concerning “Wearable cardioverter defibrillators for the prevention of sudden cardiac arrest: a health technology assessment and patient focus group study.” Med Devices. 2018 Jun 8;11:201–4.

Impact Factor: not available

Whitty JA, **Gonçalves ASO**. A Systematic Review Comparing the Acceptability, Validity and Concordance of Discrete Choice Experiments and Best–Worst Scaling for Eliciting Preferences in Healthcare. The Patient - Patient-Centered Outcomes Research. 2018;11(3):301–17.

Impact Factor: 3.481

Books

Bertram, N., Püschner, F., **Oliveira Gonçalves, A. S.**, Binder, S., & Amelung, V. E. (2019). Einführung einer elektronischen Patientenakte in Deutschland vor dem Hintergrund der internationalen Erfahrungen. In Krankenhaus-Report 2019 (pp. 3-16). Springer, Berlin, Heidelberg.

Other

Pontinha, V., **Oliveira Gonçalves, A. S**., Tran & J., Gohil, S. (2020). By the Numbers: Vaccines HEOR. Value & Outcomes Spotlight. https://tinyurl.com/c67xcte3

Pontinha, V., **Oliveira Gonçalves, A. S.** & Gohil, S. (2020). By the Numbers: Global Perspectives on Precision Medicine. Value & Outcomes Spotlight. https://tinyurl.com/vdmt4nn4

Pontinha, V., **Oliveira Gonçalves, A. S.**, Tran, J. & Gohil S. (2020). By the Numbers: The Current State of Real-World Evidence. Value & Outcomes Spotlight. https://tinyurl.com/tpx3tkkr

Oliveira Gonçalves, A. S., Bertram, N. & Amelung, VE. European Scorecard zum Stand der Implementierung der elektronischen Patientenakte auf nationaler Ebene (2018). Stiftung Münch.

Oliveira Gonçalves, A. S. (2016). Two Economic Paths out of the Crisis? Greece and Portugal in comparison. Bertelsmann Stiftung.

Martins, A. & **Oliveira Gonçalves, A. S.** (2014). Turismo em Portugal e na Bacia do Mediterraneo. Boletim Mensal de Economia Portuguesa – Ministério da Economia

14. Acknowledgments

During the last couple of years, I had the opportunity to cross paths with incredible people, without whom this work would not have been possible.

Firstly, I would like to thank my great supervisor team. I would want to convey my sincere thanks to Prof. Tobias Kurth. First I would like to express my gratitude for advising me to apply to the Heath Data Sciences PhD programme and for taking me on as a PhD student. Thank you not only for all the great advice of methods, but also on how to be pragmatic and get things done. I wish to sincerely thank Dr Annette Aigner. Thank you for all the statistical help and for your R expertise, I will miss those direct comments. I would also like to express a special thanks to Dr Dimitra Panteli for helping me to write better and transform any blunt topic in an interesting subject, skills I definitely want to master. It's been a pleasure working with the three of you.

Beyond my exceptional advisors there have been many other people who throughout these years made a great contribution to my work.

I would like to thank all my co-authors for their unconditional support. Their advice and perseverance during the numerous revisions of the manuscripts have been invaluable.

My thesis benefited greatly from the data of the SMARTGEM and B_PROUD groups and I thank all the persons involved in the collection and management of these data, as well as the involved patients.

I could not forget to thank the HDS Programme, especially to Prof. Tobias Kurth, Dr Jessica Rohmann and Muhammad Barghouth, for helping me navigate throughout the organisational and bureaucratic issues along the way. Furthermore, thank you for the incredible courses you have provided the students with.

Outside academia I am lucky to be surrounded by great friends. To them all, I am extremely grateful.

I could also not have done without my family. Obrigada a todos pelo apoio durante estes longos anos. And for my partner Ángel, thank you for your unconditional support.