Juulia Suvilehto

Using Bring-your-own Device Model in Clinical Trials to Capture Patient-reported Outcomes

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 10.2.2014

Thesis supervisor:

Prof. Jouko Lampinen

Thesis advisor:

M.Sc. Rauha Tulkki-Wilke

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Juulia Suvilehto

Title: Using Bring-your-own Device Model in Clinical Trials to Capture Patient-reported Outcomes

Date: 10.2.2014 Language: English Number of pages:6+51

Department of Biomedical Engineering and Computational Science BECS

Professorship: Computational Science Code: BECS-114

Supervisor: Prof. Jouko Lampinen

Advisor: M.Sc. Rauha Tulkki-Wilke

There is clear interest in the pharmacological industry for using subjects' own devices for capturing patient reported outcomes in clinical trials. This has potential to reduce costs and increase the quality of the captured data. The potential gains are especially great in late-phase studies.

The practical part of this thesis consists of implementing Amsterdam IADL Questionnaire® for the CRF Health TrialMax Web™ platform. Attention was paid to ensuring uniform presentation on a range of screen sizes while also providing a tailored look-and-feel to each device. The implementation was tested with cognitive interviewing methodology. The subjects (N=6) filled in a subset of the items on two different devices (randomized crossover design). The interviews were analyzed for potential differences in the two administrations.

In this thesis, the author inspects using bring your own device model in clinical trials from regulatory and practical points of view. The current technology enables capturing source data with BYOD in a way which is in keeping with the relevant regulations. The results from cognitive interviews were not comprehensive enough to conclude whether or not the device affects the psychometric properties of the instrument. The quality, especially cross device comparability, of the captured data is still debatable. Further cross platform validation work is is advised before deciding to use the model for formal instruments. This is important to ensure that BYOD does not introduce any unnecessary bias in the measures.

Keywords: Bring Your Own Device, Clinical Outcome Assesment, Clinical Trials, electronic Patient Reported Outcomes

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Juulia Suvilehto

Työn nimi: Omien laitteiden käyttäminen kliinisissä lääketutkimuksissa potilasperäisen datan keräämiseen

Päivämäärä: 10.2.2014 Kieli: Englanti Sivumäärä:6+51

Lääketieteellisen tekniikan ja laskennallisen tieteen laitos BECS

Professuuri: Laskennallinen tiede Koodi: BECS-114

Valvoja: Prof. Jouko Lampinen

Ohjaaja: DI Rauha Tulkki-Wilke

Lääketeollisuuden yrityksiä kiinnostaa mahdollisuus käyttää koehenkilöiden omia laitteita potilasperäisen datan keräämiseen kliinisissä lääketutkimuksissa. Tällä menetelmällä on potentiaalia laskea kliinisten lääketutkimusten kustannuksia ja parantaa kerätyn datan laatua. Potentiaaliset edut ovat merkittäviä erityisesti loppuvaiheen kliinisissä kokeissa.

Tämän työn käytännön osuuteen kuului Amsterdam IADL Questionnaire® instrumentin toteuttaminen CRF Healthin TrialMax Web™ alustalle. Huomiota kiinnitettiin erityisesti kyselyn yhdenmukaiseen ulkoasuun erikokoisilla näytöillä siten, että samalla säilytettiin vaikutelma juuri kyseiselle laitteelle suunnitellusta näkymästä. Toteutusta testattiin kognitiivisilla haastatteluilla. Koehenkilöt (N=6) täyttivät osan kyselyn kysymyksistä kahdella eri laitteella (satunnaistettu vaihtovuoroinen koeasetelma). Haastatteluiden analyyseissä pyrittiin löytämään eroja kahden haastattelukerran välillä.

Tässä diplomityössä kirjoittaja tutkii mahdollisuutta käyttää käyttäjien omia laitteita kliinisissä lääketutkimuksissa sääntelyn ja käytännön näkökulmista. Nykyinen teknologia mahdollistaa myös potilaiden omilla laitteilla kerättyjen lähdetietojen käsittelyn tavalla, joka vastaa oleellisien säädösten vaatimuksiin. Työn puitteissa toteutettujen kognitiivisten haastattelujen tulokset eivät riitä tekemään johtopäätöksiä siitä, vaikuttaako käytetty laite instrumentin psykometrisiin ominaisuuksiin. Jatkotutkimuksia alustojen välisestä validoinnista suositetaan ennen kuin omien laitteiden käyttö sallitaan määritellyille instrumenteille. Tämä on tärkeää jotta voidaan varmistua siitä, ettei omien laitteiden käyttö aiheuta systemaattista virhettä tuloksiin.

Avainsanat: elektroniset potilasperäiset tulokset, hoitotuloksen arviointi, kliininen lääketutkimus, tuo oma laitteesi

Preface

I would be remiss if I didn't acknowledge the help and support I have received from so many different directions. First and foremost, I would like to thank CRF Health for the opportunity to work on this thesis and make it my own. It is with immense gratitude that I acknowledge the support and help of my advisors Paul O'Donohoe and Rauha Tulkki-Wilke. Their support and advice have greatly contributed to this thesis. All of my colleagues at CRF Health have been very generous with their knowledge. I would like to give special thanks to Olli Kotiranta for his unfailing patience and technical assistance with the practical implementation.

My most sincere thanks go to my supervisor, prof. Jouko Lampinen, for his valuable feedback and for taking on a master's thesis that is not directly in his own field of study.

I owe a great debt of gratitude to my family and extended family, who have supported me through my studies in the university as well as in schools before that. You, with your personalities, different brands of brilliance, and kindness, have inspired me more than I can express.

During my studies I have also managed to amass a circle of friends, who I consider my chosen family. Thanks to all of you, my time as a undergraduate student is something to look back on with fondness and a dash of feigned scandalization. Especially I would like to thank Venla and Anna for their unfailing support during the thesis process and for being, like, totally awesome.

Otaniemi, 10.2.2014

Juulia T. M. Suvilehto

Contents

Symbols and abbreviations

Symbols

n	number of subjects per group
n_{tot}	total number of subjects
s^2	estimated variance of the observations
Δ	equivalence margin
α	probability of type I error
β	probability of type II error
s_d^2	variance of the difference between administrations (within subject)
ϱ	estimate of the correlation between observations
ϱ_0	the lower bound for ICC
ϱ_1	estimated ICC in population

Abbreviations

AD	Alzheimer's disease
BADL	basic activities of daily living
BYOD	bring-your-own-device
ClinRO	clinician reported outcome
CI	confidence interval
(e)COA	(electronic) clinical outcome assessment
EDC	Electronic Data Capture
EMA	European Medicines Agency
FDA	The United States Food and Drugs Administration
HRQOL	health related quality of life
IADL	instrumental activities of daily living
ICC	intraclass correlation coefficient
IND	Investigational New Drug
ISPOR	International Society For Pharmacoeconomics and Outcomes Research
IVR	interactive voice response
MID	minimal important difference
MTD	maximum tolerated dose
NME	new molecular entity
ObsRO	observer / observed reported outcome
OS	operating system
(e)PRO	(electronic) patient reported outcome
QALY	quality adjusted life years
QoL	quality of life

1 Introduction

Clinical trials proving the efficacy, i.e. the treatment benefit, and safety of a newly formulated treatment are necessary in order to get sales permits in the US and in Europe. The trials are mostly carried out by pharmaceutical so-called sponsor companies, which wish to bring specific drugs to the market. To ensure the wellbeing of potential future patients, Regulatory authorities like the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) closely inspect the reported methods and results to assess whether the efficacy and safety of the new treatment has been reliably demonstrated. These time-honored measures balance between the desires of regulatory authorities for conclusive and statistically significant results and those of sponsor companies for obtaining marketing permits in a cost-effective manner.

In clinical trials, certain pursued treatment benefits may be best tracked and measured by the patients themselves, rather than by objective, physiological measures or by a physician's professional judgement. Such benefits called patient-reported outcomes (PRO) include how patients feel (for example how much pain they experience), function (for example whether their ability to dress themselves has been impacted) and conduct themselves (for example how many pills they have taken). Specialized instruments are commonly used to measure PROs, for instance carefully developed and validated questionnaires with known psychometric properties.

Traditional paper versions of PRO instruments are being increasingly criticized for not providing reliable information about who filled in the instrument and when, making patient compliance to study protocol difficult to assess. Additionally, the implementation of paper tends to create messier data with accidentally or intentionally missing data points. Combined, these caveats decrease the achievable statistical power in clinical studies and limit conclusions that can be drawn about the efficacy and safety of the treatments being tested.

The pharmacological industry as whole is slowly moving from paper instruments towards electronic data capture of a wider range of outcomes, including PROs. This shift is partially due to the fact that the requirements of regulatory authorities for patient-captured data practically enforce researchers to use electronic data capture, since paper versions cannot meet the required standards. However, it is time-consuming for any industry to adapt to this turnround.

The implementation of a paper-and-pen PRO instrument to electronic format is typically done on a single device at a time, usually as an application. This is installed in such a way that prevents the users from using the device for anything other than the application. While this is a significant improvement on using pen-and-paper, the use of single dedicated devices is not without problems.

One of the key challenges is the cost of the devices. In industrialized countries, it is increasingly common for most people to have a computer at home. Most middle-aged and younger individuals also have a mobile device (e.g. smartphone, tablet, laptop). Utilizing these devices has potential to significantly reduce the costs associated with ePRO. In bring-your-own device (BYOD), patients are able to fill

in the diary using their own mobile device. Consequently, sponsors of clinical trials are increasingly interested in adapting a BYOD model for clinical trials.

This master's thesis was done for CRF Health, a leading provider of systems for electronically capturing Clinical Outcome Assessments, including PROs. The research question in this thesis can be divided into three interconnected subquestions:

1. Would it be possible, in theory, to use BYOD in clinical trials to capture patient reported outcomes?

2. How should instruments be designed to ensure measurement equivalence across a range of devices?

3. What kinds of validation are necessary to prove measurement equivalence of a multi-platform implementation of a PRO instrument?

The author provides one possible model for designing and testing for measurement equivalence over an innumerable range of devices. Technical issues of BYOD are also touched upon. More advanced technical aspects of the system, such as security and authentication of electronic data capture, were considered out of scope of this thesis. The interested reader is referred to the FDA draft guidance Electronic Source Data in Clinical Investigations (US Department of Health and Human Services Food and Drug Administration, 2013a) to see what needs to be taken into account.

The author presents an implementation of the devised model to one instrument, the Amsterdam IADL Questionnaire®. The implementation was done in an iterative manner, with feedback from the instrument author following each iteration. The resulting adaptation was tested with end users, and the suitability of different testing paradigms for qualitative testing of measurement equivalence over a range of devices is discussed. Finally, the author also discusses the plausibility of the suggested model and BYOD in general, in reference to clinical trials.

2 Background

In this chapter we will first go over drug development in general and clinical trials for new drugs in particular. From there we will consider measuring health and how health measurement has evolved over the years. Furthermore we will discuss patient-reported outcomes (PROs) and other clinical outcomes assessments (COAs), which are commonly used to capture patient data in clinical trials. After that we will go over what kinds of proof are needed to consider a formal, multi-item PRO scale to be reliable and valid. We will also discuss the equivalence of two implementations of the same instrument and how this equivalence can be formally proved. We will use implementing originally paper-and-pen instruments to electronic mode of administration as a benchmark, since it has been so thoroughly researched in the past, and then move on to discuss equivalence in the same format but in different contexts, particularly electronic implementations which are not tied to a particular device.

Clinical Trials

Testing a new drug therapy in people in clinical trials is necessary for all new drug compounds and formulations before any wide scale use of the new drug in Western countries. These clinical trials are required and controlled by regulatory authorities, most importantly by U.S. Food and Drugs Administration (FDA) in the United States and by the European Medicines Agency (EMA) in the European Union.

The regulatory authorities' main concern is patient safety. The initial task of clinical testing is to ensure a full understanding of the effect of the new treatment on humans. The sponsor company also has to prove with sufficient statistical power that for the intended user group their proposed treatment is either more effective than the best current treatment or,in some cases, as effective but with a different treatment mechanism or less severe side effects. Based on safety and efficacy data gathered in clinical trials, regulatory authorities grant marketing permission and dictate what health-related claims the sponsor company may use in product labels and advertising. The ways in which the effect of a treatment can be measured is discussed in the next section, Measuring Health. Before that, we will go over the general outline of clinical trials.

The development of a new drug is an expensive process. A study by DiMasi et al. (2003) states that average out-of-pocket pre-approval cost of a drug was approximately US$ 400 million and the cost has been rising at 7.4% above inflation rate. When one accounts for the drugs that never make it to the market, the cost of finding and bringing to market one successful drug can be over US$ 5 billion. (Forbes, 2013) Discovering and bringing a new drug to the market is a long process with several distinct phases.

Drug testing begins with identifying a new molecular entity, which shows promising results in pre-clinical studies. The new compound is tested in vitro on cell populations and in vivo on animals to screen for toxicity and pharmacological activity in a simplistic model. Once a drug shows acceptable results in lab tests, the sponsor

company starts testing the drug on people. In the US, the sponsor company has to fill in an Investigational New Drug (IND) Application if they intend to run trials in more than one state, i.e. if they will transport the new compound over state borders. This is to be done prior to commencing clinical trials.

The proceedings related to a single trial are called the study protocol. Protocol dictates for example how the treatment is administered and what is measured, how often and by whom. Throughout the trial, adherence to the protocol is tracked to ensure that the results obtained describe the actual effect of the treatment. For example, if the protocol states that subjects should fill in patient diary within 1 hour of waking up to measure morning sickness, it might compromise the validity of the results if the subject instead filled in their diary in the afternoon.

The success of a clinical trial is measured with *endpoints*. The targeted endpoints, i.e. those aspects of the disease the sponsors expect the new treatment to influence, need to be declared to the regulatory agency while applying for permission to conduct the clinical trial. The primary endpoint is the most crucial result of the clinical trial, and it is related to either the safety or efficacy of the treatment. If the primary endpoint is not met, the rest of the results will not even be considered.

Secondary endpoints are less important results, but they can still provide a competitive advantage for the tested treatment. A clinical trial can have and often has several secondary endpoints defined, these can be used as labeling claims if the primary endpoint is first met. Labeling claims, i.e. what can be said of the treatment's effect in its packaging and marketing, are also controlled by the regulatory authorities. These can be for example the speed of recovery or health related quality of life (the concept of HRQoL is explained more thoroughly in the following section). In addition to primary and secondary endpoints, a trial can have exploratory endpoints, the purpose of which is to serve as a basis for a new hypothesis.

Trials are designed in a manner that reduces unnecessary bias and is in keeping with good clinical practice. In order to determine the actual effect of the new drug, subjects are often divided in different groups called treatment arms. One group receives the new treatment and the other group, called a control group, receives standard treatment for their condition or, where there is no standard treatment available, a placebo. Placebo-controlled trials are less common with serious illnesses since it is considered unethical to withdraw treatment from the patients even for a limited amount of time. It is also possible, although not as common, to have trials with several treatment groups. This is done when the optimum dosage of the new drug is not yet known.

Trials are generally conducted as blind studies. Single blinding means that the subject does not know if they are in the treatment group or the control group. In double blinding even the treating practitioner (investigator) does not know in which treatment arm each patient is. The aim is to make sure that the evaluation of the health effects of the drug / placebo are not affected by subjects and clinicians knowing that they are (not) dealing with a new treatment.

Clinical trials are commonly divided in phases I-IV (sometimes denoted phases 1-4), where each phase has its own aim and challenges. After each phase the drug is evaluated and a decision is made on whether or not the clinical trial will proceed

to the next phase. The particular research questions of each trial type are briefly explained next.

Phase 0, or first-in-man studies are sometimes carried out before phase I. They represent the first time a new drug is used in human subjects. The doses are microscopic compared to the doses calculated to have a pharmacological effect. First-in-man studies are aimed to gain preliminary data on the toxicity of the new compound, first in one healthy subject and then one or two more.

Phase I trials are focused on finding a safe dose of the medication. These are often done on-site in a hospital in order to keep close track of the effects of the new drug, especially any possible immediate harmful side effects. Commonly phase I trials are conducted with a small number (between 20 to 80) of healthy adults or, in some rare cases, patients with the target disease (e.g. cancer, hematologic malignancy), who are given slightly different doses of the treatment. The first subjects are given the lowest dose level, with doses escalating and de-escalating for subsequent cohorts of subjects based on the previous group's reaction to their treatment. The aim is to find the maximum tolerated dose (MTD), which does not cause unacceptable toxicity. The research questions in phase I trials are related to patient safety and do not require control groups or blinding.

Phase II trials aim to establish the effects of the drug on target population i.e. people with a specific condition or disease. Here researchers seek for further safety data and proof-of-concept of the treatment's beneficial effects. Phase II trials usually have between 100 and 300 subjects and they can have single treatment arm or compare multiple arms, e.g. different dosages. In phase II the research questions are generally related to the short-term efficacy of the drug. If the trial indicates that the treatment might be effective and the risks are acceptable considering the severity of the disease, the drug moves on to phase III. If sufficient proof cannot be gathered, the endpoints are to be reconsidered or further trials on that drug are halted. The decision is especially important at this stage, since phase III trials are generally very expensive.

Phase III trials are very resource intensive, with a greater number of subjects, from several hundred to thousands, and longer duration. The aim is to establish the actual treatment effect of the new drug compared to standard treatment, or in some rare cases, placebo. With increasing number of subjects and longer duration, also less common side effects are more likely to be noticed and the statistical power of the trial is increased. The primary endpoint in phase III trial is tied to the treatment efficacy, e.g. survival is a common endpoint in cancer trials. Phase III trials are generally double blind with randomized assignment to control and treatment arms. If the endpoints are met in a satisfactory manner and the trials have been conducted properly, the regulatory agencies will consider granting market permission at this stage.

Sometimes the market approval can be conditional to continuing safety evaluations after marketing authorization. When post-marketing surveillance studies are conducted, they are referred to as phase IV trials. Phase IV trials are needed in order to detect adverse effects of long-term continuous use of a new treatment. This is especially true for interventions targeting chronic conditions, where phase III trials

do not last as long as the typical length of a treatment.

The regulatory authorities have very specific regulations and recommendations on practical issues of clinical trials, such as guidelines for data capture and subject compensation. According to FDA, subject compensation should be considered only when health benefits to the subjects are remote or nonexistent. Other than phase I where the subjects are healthy adults, compensating subjects for participation is not advised. The reasoning behind this is that people should not participate in clinical trials for personal gain but out of willingness to help science. Rules relating to data capture will be further discussed under section Electronic Data Capture.

In addition to how data is captured and stored, it is critical to understand what types of data can be used in clinical trials. It is crucial that the measures used to evaluate the endpoints give accurate and relevant information about the effects of the treatment. The following section will explain approaches to measuring health in more detail.

Measuring Health

Health can be measured at the individual level and at the population level. These two are strongly related, even though measures that are useful for one might not be as helpful for the other. Population level health indicators are often used for comparing different ethnic groups and nations.(McDowell, 2006)

The simple act of measuring also influences health on population level: The measures traditionally used for indicating population health are straightforward numerical measures such as mortality rate. Death is easy to confirm and in most places all deaths are noted in public records, so the data are generally complete. For example in pre-industrial societies, infant mortality rate is commonly used to indicate overall health levels. However, as the mortality rates begin to drop they become less informative. With declining infant mortality rates, health issues related to prematurity and low birth weight will become more prominent. What happens is that with the resolution of one health problem new ones emerge, these become more prominent and the usefulness of the earlier health indicator will be reduced. In this way, using a specific indicator draws attention to the problem and the (successful) resulting interventions reduce the prevalence of the problem thereby reducing the information content achieved from the indicator. (McDowell, 2006)

Owing to the impact of measuring on the population being measured, health indicators are continuously evolving. When mortality is no longer sufficiently informative, morbidity is taken under consideration. Since all pathological states cannot be measured using the same instruments, one needs to devise a way to set these on a scale for comparison. The question of comparing different morbid states against each other is not trivial and it is closely related to the concept of "health"(Goldsmith, 1972; McDowell, 2006). World Health Organization and others (1948) define health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."

We will now take a closer look at how one can compare different states of health or people's experiences of their own health. In clinical trial?s setting, it is especially

interesting to track the changes in an individual's state before, during, and after a treatment. There are two distinct ways to measure health: one is objective, physician-led tracking of observable symptoms, the other is asking the patient to subjectively evaluate or describe their symptoms.

Particularly in western countries, it is very common to measure health in terms of physiological changes. Blood pressure is a prime example of a very commonly used physiological measure. It is mostly objective and can be repeated any number of times. The connection between a blood pressure measure and physiological blood pressure is well known and the measure can be used to draw direct conclusions of the patient's state. There can be slight variation in the results, depending on who measures and what equipment is used, but the errors are small compared to the actual variation.

Although there are a number of measures that can be easily observed and interpreted from the patient, not all changes can be measured in the same way. For instance, let us consider testing the memory of an Alzheimer patient. Usually such measurement includes a clinician having the patient perform a memory task and then drawing conclusions about the state of the patient from the results. The result depends greatly on several factors, including but not limited to, on patient's state of alertness during testing, the instructions that the clinician gave to the patient, and on the amount of experience that the clinician has with that measure.

Patients show increasing interest in results obtained with subjective measures in clinical trials (Guyatt et al., 1993; Fayers and Machin, 2007; Doward et al., 2010). There are several illnesses and treatments the actual impact of which on the patients' lives cannot be quantified with physiological measures, at least not in a way which would respond to the patient's own perception. Subjective measures can be related to the symptoms that a patient is experiencing. These are difficult to measure objectively (e.g. pain, dizziness). Another such measure is quality of life (QoL), or how the subjects experience their own wellbeing and ability to function in the society. In fact, in some cases such as cancer chemotherapy, continuing a treatment might not be worthwhile due to the deteriorating QoL and treatment costs even if the objective condition of the patients might slightly improve (Buccheri et al., 1989).

QoL does not have a definition that is widely accepted. It is generally considered to encompass components of happiness and satisfaction with one's life. QoL can be defined depending on the area of application, e.g. to a city planner QoL might include access to green spaces. Such a broad definition is not necessary in clinical trials. The term health related quality of life (HRQoL) was coined to remove the ambiguity and describe the focus of interest (Fayers and Machin, 2007). Some researchers argue that narrowing focus down to HRQoL focuses too much attention on clinically defined impairment and disability, disregarding the patient's own areas of interest (Doward and McKenna, 2004).

Even with the focus narrowed to HRQoL, the exact content of the term is still open to debate. Many researchers circumvent this issue by stating their own definition and matching it to their own purposes. Most researchers agree, however, that HRQoL is a multidimensional concept and that it should include at least some of the following aspects: toxicity, general health, physical functioning, emotional function-

ing, cognitive functioning, role functioning, physical symptoms, social well-being, sexual functioning, and existential issues. (Fayers and Machin, 2007)

HRQoL is a good example of a measure that is hard to quantify objectively, as it is very subjective and best measured from the patient perspective. There is a body of research stating that clinicians and other observers give very different evaluation from that of the patient when queried about HRQoL. It has been suggested that observers tend to overestimate the impact of obvious (physical) symptoms and underestimate the importance of psychological issues. Medical professionals tend to also underestimate the impact of expected symptoms and toxicity, even though these can greatly diminish HRQoL from the patient's perspective. (Fayers and Machin, 2007; Doward et al., 2010)

As an example, studies of vomiting and nausea in patients undergoing chemotherapy indicate that medical personnel expect these symptoms to appear and therefore do not report them unless the case is exceptionally severe. On the other hand, a patient might self-report having experienced quite a lot of vomiting, even if a doctor would report no problems. (Fayers et al., 1991) Focusing on clinician-reported outcomes in assessing any treatment might leave out information that is crucial to the patients.

HRQoL is here used as an example of a phenomenon best measured from the patient perspective. The challenges and benefits of measuring HRQoL from patient perspective presented here can be adapted to other similar multifaceted health-related phenomena and different subsets of the aspects of HRQoL. The concepts that are best measured from patient perspective are often assessed via event logs and patient diaries or standardized questionnaires, collectively called instruments.

The instruments can either be general or disease-specific. Disease-specific instruments focus on the most important effects of that disease, and are not suited for comparing treatment results of several diseases. Generic instruments are suitable for patients independent of the underlying disease. Generic instruments which produce a single score on a 0-1 scale can be used to calculate quality-adjusted-life-years (QALYs). QALYs are used to evaluate the effectiveness of healthcare interventions in terms of change to both length and quality of life. (Räsänen et al., 2006)

Patient-Reported Outcomes

As we have discussed above, many of the changes tracked in clinical trials can be directly measured, e.g. blood pressure after using antihypertensive agents or survival rate for new cancer medication. For other aspects, such as the sensation of pain or depression, the best, and sometimes even the only way to find out is to ask from a patient.

In addition to physical changes in a patient, it is often of interest to know what they are experiencing and how they are coping with their everyday life as a result of the symptoms of the disease. This scope can be divided into three subcategories: Firstly how do subjects feel i.e. the sensations and emotions they experience. Secondly, how do they behave i.e. actions, which have occurred within a predetermined recall period. Finally, it can be of interest to know how the patients function, i.e.

can they work and do the necessary daily chores. (Byrom and Tiplady, 2010) These concepts can be measured in absolute terms or as change occurred since previous measure.

The phenomena listed out above can be called patient reported outcomes (PROs). The umbrella term PROs describe any outcome evaluated directly by the patient based on their perception and with no interpretation from third party. PROs can cover health related quality of life (discussed in section Measuring Health), symptoms, health status, satisfaction with treatment, etc. PRO scales can be either single dimensional or multi-dimensional. (European Medicines Agency, 2005; US Department of Health and Human Services Food and Drug Administration, 2009)

PROs are the best way to measure sensations and emotions, and they can be used to measure behavior, especially behavior which is rarely observed. Byrom and Tiplady (2010) advise against using PROs to measure function. In some cases, such as when the subject has some form of degenerative nerve disease (e.g. Alzheimer's) or when the potential loss of function has stigma (e.g. incontinence), the reliability of PRO function measures is a valid concern. The patient can sincerely believe that they can function normally or bend the truth in order to appear more competent (social acceptability bias) even though it is clear to caregivers that this is not the case.

Instruments similar to patient reported outcomes can be observer reported outcomes (ObsROs), clinician reported outcomes (ClinROs) or proxy-reported outcome. Using a proxy (proxy-reported outcomes) means that someone else, e.g. the parent of an infant, responds to the PRO questionnaire as if they were the subject. Proxies are generally discouraged: "We discourage proxy-reported outcome measures ... [f]or patients who cannot respond for themselves (e.g., infant patients), we encourage observer reports that include only those events or behaviors that can be observed." (US Department of Health and Human Services Food and Drug Administration, 2009, p. 21)

Observer-reported outcomes are filled in by a person who does not have relevant clinical training, commonly a caregiver or a family member. It is recommended that especially ObsROs include only behavior and events, which can be observed. Clinician-reported outcomes are instruments, which are filled in by trained clinical personnel. Many standardized tests of cognitive capability, e.g. tests measuring symptoms for dementia, are ClinROs. Unlike PROs, ObsROs and ClinROs can include interpretations or opinions on the condition of the patient. (US Department of Health and Human Services Food and Drug Administration, 2009) It is to be noted that objective clinical endpoints are preferred as primary endpoints and PRO endpoints are considered when suitable objective clinical endpoints do not exist.

FDA has given explicit guidelines about where, when and why PRO measures can be used in clinical trials. The sponsor companies generally consider "Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims" the most important regulatory guidance paper on PRO use in clinical trials. It states that "use of a PRO instrument is advised when measuring a concept best known by the patient or best measured from the patient perspective . . . PRO measures often represent the effect of disease (e.g., heart

failure or asthma) on health and functioning from the patient perspective."(US Department of Health and Human Services Food and Drug Administration, 2009)

At the time of writing, EMA has not released a similar formal guidance. However, European Medicines Agency (2005) has published a reflection paper on using HRQoL measures in clinical trials.

PROs can be considered as endpoints in clinical trials in a number of cases. Most often, the purpose is to devise a value proposition which is more comprehensive than the traditional safety and efficacy data (Gnanasakthy et al., 2013). Fayers and Machin (2007) list out the following cases specifically relating to QoL:

1. QoL can be used as the primary endpoint. This is often the case in palliative care and with some chronic, incurable diseases.

2. The expected efficacy of the new treatment is very similar to the efficacy of products already in the market. Having QoL as a secondary endpoint might help establish the new treatment as preferable.

3. With treatments where the overall failure rate is high, QoL should be considered in addition to efficacy.

4. Even if the benefits in cure rates or survival for a new form of treatment might be marginally better than an established treatment, it can be offset by QoL deterioration.

Especially in cases where the treatment efficacy is not significantly different from that of a competing treatment, PRO endpoints can provide extra value to the sponsor company in form of more beneficial pricing decisions (Doward et al., 2010).

The use of PROs as endpoints in clinical trials in USA has been tracked by Willke et al. (2004) and Gnanasakthy et al. (2012) for new molecular entities (NMEs). The reviews spanned 1997-2002 and 2006-2010, respectively. According to the reviews, the use of PROs depends greatly on the therapeutic area. The NMEs that had only PROs as their endpoints were mainly symptomatic in nature (Willke et al., 2004). Pain was among the most prominent symptoms to get PRO based label claims in both reviews. Pain is an excellent example of an area where PROs provide valuable insight in the patients' experiences. It cannot be measured objectively and a clinician or other observer can, at most, describe patient pain-related behavior (e.g. "the patient is holding his head as if in pain") or ask the patient to describe their sensations.

The amount of FDA approved product labels that included PRO claims was around 30% in 1997-2002 (Willke et al., 2004). In 2006-2010 the number was approximately 25% (Gnanasakthy et al., 2012). The percentage of approved PRO label claims reduced slightly in the latter overview, and at the same time the amount of approved new drugs was dramatically smaller (215 in 1997-2002 and 116 in 2006-2010). Of the NME products with PRO endpoints granted between 2000 and 2012, 81% were included in the primary endpoints. Furthermore, most of the primary PRO endpoints were about disease specific signs and symptoms (Gnanasakthy et al., 2013).

The cited research strongly suggests that there is a need for patient level data in developing new treatments. It is important to note that the psychometric properties of a PRO instrument, i.e. the aspects which influence the quality of the data, are related to the scale itself, the population and the context of administering the scale. This causes some challenges with ensuring the quality of the data. These challenges, particularly reliability and validity, will be further discussed in the following sections.

Reliability

The aim of any health measure is to produce results which provide enough separation between people in different states of health but which are fairly consistent within a person who is in a certain state. That means, that when comparing two persons for any measure, e.g. depression, the measurement should reflect the "true" differences between people and "true" consistency within a person. Terms *reliability* and *validity* are used to describe these properties. Essentially, validity describes how well a scale or measurement measures the phenomenon of interest. Reliability describes how reproducible and consistent the scale or measurement is. This section covers different forms of reliability and the following section discusses validity.

So, if out of two persons one is clinically depressed and the other one isn't, the test scores for a depression instrument should reflect this (validity). When the same test is administered a week later to the same two persons, the difference between them should still be there and in addition, if the mental state of either has not changed significantly, their scores should be approximately the same as they were a week earlier (reliability). Also, if the depressed person goes to therapy and gets better, the test scores should change accordingly and be closer to the non-depressed range than in the beginning (validity and reliability). (Streiner and Norman, 2008)

In classical test theory, a score is determined by a underlying "true score", which we are unable to tap into directly, and a variable degree of error. The error can be caused by e.g. the measurement method or the measurer. Errors in measurement can be divided into two subgroups. *Systematic error* or *bias* arises every time a measurement is made in a particular setting, e.g. by the same measurer or with the same equipment. Bias always acts in the same direction and is the same size. *Random error*, or *noise*, is non-predictable and can come about due to e.g. inattention or mechanical inaccuracy. Random error cancels itself out with enough repeated measurements. Reliability is a way to describe the amount of both systematic and random error inherent in a measurement.

Even though biases occur in all kinds of measurements, there are a number of biases which are specific to PRO measures. Rather, biases in subjective measurements can be caused by, e.g. the way the subjects interpret the items in the instrument, the circumstances of their illness or even the respondents' personalities (McDowell, 2006). One of the common challenges with PROs is that people in general are not good at comparing their current sensations, e.g. feeling of pain, with a previous state. Because of this, it is a standard in the industry to always relate standardized PRO items to a specific time period, for example "today", register the item at regular intervals over a period of time, and then compare the responses afterwards,

instead of letting the patients themselves do the comparison.

It is important to note that reliability is not a quality of the instrument alone. Reliability is an interaction between the instrument, the group of people using the instrument, and the situation. Therefore, the reliability of an instrument is always related to a specific sample and circumstances. (Streiner and Norman, 2008) Moreover, in many branches of medicine the increase in treatment benefit is likely to be very small so all possible sources of bias in clinical trials are avoided extremely carefully. This is especially important to remember when we move forward to discuss measurement equivalence.

A significant part of reliability is the ability to detect meaningful change. Minimal important difference (MID) is the smallest change in score or the smallest score that is likely to be meaningful from the patient's or clinician's perspective. Measuring MID is not a straightforward process and it can vary depending on the direction of the change, the population, and contextual characteristics. There is no one universally agreed way to establish a MID for an instrument and population, but Revicki et al. (2008) have gathered some guidelines for MID evaluation.

MID in a new instrument is recommended to primarily be based on several anchors, which are already known to be responsive to change. These anchors can be clinical or patient-based. With the help of anchors the patients are divided into different groups (no change, small positive changes, large positive changes, small negative changes, and large negative changes). The groups identified as tose with small positive and small negative changes are then further evaluated to determine MID in the instrument of interest. Sometimes the anchor, which is used to do the grouping, is simply asking patients to evaluate how they feel compared to for instace two weeks ago. It should be kept in mind that asking patients themselves to evaluate whether their condition has improved, worsened, or remained the same is very prone to recall bias. (Revicki et al., 2008)

Most forms of reliability can be measured by repeating the same measurement for a number of times. If the measurements are done in exactly the same way, e.g. by the same person or using the same settings, the systematic error should remain the same and it is possible to examine just the random error. This is called *test-retest reliability*. Each instrument has their characteristic accuracy, which determines how detailed data can be gathered by using that measurement. In an optimal instrument, the test-retest reliability would be very good (i.e. there would be very little variation) compared to the scale and the variation from actual changes in subject's condition. Measuring for test-retest reliability on large enough a group provides the instrument developers useful information about the amount of noise that is to be expected. If an instrument is then used in a clinical trial to prove treatment effect, the improvement of scores needs to be beyond the expected test-retest variability. (Fayers and Machin, 2007)

Inter-rater reliability describes the agreement between two raters. PRO measures are generally self-administered and thus inter-rater reliability as such is not an issue, but for example ClinRO scores can greatly differ depending on the clinician doing the evaluating. Usually, reliability is considered to be bound to agreement, i.e. if two raters or repeats would give the same results. While agreement is an important

factor in inter-rater reliability, one should keep in mind that the ultimate question of interest is the difference between individuals. Even absolute agreement over raters is useless, if the thing to be measured cannot differentiate between the subjects in a meaningful way.

Another aspect of interest is whether that difference between individuals is related to the phenomenon of interest. This is the core of validity and it will be discussed in detail the next section, Validity and validation.

Validity and Validation

Validity is often described as the extent to which a test measures what it is alleged to measure (McDowell, 2006). For a valid instrument the degrees of change in scale co-vary with the change in attribute. It is easy to understand and prove the relationship when measuring a physical quality (e.g. temperature with a mercury thermometer). The task becomes harder when measuring attitudes, beliefs, and feelings or complex concepts such as "Quality of Life" and "amount of cognitive decline", since the correctness of the measure becomes dependent on the definition of the attribute. (Streiner and Norman, 2008) Often it is not possible to choose one instrument that would always be the best way to measure fuzzy concepts. The validity is also dependent on the context of the measurement. Different ways to validate instruments, i.e. prove their validity, are discussed next.

Validity is commonly divided to the three "Cs", *content validity*, *construct validity*, and *criterion validity*. These all need to be established separately in order to be able to say that an instrument has been validated. (Fayers and Machin, 2007)

FDA regulations for Patient Reported Outcomes (PROs) define content validity as the extent to which the instrument measures the concept of interest (US Department of Health and Human Services Food and Drug Administration, 2009). Content validity is specific to a given population, condition and treatment.

FDA also emphasizes target patient population input in item generation, which is a crucial part of establishing content validity. In order to ensure that the items in an instrument capture the concept of interest the target patient population should be given a chance to help generate item wording, assess item clarity and readability, and evaluate whether the items cover the entire range of on the phenomenon of interest. Patient input should be sought until saturation, a point where no new or relevant information emerges, is reached. Patient involvement is considered so important, that US Department of Health and Human Services Food and Drug Administration (2009) state in the PRO Guidance that "[b]ecause the purpose of a PRO measure is to capture the patient's experience, an instrument will not be a credible measure without evidence of its usefulness from the target population of patients."

Meehl and Cronbach (1955) advise that construct validity should be considered when measuring an attribute which is not operationally defined. Construct validity offers a framework for hypothesis testing which is based on our understanding of the constructs underlying the phenomenon of interest (e.g. quality of life). Construct validation (testing construct validity) requires firstly a conceptual definition of the construct to be measured. A conceptual definition should describe the internal struc-

ture of the concept and the relations to other constructs. Based on the definition, it should then be possible to formulate hypotheses such as which respondents should score higher or lower on that scale and what kinds of correlations should be present between related scales or sub-scales. (McDowell, 2006)

Moving from the 1960's onwards, the focus has shifted from "is this scale valid" approach, described above, towards "what do the test scores tell us about the people who took the test". The consequence of this is a change in mindset This results in validation being more like hypothesis testing and the whole concept going from validity (is this scale valid) to validation (what can we infer from these results). Following this reasoning, when content validation fails it is our inferences and not the instruments that should be considered invalid. Streiner and Norman (2008) explain that unlike other forms of validation, content validation is not related to the scores of a scale or score differences between respondents, instead it is constructed from the judgment of experts on the content of the items.

Even with the wide array of definitions of and tests for validity in the literature, one conclusion is obvious: However one defines validity, it is not possible to say that an instrument is ultimately valid. The best we can say is that it is validated within a specific group of people and in a specific context. This is important to remember when adapting an instrument from one format to another. This is discussed in more detail in the section Measurement Equivalence.

Electronic Data Capture

With advancing technology, it is becoming increasingly common to use technical devices to capture, store, and transport data in clinical trials. For example, most patient reported outcomes (PROs) are originally developed for paper-and-pen administration. If they are adopted to be filled in by electronic device, they are called electronic Patient Reported Outcomes, ePRO. This section will deal with describing what kinds of options there are for electronic data capture (EDC) of PROs and discuss relevant regulatory restrictions.

There are a number of regulatory considerations specifically related to EDC. Since it can be easier to forge large numbers of electronic data points than paper records, most of the concerns are related to the audit trail of the source data during and after the trial. Each data point recorded electronically should be associated with a data originator, i.e. the person or computer system who entered the data point into the system, as well as time stamped. The access to data entry should be limited and somehow controlled e.g. with passwords. Modifications and corrections of the data are only allowed to be done by authorized study staff. The identity of modifier and the reason for the modification/correction should be stated. All additions and modifications to data related to the trial should be tracked with time stamps. These audit trails help ensure that only authorized changes are being made. To minimize need for corrections, FDA recommends use of prompts and data quality checks prior to saving data entries. (US Department of Health and Human Services Food and Drug Administration, 2013a; International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1996)

Additionally, good data handling practices such as backup copies of all electronic records are needed.

The aim of adapting PROs to electronic administration is to produce data which are equivalent or superior in quality to the data produced by the original version (Coons et al., 2009). There is a number of potential advantages of using electronic data capture to gather PROs. Adapting existing instruments to electronic data capture can reduce administrative burden, enhance subject compliance, prevent secondary data entry errors, enable or improve usage of complex skip patterns, and reduce missing data (Eremenco and Revicki, 2010). It is worth noticing that EDC can and does also contribute to better quality data: Using EDC enables eliminating response values which are out of range by only accepting valid values. It is also possible to time stamp data entries and track subject compliance to study protocol. (Gwaltney et al., 2008)

In a clinical trial, the compliance to study protocol is paramount for the reliability of the data. With paper diaries the actual compliance can be hard to estimate: in a study by Stone et al. (2003) comparing subject compliance with paper and electronic diaries, the paper records were filled in for 90% of the designated time points but the electronic tracking of accessing the papers in the study binder revealed the true compliance to be 11%. Such low compliance risks invalidating the findings of the study. In the electronic diary, the actual compliance was 94% and it could be easily tracked without additional pieces of equipment.

Patients are surprisingly willing to use ePRO. Even though it takes longer for some patients to complete ePRO, they report thinking that it takes less time. (Eremenco and Revicki, 2010) Computerized diaries are often preferred by the patients and are considered easier than their paper counterparts. This in turn increases compliance: with computerized diaries compliance is often as high as 90% whereas some studies have documented only 20% compliance with paper diaries. (Gwaltney et al., 2008)

There are two main groups of ePRO: voice/auditory systems and screen text systems. Auditory systems are generally telephone-based and they are often referred to as interactive voice response (IVR). Screen text systems offer a digitized version of the original instrument's items and responses in a visual format. Screen text devices come in different sizes: desktop and laptop computers, tablet computers and handheld devices such as mobile phones and personal digital assistants. The devices can utilize keyboard and mouse or they can be fully touch-screen based so that commands are given by pressing the screen with a finger or with a stylus. (Coons et al., 2009)

Desktop computers are fairly ubiquitous but they are generally stationary. Laptop computers and tablets offer a great deal of screen space. This offers possibility to display the instructions/question text and response options in larger font. Handheld devices offer the advantage of being lightweight and therefore most mobile. The downside is the limited screen space on these devices, which can mean smaller fonts or questions split over several screens.

A screen text ePRO can be a native application to that device or web-based, where the user accesses the instrument via web browser. An application does not

need Internet connection to work and it can be tailored specifically to that device. However, completing instruments offline requires a separate step of sending the data to a central server and this has been known to cause challenges to study personnel and subjects. A web-based system does not require a separate step of sending data to server, but a constant Internet connection is required for the duration of filling in the instrument. Web-based instruments respond better to potential changes in study protocol in the middle of the study. (Coons et al., 2009)

IVR systems are conceptually different from pen-and-paper or screen text administration. They are automated telephone-based systems where callers interact with pre-recorded voice questions. One advantage of IVR is that the subject only needs a functioning telephone to respond to the questions. They can be used with e.g. groups with low literacy or problems with eyesight. IVR systems can record voice input and in addition, touch-tone keypad selection can be used to facilitate the completion of the instruments. The auditory presentation of the items departs from the visual medium with which most PROs are developed. (Coons et al., 2009) There is not enough research to determine whether and under which conditions visual PRO and IVR produce comparable data.

According to the golden rules of ePRO, set by International Society For Pharmacoeconomics and Outcomes Research (ISPOR) task force, the choice between different possible ePRO platforms should be made based on the target population. The complexity of data capture requirements and time frame required for subject reporting are also to be considered before committing to a platform. (Coons et al., 2009)

Measurement Equivalence

In previous section, the difference between electronic and paper administrations of a PRO measure was already touched upon. As was mentioned in section Reliability, it is important that changes in the scores of an instrument reflect actual changes and that when no change has happened in the phenomenon of interest, there should be no changes in the scores either. When adapting an instrument from one form of administration to another one, we also want to ensure that the psychometric properties of the instrument have stayed intact and that the results obtained from the new implementation are comparable with the scores from the original test, i.e. measurement equivalence. Measurement equivalence has been defined as "a function of the comparability of the psychometric properties of the data obtained via the original and adapted administration mode" (Coons et al., 2009). In this section we will consider measurement equivalence using adopting paper instruments to electronic format as an example.

There are two potential sources of difference between a pen-and-paper and a screen text instrument. One is the way the items are formatted and presented on a screen. There may be a number of adjustments that need to be made in order to computerize a questionnaire. For example, paper instruments often have several items on a single page, whereas it is common to show the items one at a time in a computerized version. Sometimes it is even necessary to split an item on several

screens because of screen size constraints. Other common changes include changing item wording to reflect the mode of administration, e.g. changing instructions from "circle on paper" to "tap on screen". (Gwaltney et al., 2008)

Besides formatting, another potential source of differences are the difficulties that some respondents might encounter when using the electronic mode. It is conceivable that a computerized instrument would be harder to complete for those respondents who have little experience in working with computers or who suffer from "computer anxiety". (Gwaltney et al., 2008)

FDA guidance paper on PRO use states, that when using a modified version of a PRO instrument, sponsors should provide evidence to confirm the new instrument's adequacy. According to the guidance changes, which can alter the way that patients respond to an instrument, include:

> - Changing an instrument from paper to electronic format
> - Changing the timing of or procedures for PRO instrument administration within the clinic visit
> - Changing the application to a different setting, population, or condition
> - Changing the order of items, item wording, response options, or recall period or deleting portions of a questionnaire
> - Changing the instructions or the placement of instructions within the PRO instrument
>
> (US Department of Health and Human Services Food and Drug Administration, 2009)

However, FDA is not explicit what kind of validation is needed to prove measurement equivalence for different types of change. Here the instrument format is of express interest and it will be discussed most thoroughly.

Following the release of FDA PRO guidance, the International Society of Pharmacoeconomical and Outcomes Research (ISPOR) established a task force to devise recommendations for testing for measurement equivalence in different forms of administration. The task force membership consisted of a heterogeneous set of related professionals with diverse backgrounds, perspectives and expertise. (Coons et al., 2009)

(Coons et al., 2009, p.422) conclude that based on the definition of measurement equivalence, "the amount of change that occurs during migration to the electronic platform/device will dictate the amount of evidence necessary to demonstrate that the change did not introduce response bias and/or negatively affect the measure's psychometric properties." To further elaborate on this point, the task force introduces categories of potential changes according to the degree of difference between the original instrument and the adapted mode of administration.

A *minor modification* is not expected to change the items and response scales. Adapting a scale from paper into screen is considered a minor modification when no changes are made to e.g. the font size, recall period, response options or item

content. Changing the display from several items on one screen/page to one per screen is also considered a minor modification. (Coons et al., 2009) Previous research suggests that these common modifications do not have a substantive effect on the results obtained from an instrument. When only minor adaptations are made to the scale, there should be no need to re-evaluate content or construct validity assuming that these were properly validated for the original scale. (Gwaltney et al., 2008) It is recommended that small-scale cognitive interviewing and usability testing be performed on target population in order to ensure that subjects are responding to the assessment items in the intended nature.

A *moderate level of modification* might introduce a subtle change in the meaning of the assessment items. Examples of changes are splitting a single item across multiple screens, changing the order of item presentation or requiring the respondent to use the scroll bar to view all response options. Changing presentation mode from visual to aural (IVR) also belongs in this category. It is advisable that the equivalence of the original measure and the modified version be formally established via statistical testing. (Coons et al., 2009)

Substantial modifications are very highly likely to change the content or meaning of the assessment. Examples of changes are removing or adding items or significantly changing item texts. When substantial modifications have been made, the question of equivalence becomes irrelevant and the modified instrument should be treated as a new measure. This includes full psychometric testing. Estimating the comparability of the scores might still be of interest for e.g. bridging scores, but main interest should lie in establishing the psychometric properties of the new instrument. (Coons et al., 2009)

Measurement equivalence between two electronic implementations or using same implementation but different device types has not yet been studied in any great depth. We will therefore take a look at paper-to-electronic transformations, since that has been an area of focus and it has provided a wealth of information, much of which can be considered to be applicable to cases where we're interested in the cross platform measurement equivalence within the same mode of administration.

In an extensive meta-analysis of measurement equivalence between paper and ePRO, it was found that the difference between electronic and paper-based data collection was minimal (0.2 per cent of scale range). Also, it seems that the size of the computer screen, respondent age, or amount of computer experience does not meaningfully influence the equivalence of ePROs. The exact changes made when adapting the instrument from one mode of administration to another was not described in detail but the authors tell that the papers included in the meta-analysis seem to be "faithful migrations", meaning that the items closely resembled each other in both modes of administration. (Gwaltney et al., 2008) This would translate to minor modifications on the ISPOR scale.

Gwaltney et al. (2008) also compared the paper vs electronic to test-retest within one administration mode where data was available. The correlation between paper and electronic scores (average 0.88, 95% CI 0.85-0.91) was very close to that of test-retest reliability of the paper measure (average 0.91, 95% CI 0.86-0.94). They conclude that even the very modest variation which can be observer between paper

and electronic administration is most likely not due to the mode of administration but instead it reflects the random variation which is seen with test-retest within one mode of administration. It is also noteworthy that the quality of data gathered by an modified instrument very strongly correlates with the quality of data gathered with the original instrument.

Testing for Measurement Equivalence

When modifying an existing instrument, one should prove that the new implementation produces results, which are equal to those of the original scale, i.e. test for measurement equivalence. The correct amount or level of testing needed in each particular case is still somewhat unresolved. The ISPOR task force is forming into a golden standard in the industry. Coons et al. (2009) point out that it is not feasible to perform full psychometric testing, as if dealing with a novel instrument, on each each electronic adaptation of each PRO instrument. The cost is prohibitive and there is potential for very little (if any) scientific gain. Likewise, it is not reasonable to expect that a simple qualitative study such as cognitive interviewing is sufficient if large structural changes have been made to an instrument.

As discussed in the previous section, the degree of modification made in the adaptation process should decide how much evidence is needed to ensure measurement equivalence. It should be noted that there is no need to prove content validity: assuming that the original questionnaire has thoroughly tested for content validity it should carry over for current adaptation. Essentially the interest lies in potential differences between different devices.

When testing for measurement equivalence between multiple modes of administration becomes necessary, there are two recommended study designs: *randomized parallel groups design* and *randomized crossover design*. (Coons et al., 2009) In randomized parallel groups the subjects are split into two groups at random, and each group will then test different implementations. In randomized crossover design, all subjects will test both implementations one after the other, with the implementation to be tested first randomized.

There is commonly a rest period or a distractor task associated with the crossover design to ensure that the subject does not respond to questions from memory but actually thinks about their answer. The rest period is usually shorter than optimal for subject convenience. That is why there is almost always a little carryover effect in crossover design, i.e. the subject can remember what they have responded before. On the other hand, in parallel groups the sample size has to be a lot larger than in crossover: Firstly, there are two groups instead of one, and secondly, in crossover design the groups are already matched (since each person acts as their own control) which also helps reduce sample size. In some cases, as little as 5% of the number of subjects in a parallel groups design are needed in a crossover design (Beatty, 2010). In practice, crossover designs are frequently favored, and in either case the required sample size needs to be calculated before commencing testing.

Recommendations for appropriate tests to establish measurement equivalence for each level of modification made by ISPOR ePRO Good Research Practices Task

Force (Coons et al., 2009) are discussed below.

Cognitive interviewing, also called cognitive debriefing, is a qualitative tool for improving questionnaire design. It consists of having subjects fill in a (draft) survey in the intended administration mode, while interviewer notes down additional information about the responses. The results can provide information about how the subjects interpret the questions, how they arrive at their answers, and any issues or challenges they had while filling in the questionnaire.

This is an important step in instrument development: due to the inherent ambiguity of language and imperfection of human memory, it is very difficult to devise effective and straightforward standardized questions. The item wording is balancing the fine line between specificity and simplicity. This is why it is advisable to test the questions or items with a sampling from the intended target population to ensure that they understand the question and understand it in the same way. (Beatty, 2010)

Cognitive interviews are qualitative in nature, i.e. no numerical data can be gathered, and they are generally conducted for a small number (5-15) of subjects at a time. (Willis, 2005) Even though cognitive debriefing is commonly considered as a tool for questionnaire design, Coons et al. (2009) argue that it can be also used to ensure measurement equivalence.

There are two alternative paradigms in cognitive interviewing: *thinking aloud* and *probing*. In thinking aloud the subject is instructed to verbalize his/her thought process while filling in the questionnaire, and the interviewer attempts to disturb the process as little as possible. In probing, the interviewer asks questions, verbal probes, from the subject as they fill in the questionnaire. These questions are designed to be as neutral as possible, e.g. "What does X mean to you, in this question?". The goal of both approaches is the same: to gather additional verbal information on subjects' cognitive processes related to the questionnaire and item wording, even though the methods are different. (Willis, 2005; Beatty and Willis, 2007)

Both methods have their supporters. The advocates of the think-aloud method point out that thinking aloud is all but free from interviewer bias, and the format is more open-ended which enables getting feedback from things that the interviewer might not even realize to ask about. The proponents of verbal probing point out that with verbal probing, the interviewer is able to better control the interchange and guide the discussion into the direction of interest, so that even areas which might not come up in a think-aloud situation can be addressed. However, it seems that a lot of cognitive interviewing is a combination of the methods: when probed, the subjects start to spontaneously think aloud to an extent and most inexperienced subjects need some probing to remember to keep thinking aloud. (Willis, 2005; Beatty and Willis, 2007)

Cognitive interviewing enables investigating some key areas of interest in subjects' response behavior. The parts of a question-and-answer process are considered to be (in order): comprehension, retrieval of information, judgment, formatting response, and editing response. Any of these parts might produce bias into the answers. In cognitive interviews, verbal probes can be targeted at the process of interest e.g. comprehension of questions. (Collins, 2003)

Since the data gathered are qualitative in nature, so also the analysis methods must be qualitative. There are several ways to conduct the analysis. Willis (2005) separate between informal analysis, i.e. writing notes during the interview and working from those, and formal analysis, i.e. taping the interview and assigning coding categories to the transcript. Formal analysis is much more labor intensive, which is why informal analysis or informal analysis with help of recording are generally used in practice.

If modifications made to the instrument are of moderate level (Coons et al., 2009, p.423) recommend that "it is advisable to formally establish the equivalence of the electronic measure." This means statistically comparing scores from the original and the modified instrument. The statistical comparison can be made using either parallel group or crossover study design.

Due to random error it is unlikely that one would get identical results when conducting equivalence testing with psychometric instruments. One needs to consider the *equivalence margin* of interest, i.e. how similar do two results need to be considered "close enough". The equivalence margin, which is commonly denoted by Δ, is often considered from a clinical point of view: the effect size which needs to be detected is often the same as MID.

Two types of error can occur when assessing the equivalence of two measures: *Type I error* occurs when we conclude that measurements are equal when in fact they are not. If we decide that two measurements are not equal, when in fact they are, that is called a *type II error*. Entirely eliminating these errors is practically impossible, so one commonly decides upon acceptable values. Jones et al. (1996) recommend using 95% confidence interval, i.e. accepting 5% possibility of type I error, and aiming for 80% or 90% power, i.e. accepting 10% or 20% possibility of type II error, respectively.

Suitable statistical analysis for randomized parallel group studies is significance testing of mean scores, with null hypothesis of equal means. The subjects can also re-fill the questionnaire after a suitable interval, which enables direct comparison of test-retest reliability across the modes. (McEntegart, 2010; Coons et al., 2009)

When estimating for the minimum number of subjects for a parallel group study, the acceptable values for equivalence margin, Δ, probability for type I error, α, and probability for type II error, β, need to be decided. Assuming that the measurement score has normal distribution, we can calculate the number of subjects needed for parallel groups trial with the following formula (Jones et al., 1996):

$$n = 2(s^2/\Delta^2)(Z_{1-\alpha} + Z_{1-\beta/2})^2 \tag{1}$$

where n = the number of subjects per group, s^2 = estimated variance of the observations (assumed to be equal in both groups), Δ = equivalence margin, α = probability of type I error, and β = probability of type II error.

Similar calculation for crossover studies can be made with the formula (McEntegart, 2010):

$$n = (s_d^2/\Delta^2)(Z_{1-\alpha} + Z_{1-\beta/2})^2 \tag{2}$$

where n = the total sample size, s_d^2 = the standard deviation of the difference between administrations and within subject. s_d^2 can also be written as $2s^2(1-\varrho)$, where ϱ is an estimate of the correlation between observations. Other variables are as above in Equation 1.

The sample size calculations as presented above are used for comparing means. For randomized parallel groups design, this is a favored method of statistical analysis. For randomized crossover design, intraclass correlation coefficient (ICC) is recommended by Coons et al. (2009). ICC compares the within-subject variance to the total variance of all subjects and all measurements, i.e. how much of the total variance is caused by variance between administration methods.

Sample size for ICC analysis can be estimated with (Walter et al., 1998):

$$n = 1 + \frac{4(Z_{1-\alpha} + Z_{1-\beta})^2}{ln(C_0)^2} \tag{3}$$

where $C_0 = \left(1 + 2\left(\frac{\varrho_0}{1-\varrho_0}\right)\right) / \left(1 + 2\left(\frac{\varrho_1}{1-\varrho_1}\right)\right)$, ϱ_0 = the lower bound for ICC, and ϱ_1 = estimated ICC in population.

There are several ways to calculate ICC (Shrout and Fleiss, 1979; McGraw and Wong, 1996). The correct way to evaluate the measurement equivalence is ICC(2,1) in Shrout and Fleiss notation and ICC(A,1) in McGraw and Wong notation (McEntegart, 2010). ICC can be calculated with (McGraw and Wong, 1996):

$$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{4}$$

where MS_R = mean square for rows, i.e. subjects, MS_R = mean square for columns, i.e. measurement methods, MS_E = mean square error, n = number of subjects, and k = number of measurements methods.

The calculated sample size can be different for ICC and comparison of means. The magnitude of the difference depends of the estimated parameters. McEntegart (2010) recommends choosing the larger of the two as the actual sample size, which will then allow for both analyses. Due to the notable difference in sample sizes, crossover design is more commonly used.

A summary of the ISPOR task force levels of modification and recommendations for establishing measurement equivalence can be found in Table 1. Coons et al. (2009) recommend including usability testing as a part of the testing for all levels of modification. This refers to the ability to complete an instrument in its modified form by a member of the target population.

It is important to remember that both validity and reliability are measures, which cannot be ultimately proven for an instrument. Instead, they need to be evaluated for a specific population, in specific context and for a specific purpose. From this, it is possible to extend the reasoning onward, e.g. if a questionnaire is validated for 15-20 year old females in New York City, it can most likely be considered also validated for 21 year old women in the same area, even though they were not part of the original population, but might need further validation for 50-55 year old men in rural Finland. Drawing the line of when something is similar enough or too different

Table 1: Instrument modification and evidence needed to support measurement equivalence, adapted from (Coons et al., 2009)

Level of modification	**Rationale**	**Examples**	**Type of testing**
Minor	The modification can be justified on the basis of logic and/or existing literature. No change in content or meaning.	1) Nonsubstantive changes in instructions (e.g., from circling the response to touching the response on a screen). 2) Minor changes in format (e.g., one item per screen rather than multiple items on a page).	Cognitive interviewing Usability testing.
Moderate	Based on the current empirical literature, the modification cannot be justified as minor. May change content or meaning.	1) Changes in item wording or more significant changes in presentation that might alter interpretability. 2) Change in mode of administration involving different cognitive processes (e.g., paper [visual] to IVR [aural]).	Equivalence testing Usability testing.
Substantial	There is no existing empirical support for the equivalence of the modification and the modification clearly changes content or meaning	1) Substantial changes in item response options 2) Substantial changes in item wording	Full psychometric testing Usability testing

is a very difficult question and should be seriously considered while planning for validation of measurement equivalence.

Bring Your Own Device

Bring your own device (BYOD) is a model where users, generally employees, are encouraged to use their own technology in a work setting. The same term is used in education when students are allowed to use their own computers in teaching setting. This model, while relatively new, is rapidly becoming more and more common. A significant amount of population in the developed countries own Internet enabled devices and have Internet access at home. In 2011 roughly 70% of U.S. households had Internet connection and the percentage has been steadily increasing in the 21st

century (United States Census Bureau, 2013). In Europe Union on average 79% of households had Internet connection in 2013.(Eurostat, 2013) Some member states had even higher connectivity; e.g. in Finland 89% of the households have Internet at home. (Eurostat, 2013; Suomen virallinen tilasto, 2013)

The growing popularity of BYOD is often attributed to the selectivity of many users. With more than 5 Internet enabled devices in every U.S. Internet household (The NPD Group, 2013), the users often have very specific wishes as to which setup or operating system they wish to use. In business context it is commonly thought that employees are more productive when they get to choose the equipment they work with than when the employer chooses for them.

In the context of clinical trials, BYOD is a very new phenomenon. It has been widely discussed in the industry, but partially due to lacking information about necessary forms of validation, it has not been widely used. The massive costs associated with clinical trials make it less likely for sponsors to take any risks in proving their endpoints, thus slowing down the speed of adapting new practices.

One of the big concerns regarding BYOD in general is information security. (Thomson, 2012) The devices, which are used are not nearly as well protected as would be suitable for the data which is accessed by them. This is especially true for use of BYOD by health care providers. Change towards BYOD is happening rapidly and it remains to be seen whether the benefits outweigh the risks and costs. (Moyer, 2013)

In the project discussed in this thesis, information security issues, such as user identification and data transfer, are dealt with according to CRF Health's quality management system and in accordance with FDA and EMA regulations. More detailed aspects of information security are considered trade secrets and out of the scope of this thesis.

There is also a variety of practicalities to consider with BYOD, such as how many back-up devices should be allocated to each site for those users who don't have their own, how to provide help desk when the range of potential devices is innumerable, and how to pay for the data transfer costs while ensuring subject anonymity. These will be further discussed in the Results chapter of this thesis.

Currently, there is no set methodology to prove measurement equivalence over a range of devices. The methods adapted in this case are further discussed in the section Validation Study Design. Further methodological development and suggestions are presented in chapter 5 Discussion.

3 Materials and Methods

In this chapter we will go over the practical scope of the project to which this thesis is related. For that end, we will introduce the Amsterdam IADL Questionnaire® and CRF Health. We will also describe the validation process which was carried out in collaboration with them to evaluate the project.

Amsterdam IADL Questionnaire®

The instrument used in the practical part of this thesis is targeted at detecting early symptoms of Alzheimer's Disease (AD). As cognitive decline proceeds, AD patients find it noticeably harder to cope with activities of daily living. In the early stages of the disease, patients encounter difficulties with more complex everyday activities, such as cooking or doing finances. Instrumental activities of daily living (IADL) are those activities of daily living, which are required for a person to function independently in a society. As the disease progresses, patients tend to also experience problems with Basic Activities of Daily Living (BADL). BADL include very basic self-care skills such as eating and washing oneself.

Amsterdam IADL Questionnaire® was developed by Sietske Sikkes after she noted that patients' coping with IADL was a good predictor of AD onset, but the available measures of that time to assess this were insufficient (Sikkes et al., 2009).

Due to the nature of the disease, patients do not fill in the questionnaire themselves. Instead, their primary caregiver (the observer) fills it. The caregiver is most likely a spouse or a relative who lives with the patient. The instrument is an observer reported outcome (ObsRO) and not a proxy-filled patient reported outcome since the caregiver is not answering on behalf of the patient but about the patient.

Amsterdam IADL Questionnaire® is a native computer-based instrument and it utilizes sophisticated skip patterns enabled by computer administration. There are 70 items altogether in the questionnaire. The skip patterns ensure that each caregiver only needs to answer those questions which are relevant to the patient in question, e.g. if the patient does not use a computer, the caregiver is not asked more detailed questions about computer usage. The questions are split to two levels: main questions and follow-up questions.

The main level question is a "yes/no/don't know" question about an activity the patient might have performed during the recall period (four weeks). The follow-up questions adapt to the answer of the main level question. If the answer to the main question is yes, the caregiver is asked if it was more difficult for the patient to do the activity. If the answer is no, the caregiver is asked for their assessment about why the patient didn't do that activity.

The data gathered from observers is limited to objective, observable phenomena, such as activities or mobility. This adheres to FDA regulations, which stipulate that observers should only be asked about those events or behaviors, which can be observed (US Department of Health and Human Services Food and Drug Administration, 2009).

The results are calculated using Item Response Theory, which enables taking

into account the difficulty levels of the activities covered by the instrument.

CRF Health

The project and the thesis were done for CRF Health. CRF Health is a software company, which provides electronic Clinical Outcome Assessment (eCOA) solutions for life sciences. The company was founded in 2000 in Helsinki, Finland by the name CRF Box and in a relatively short time it has risen to become a global leader in eCOA. CRF Health has grown rapidly and it currently has offices in Philadelphia (USA) and in London (UK) in addition to Helsinki. According to their company agenda, CRF Health is dedicated to bringing solutions that fit right into the users' lives and is constantly working on improving the product family. The work outlined in this thesis was mainly done in CRF Health Research & Development Center of Excellence in Helsinki.

The solution outlined in this thesis is built on TrialMax Web™ eCOA solution by using the company's internal instrument development tools. The user interface of TrialMax Web™ works with all modern web browsers, and thus the system can be used on almost any Internet capable device as long as the device is online.

Some crucial issues in eCOA, such as data transfer and safety, are already built into the TrialMax® product family and were therefore left out of the scope of this Thesis. In accordance with CRF Health company agenda, committed to meeting all regulatory requirements for clinical trial data while driving global improvement in outcome quality and efficiency in paper-free clinical trials, this thesis aims at providing a process for late-phase clinical trials for more efficient, user-friendly, and more cost-effective eCOA capture. The solution presented here is heavily building on the TrialMax™ suite and the possibilities and limitations of the TrialStudio™ development environment.

Scope of Work

The practical scope of the work consists of making an implementation of the Amsterdam IADL Questionnaire® to be used on both tablet and desktop computers. The instrument selection was based on business decisions, and the suitability for cross platform use was a request from the instrument author.

The implementation will be used by the instrument author, Dr. Sikkes, as well as by other researchers and clinical professionals. The questionnaire is first filled in at a hospital or memory clinic on a tablet computer. If need be, follow-up will be done at patients' home on their own device, most likely a laptop or desktop computer.

While designing the user interface, the second research question "How should instruments be designed to ensure measurement equivalence across a range of devices?" was carefully considered. TrialMax Web™ was chosen as the platform because, out of the existing members of TrialMax™ product family, it is best suited for cross device administration. The implementation was made using CRF Health's proprietary development tool, TrialStudio™.

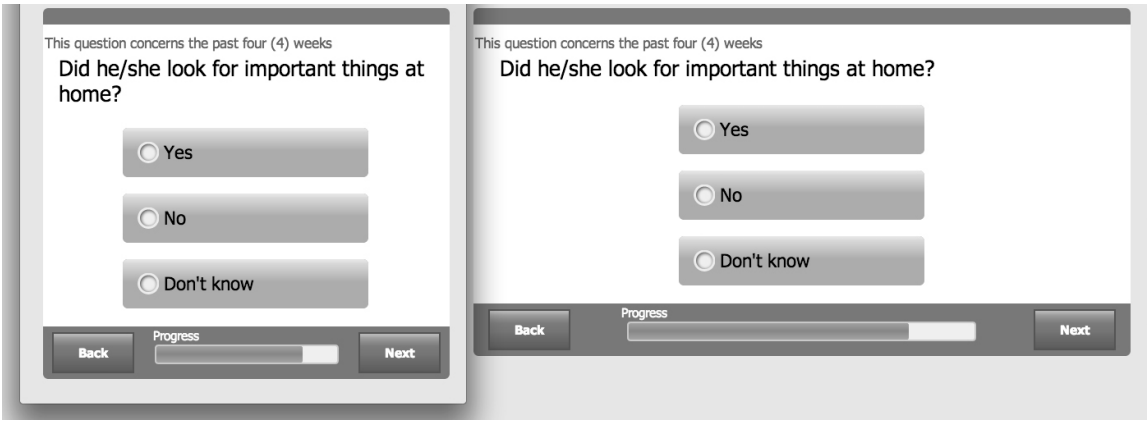

Figure 1: Screenshot of fluid layout displayed in two browser windows: one on the right is the smallest size which does not require scrolling and window on the left is the largest size.

Most emphasis in the development process was put on ensuring layout equivalence across devices while aiming for a tailor-made user experience, which would not have the feel of being designed for another device. The goal was to reduce or eliminate variability in the psychometric properties of the instrument and thus increasing measurement equivalence. As part of the project, the author attempted to assess to what extent this is possible with the required tools. Generalizability of the results was also taken into account, and solutions deemed too complex or labor intensive to use in fast-paced client projects were excluded if possible.

The implementation of the instrument was done in an iterative manner and each new iteration was demonstrated to Dr. Sikkes. Several of these iteration rounds were done by email or a web conference with shared display. The penultimate design was demonstrated in person to Dr. Sikkes and other staff members at Amsterdam VU Medical Center Alzheimer Center and the feedback from that round was incorporated into the final design.

According to Dr. Sikkes' wishes the design took into consideration the target user group, who are elderly and likely to have limited amount of computer experience. In practice, designing for the elderly users meant sacrificing the equal size of response buttons in favor of largest possible font that would nicely fit even in smaller tablet screens. Guidelines for designing clinical systems for elderly users (Demiris et al., 2001) and tenets of inclusive design were followed wherever possible.

The scope of likely devices was limited to desktop, laptop, and tablet computers, with an active decision to exclude devices with a small screen (i.e. mini tablets and smartphones). The original implementation of the instrument had also excluded small screen devices from the scope, and according to the instrument author this has not been problematic. The proportions of the layout were made fluid, so that the layout would seem tailored both for a portrait mode tablet as well as traditional landscape mode screens. The layout can be seen as a screenshot in Figure 1 and in in context in Figure 2.

A minimum size was fixed, so that if viewed with a less than 9" screen, the

user would see a 9" screen version of the layout. This choice was made due to the exceptionally large font size, which was chosen to accommodate elderly respondents. Different user groups and the limitations imposed by them are further discussed under Chapter 5, Discussion.

Validation Study Design

The third research question, "What kinds of validation are necessary to prove measurement equivalence of a multi-platform implementation of a PRO instrument?", was approached by testing the implementation on users.

As already discussed in section Testing for Measurement Equivalence, there are two recommended approaches to testing two versions of the same instrument. If the change between original and modified instrument can be considered minor, then cognitive interviewing can be sufficient. If the difference is moderate, i.e. larger than minor but not so substantial that it would necessitate considering the modified version as a completely new instrument, statistical validation is recommended.

The line between the above-mentioned two magnitudes of modification is not clear and therefore it can be difficult to choose the suitable method of testing for measurement equivalence. In order to conduct statistical testing, we first need to calculate suitable sample size to achieve desired statistical power.

For parallel groups design, we can use Equation 1. The parameters to be decided are desired confidence interval, statistical power, standard deviation Confidence interval is commonly set to 95%, which would mean $\alpha = (1 - 0.95)/2 = 0.025$. This gives us $Z_{1-\alpha} = Z_{0.975} = 1.96$. β is often set to 10% to give power of 90% or 20%, to give power of 80% (Jones et al., 1996). Selecting $\beta = 10\%$ gives us $Z_{1-\beta/2} = Z_{0.95} = 1.65$. We do not know numerical values of standard deviation or MID of the Amsterdam IADL Questionnaire®, but both Coons et al. (2009) and McEntegart (2010) use an example of 0.3 standard deviations as the equivalence margin, i.e. $\Delta = 0.3s$. Inserting these values into Equation 1 gives us

$$n = \frac{2s^2}{(0.3s)^2}(1.96 + 1.65)^2 = 289.60 \tag{5}$$

i.e. there would need to be at least 290 subjects in both groups, giving a total subject count $n_{tot} = 580$.

For comparing the means with crossover design, we can use Equation 2. Using the $s_d^2 = 2s^2(1-\varrho)$ substitution, we see that $n = [(1-\varrho)/2]n_{tot}$, where ϱ denotes the estimate of correlation between the two modes of administration. We do not know the actual value of ϱ for Amsterdam IADL Questionnaire®. In the meta-analysis of paper-and-pen versus electronic administration by Gwaltney et al. (2008), the average value was 0.9, which is why we choose $\varrho = 0.9$. This gives us

$$n = [(1 - 0.9)/2] * 580 = 29 \tag{6}$$

i.e. in crossover design, altogether 29 subjects are needed. Selecting a different ϱ will change this number, e.g. selecting a more conservative $\varrho = 0.8$ would double the sample size.

As was mentioned earlier one can also analyze results from crossover design by using ICC. For estimation of sample size for ICC we need to decide the ϱ_0, i.e. the lower bound for ICC, and estimate the value of ICC in population, ϱ_1. Coons et al. (2009) use $\varrho_0 = 0.7$ and $\varrho_1 = 0.85$ as examples. Using these, and $\beta = 10\%$ and $\alpha = 0.05$ as in the above calculations, we get

$$C_0 = \left(1 + 2\left(\frac{0.7}{1-0.7}\right)\right) / \left(1 + 2\left(\frac{0.85}{1-0.85}\right)\right) = \frac{17}{37} \tag{7}$$

Inserting this into Equation 3 gives

$$n = 1 + \frac{4(1.64 + 1.28)^2}{ln(\frac{17}{37})^2} \approx 57.4 \tag{8}$$

thus 58 subjects would be needed with these parameters. Keeping other parameters the same, but aiming for 80% power (as opposed to 90% in the equation 8), the sample size would be 43. For crossover design, the larger of ICC and means comparison sample sizes is recommended. In this case, the sample size would then need to be 58 subjects altogether, assuming that the selected parameters are valid.

In cognitive debriefing, the decision of sample size is based on heuristics. Coons et al. (2009) recommend 5-10 subjects. The validation work described in this thesis consists of cognitive interviews. Qualitative validation was chosen based on practical considerations. The implications of this choice are discussed in Chapter 5, Discussion.

In validating the questionnaire the independent variable was considered to be the screen size and screen proportions. In addition to these, there is a theoretical possibility that things like the operating system, device type (smart phone, tablet computer, laptop computer or desktop computer), and screen manipulation method (touch screen or mouse) might influence the responses.

In validating the questionnaire, the independent variable was considered to be the screen size and screen proportions. In addition to these, there is a theoretical possibility that for example the chosen operating system, device type (smartphone, tablet computer, laptop computer or desktop computer), and screen manipulation method (touch screen or mouse) might influence the responses. In an optimal case, all of these variables would be individually manipulated to see actual effects. Each pair of variations should have been tested qualitatively with at least 5 subjects or quantitatively with approximately 30 subjects, if using crossover design. Unfortunately, using a sample size this big was impossible due to resource limitations. To maximize the possibility of detecting differences, the author chose to use systems as different from one another as possible. It is possible, although unlikely, that the effects from some of these variables might have reversed each other.

Having two implementations tested back to back (even with a distractor task) is bound to influence the latter administration (carryover effect).McEntegart (2010) recommends the duration between administrations to be between a day and a week. The duration depends on the type of measurement: in order to test for measurement equivalence, the underlying condition should not change between the administrations.

Initially, the author hoped to find subjects who would come in on two consecutive days. The Amsterdam IADL Questionnaire® has a recall period of 4 weeks and therefore it is unlikely that many of the answers would have changed in a day. The company responsible for the recruitment of users gave their professional opinion that it would be impossible to find participants who would come in on two different days.

Durations shorter than a day between administrations can also be found in the literature. When testing for a scale for motivation to eat,Whybrow et al. (2005) had the subjects fill in pen-and-paper and electronic implementations right after one another. Chen et al. (2007) used a duration of "at least 10 min" between pen-and-paper and electronic administrations of a quality of life scale. This does not mean that a very short break would be recommended, but it is not unheard of either.

After consulting literature and discussing with the user researchers in the recruiting company, the length of the session was set as 3 hours, to allow for a longer subset of questions and a more substantial distractor task or lunch break in between the administrations, which hopefully would have helped to improve the quality of the data. Evidently, this was asking too much of the potential subjects and not a single subject was found in over a month's time.

Finally, in order to have any subjects at all, a 1,5 hour setup was used, with a 10 item subset and just 30 minutes between the implementations. This is much shorter than what would have been optimal, but 30 minutes did allow for a distractor task to be used.

Study Implementation

Measurement equivalence was measured by conducting cognitive interviews with potential end users. For the interviews the users answered the same questions using the 9,7 inch iPad (3rd generation) and Dell-laptop with an external 20.1 inch display (Dell 2007FP). On the computer, the operating system was 64-bit Windows 7 Enterprise and the browser was Firefox 24.0. On the iPad, the operating system was iOS 7.0.4 and browser was Safari. The interviews were conducted in a user research room set up for this purpose and by the author. The interviews and on-screen activity were recorded using Camtasia Recorder and an external microphone. The subjects were allowed to hold the tablet computer in the way most natural for them, most (n=5) opted for horizontal orientation. The experiment set-up is displayed in Figure 2.

The subjects (N=6, 2 men) filled in a 10-item subset of the Amsterdam IADL Questionnaire®. The subset consists of items number 6, 7, 8, 9, 22, 24, 25, 30, 68, and 69 in the instrument. The subjects were first introduced to the thinking aloud process by a practice question ("Please tell me how many windows there are in the house or apartment where you live") and then instructed to think aloud while processing and answering the questions. The interviewer also used classical cognitive interviewing probes, e.g. "How did you arrive at your answer?", to aid the thinking aloud process. The probes were mostly directed at comprehension and retrieval of information. This mixing of think-aloud and verbal probing is common practice in the industry (Willis, 2005).

Figure 2: Experiment setup. Clockwise from far left, tablet, computer screen, external microphone, mouse. During interviews, subjects saw the instruments one at a time, here displayed side by side for the benefit of the reader.

The subjects first filled in the subset of the questionnaire on one device. This lasted for about half an hour (25-40 minutes). Then, the subjects rated emotional sounds for the following half an hour as the distractor task. Finally, they returned to the same subset of the questions on the other device. Half of the subjects filled in the questionnaire first on a computer and then on the tablet and the other half in reverse order.

Inclusion criterion for the subjects was fluency in English. All of the subjects had prior experience in using a computer, though not all were familiar with touch screen devices. All of the subjects were also living together with a spouse, who was over 60 years old. The profile of the interviewees closely mimics the demographics of the original validation study conducted in the Netherlands (Sikkes et al., 2010). A summary of subject information can be found in Table 2.One important difference was the language used, which was Dutch in the original implementation. The translation of the questionnaire had been done by professional interpreters prior to the onset of the project and was used as-is.

The prior expectation was that there would be no clear distinction between the devices but that there would be some differences between first and second administrations. The differences between the two interviews were expected to be similar to test-retest variability within the same method of administration. Given the short

Table 2: Ages of subjects and spouses, plus first administration method

Subject Code	Age of Subject	Age of Spouse	Which Device First
S1	66	75	computer
S2	52	61	tablet
S3	57	66	computer
S2	67	62	tablet
S5	65	65	computer
S6	64	67	tablet

break between the administrations, it was also expected that there would be some recall effect, which would be verbalized during the second administration. It was hoped that including a distractor task between the administrations would dampen the recall effect.

Analysis

In survey design, the goal of analyzing cognitive interviews is to discover difficulties that the subjects have with understanding the items. In this project the aim was to compare two administrations. While testing for measurement equivalence, it does not matter whether the subjects understood the questions the way the instrument author had meant them, however they consistently need to understand the questions the same way.

The interviewer took notes during the interviews and these were utilized in the analysis to detect potential interesting patterns. The interviews were informally analyzed by following particular themes in the recordings and writing down key words and key phrases of each answer. These themes were focused on comprehension and information retrieval parts of the question-and-answer process. No conclusive transcripts of the records were made. This is the industry standard: the gains from transcribing whole interviews are generally not considered substantial enough to justify the workload (Willis, 2005).

The items included in the subset were selected so that there were both questions which the author considered to be fairly straightforward ("Did he/she use cash?") and those which could have more variable interpretations ("Has he/she experienced unexpected circumstances?"). Each question in the 10-item subset was separately inspected both within subject and across subjects. Particular attention was paid to test-retest variability and whether or not the order of devices might have any contribution to that.

4 Results

The main result of the thesis is the actual implementation of the instrument. The project showed that a cross device implementation was indeed possible to do for the TrialMax Web™ platform using CRF Health proprietary development tools, and that the cross device implementation work was not too labor intensive or slow even for client projects. Design recommendations based on insights gained from the practical development work and literature are presented in Chapter 5, Discussion.

The user testing done in the scope of this work should be considered a pilot rather than an exhaustive validation. As explained in the previous chapter, the cognitive interviews were analyzed using qualitative methods. The interviewees comprised too small a sample and they filled in too small a subset (10 out of 70 items) to say anything statistical about the data or the differences between the administrations. In the full version of Amsterdam IADL Questionnaire®, the scores would be scaled using item response theory, but the 10 item subset was considered too small to give a properly scalable score.

The differences between subjects were far more pronounced than within subject variation. The two administrations varied slightly in a mostly predictable pattern. Most of the variation arose between the first and second interview, regardless of the order of devices. The second round of interviews was on average 5 minutes faster than the first one. This might be caused by the subjects being more familiar with the thinking aloud method. The subjects spontaneously displayed effects of recall effect with phrases such as "Yes, I answered yes and I answer yes now" (S1, Table 2) and "Oh, we were discussing this [earlier]" (S1, Table 2). The effect was more pronounced in some subjects, but five of the six subjects brought up the previous administration at least once during their second interview.

Five out of the six subjects did respond to at least one question differently in the second round than they had in the first round. The differences generally reflected the discussion that they had had with the interviewer. For example, the interviewer would probe about what is included in a task, and after listing out the aspects they considered, the subjects might change their answer. "It [is] tempting for me to, to answer 'No' but let's say, I could answer more precisely and according to the actual situation, I put 'Yes, more difficult' but [it's] because of her physical [condition]' (S5, Table 2).

The differences were small, mostly shifting between "no" (i.e. not more difficult than before) and "slightly more difficult" or "slightly more difficult" and "more difficult", which would represent 1 point difference in the raw scores. There was no clear device-related direction of the change, i.e. there were approximately as many harsher judgments as there were milder judgments on both devices.

The cognitive interviews gave valuable insight to the instrument and its English language version. The subjects made a number of comments on the general content and wording of a number of questions. This was not related to the actual goal of the interviews, but nonetheless provided valuable information. For example, two of the six subjects did not seem to pay any attention to the four-week recall period, which was mentioned twice in the instructions and displayed on every main question page.

The interviews also functioned as light usability testing, even though the interviewer was quicker to help the subjects than would have been ideal in usability research.

Some of the subjects were fixated only on the cognitive difficulties and cognitive capability of their spouse, after hearing that the instrument was intended to detect AD. A lot of subjects also answered based on an idealized situation, and explained at some point that even though their spouse had had a major operation or had been in bed rest during the recall period, they theoretically could do a lot of things. This effect was more pronounced on the male subjects.

Interestingly, there were some items that brought on comprehension difficulties to most subjects, causing them to ask for interpretation from the interviewer. This, from a cognitive point of view, should most clearly predispose to variation. Since the thought processes regarding and interpretation of unfamiliar terminology is not automated, these may most easily be influenced for instance by layout.

As an example, one confusing item used the term "sandwich meal". This is not a common concept in Finland, so all but one subject reacted to the item by first asking for a definition to the term. When the interviewer declined to give an explanation and then asked for the subjects' own interpretations, the responses varied greatly. Some took it to mean sandwiches eaten at a specific time of the day (especially at mealtime), some subjects said that it was a more substantial or a particular type of sandwich, and one subject considered it a synonym for any type of light lunch. Notably, the interpretation remained the same for each subject over both administrations. This might be partly attributable to the carryover effect.

As expected, the less specific questions elicited more inter-subject variability in the answers. However, the within subject test-retest comparison revealed no such effect. Occasionally, a subject would change their interpretation of the meaning of the question while thinking aloud. In two subjects, this happened with the phrasal verb "look for" in one particular question, the subjects would fluctuate between "look after" and "search for" as their interpretation. Interestingly, during the latter question round, they would interpret the question in the same way that they did in the beginning of the previous round.

Also, during the second administration round, the subjects often were more analytic in their explanations and they used broader terms whereas during the first round, the subjects generally used more concrete examples. One subject answered a probe to define unexpected circumstances first as a list "Something happens to our children, somebody drives in our car, like an accident . . . [or] something happens to his family." (S2, Table 2) During the next round, the subject defined the term without a probe. The subject used more general terms but they were in keeping with the earlier answer: "It sound so ominous, I would say it is an accident or that kind of thing that [is] so out of the ordinary" (S2, Table 2).

The subjects might even have completely opposite views of how to answer the questions. In the follow-up question to the one judging the patient?s need to look for important things at home, one explained that "no, [searching for things] has been going on for five years now . . . but it is not more difficult" (S4, Table 2) whereas another subject commented, on the same question, "yes, it is more difficult for him . . . it is nowadays more often" (S3, Table 2). So here one subject considers frequency

of a behavior as a sign of trouble, whereas another one thinks that even though the frequency has increased, it does not mean that their spouse would find the action more difficult. Notably, both subjects were consistent with their interpretation on both administration rounds.

The subjects' attitudes also somewhat influenced their answers. One subject considered the options of difficulty levels, and concluded that "he's not there yet" (S2, Table 2), i.e. she is expecting this kind of behavior to occur at some point but has not noticed in her spouse. Another subject was very relative in her evaluation, looking for anomalies. Instead of answering the question based on observed behavior, she would compare her spouse to other people, and base her decision on that "he has done it . . . everybody in every age do [search for things] but not unusually so [I'll answer] no" (S6, Table 2), i.e. she would disregard events which were expected. These both traits remained constant within the subject.

Filling in the questionnaire subset with an iPad was for some of the subjects their first encounter with a tablet, and even the first time using any touchscreen device for one. This caused some usability challenges in the beginning of the tablet administration to these subjects, as they simultaneously had to learn technical matters, for example to tap the screen slowly enough. There was no discernible effect on the answers from this. The only comments the subjects made relating to the used devices stemmed from slow Internet connection and resulting slow page loading of the tablet.

The results from the cognitive interview are mostly centered on the language and comprehension of items. These findings are not conclusive enough to give sufficient proof for or against measurement equivalence.

5 Discussion

The results from the cognitive interviews revealed that the differences between subjects were far greater than the differences between the two modes of administration.

In this chapter, the author discusses possibilities and challenges revealed by the project. First the development project, validation and results gained are critically evaluated. Then, the plausibility of using BYOD model in clinical trials from regulatory perspective is further assessed. Here, the required levels of validation for a multi-platform implementation of an instrument will also be considered. Lastly, practical issues which will likely arise when BYOD is used in a trial are discussed.

Evaluation of Work

The decision to use Amsterdam IADL Questionnaire® was based on the needs of CRF Health. The practical implementation presented in this thesis was constructed in accordance with the requirements of the chosen project and within its constraints.

Several design decisions made in this project were targeted at the actual user population. Especially the choice to use larger fonts instead of having uniformly sized buttons is open for debate. The font limitation is not unique to this study and will be encountered with other studies and instruments where the subjects are likely to be elderly. The choice to use large fonts was made keeping in mind that the instrument users would be elderly people and that the instrument is observer reported outcome instead of a patient reported outcome. PROs may include items containing questions on more subjective attitudes and feelings, which in turn may be more prone than ObsROs to bias caused by for example layout and administration method of the questionnaire. Since items in ObsROs concern more concrete, objective and easily observable things such as function and behavior than items in PROs, the present findings are not directly applicable to projects using PROs items. Neither should all choices made in this thesis project thus be considered as general guidelines for similar projects on PROs and in all age groups.

The introduction of BYODs into clinical trials by sponsor companies will most likely take place in trials with younger subjects who most likely will also have more ready access to and be more comfortable with electronic devices than the elderly subjects of this project. For younger subjects, to minimize bias respond options of uniform size should be used instead, as it has been shown that scales with variable size responses can have effect on scores (Smith, 1995). This would be especially important if PROs are studied, due to the more subjective nature of PRO items.

The validation work outlined in this thesis is of qualitative nature. Several points may have affected the outcome of the cognitive interviews and should be taken into account when evaluating the results.

In the performed cognitive interviewing, the subjects used both a tablet and a computer. According to the hypothesis, if noticeable discrepancies occurred between administrations these would most result from the different screen sizes and proportions. However, also the method of manipulation (touchscreen or a mouse), operating systems and browsers were different between administrations. From the

literature, it is not clear whether these additional potential caveats can have an effect on the psychometric properties of an instrument. It was however expected that by creating two extreme examples, it would be possible to mimic a true BYOD situation and detect potential differences. No differences that would have affected measurement equivalence were detected. It is possible, although unlikely, that some of these properties may counteract each other. Having only six subjects for initial testing of a theory is not sufficient to exclude the possibility of interpretative differences between the two modes of administration. Further quantitative work is recommended to gain conclusive evidence. Due to the small sample size, different results might have been detected with a different sample of subjects.

Recruiting voluntary subjects who fulfilled the inclusion criteria turned out to be extremely challenging. Due to this, the design of the validation study had to be modified from the initial idea.

The duration between the two consecutive administrations was shorter than optimal, as explained in Chapter 3, Materials and Methods. The recommended time between two administrations is one day or more, whereas we were constrained to use 30 minutes. This led to most of the subjects verbalizing presence of carryover effect when thinking aloud for the latter implementation. This leads to the conclusion that the duration of 30 min was too short. However, the subjects did verbalize their thought processes slightly differently the second time around, as was explained in Chapter 4, Results. It is unclear to what extent the subject genuinely forgot what they had answered the first time around, how many of these differences were affected by the interview process and the verbal probes used by the interviewer.

Additionally, having only six subjects for initial testing of a theory is not sufficient to exclude possibility of differences in interpretations of the two modes of administration. Further quantitative work is recommended to gather conclusive evidence.

Since the English cognitive interviews were conducted in Finland, finding suitable subjects was not a trivial task. None of the subjects were native speakers of English, even though they all professed themselves as fluent in English. The actual language skill levels (as evaluated by the author) seemed to vary from barely conversational to fluent. Subjects' often seemingly limited vocabularies might have had an impact on their answers. In particular, some of the subjects either misunderstood or requested a translation of some of the words (especially the phrasal verb "look for") in the questionnaire. It is also possible that some of the subjects could not find the words to thoroughly express potential slight changes in their interpretations, i.e. their language skills didn't allow for discussion in a sufficient level of granularity.

The author is not a native speaker, either, so there is a possibility that the probe questions or instructions unintentionally suggested a particular interpretation. In addition to this, the author is not particularly experienced in cognitive interviewing. Willis (2005) states that cognitive interviewing is a skill, which evolves over time. The lack of relevant experience might thus have had an effect on both the interviews and their analysis.

Future work to establish a routine way for validation in multiple devices simultaneously is required. First, however, sufficient sample size and the usage of different

instruments on a variety of devices are needed to perform in-depth statistical analyses of possible differences between devices or device groups. Power calculations for statistical analysis between two versions are presented in Chapter 3, Materials and Methods.

Another option is to allow subjects to fill in the study instruments on their own devices. This would reveal the actual range of devices likely to be encountered with BYOD, but would hamper with pre-study estimates of necessary sample sizes when devices are not known in advance. It would also be impossible to perform proper randomization if statistical validation of BYOD were performed using actual devices already owned by study subjects. These may not follow random distribution and would likely differ from geographical region to another. A very thorough controlling for confounding variables would need to be performed, but such analysis would mimic real-life scenarios closely. If no discernible differences arose, this would help to convince different regulatory authorities and the scientific community of the reliability of results obtained with the BYOD model.

Plausibility of BYOD in Clinical Trials

A central aim of this thesis was to gauge and appraise the possibility to adopt BYOD model to clinical trials for those patients who have their own web-capable device(s). The question is essentially dual in nature.

Firstly, it needs to be established *if* BYOD can be used in clinical trials. In practice, demonstration that the regulatory requirements for eCOA can be met with a BYOD model needs to be achieved and that the scores obtained by cross platform PRO instruments are as reliable as their single device or paper counterparts. This hurdle to prove measurement equivalence is of the essence, and theoretically only needs to be overcome once. Secondly, a number of practical issues need to be considered concerning *how* BYOD can be used in clinical trials. Unlike measurement equivalence, the proof of which can then be used by all of the industry once it has been established, the practical issues need to be considered separately for each trial. These will be discussed in more detail in the following section, Practical Considerations of BYOD in Clinical Trials.

It is important to note that most of the regulations relating to clinical trials are under the responsibility of the sponsor company. CRF Health, a software company, funded this thesis; therefore a software development point of view is used when considering the relevant regulations. The regulatory requirements for using BYOD in capturing PROs are twofold. First, the demands on information security and audit trails need to be covered. These have been previously discussed in Chapter 2, section Electronic Data Capture. These regulations are relevant to all electronic data capture in clinical trials, with some issues that require extra attention when discussing BYOD. Second, both cross platform instrument validity and validation need to be proven. This is only relevant for those PRO measures which use standardized, validated instrument. We will discuss these two separate aspects below.

In the relevant guidances, such as Guidance on Good Clinical Practice as formulated by International Conference on Harmonisation of Technical Requirements

for Registration of Pharmaceuticals for Human Use (1996) and Investigational New Drug Application, which is part of Code of Federal Regulations in the United States (21CFR312, 2013), the key areas of interest concern the audit trail. Essentially, each bit of data needs to be accompanied with information about who has created it and when, who have changed it and when (if changes have been made) and all values it has had. Also, sponsors should never be granted exclusive control over eCOA source data. It is also notable that even if an ePRO application is designed for mobile devices, it is not a Mobile Medical Application as seen by US Department of Health and Human Services Food and Drug Administration (2013b).

Also, regulatory approval of ePRO data gathered with subjects' own devices needs validation of relevant instruments to all potential devices. This is especially important to those PRO measured with validated, formal instruments.

As discussed above in Chapter 2 in section Measurement Equivalence, modifying an instrument, including going from pen-and-paper to electronic administration, requires one to demonstrate that the psychometric properties of the original instrument have not been affected. It may be assumed that this logic, to a certain extent, is applicable to BYOD model where questionnaires are being deployed on a wide range of similar devices with slightly different properties.

Measurement equivalence is not crucial only for regulatory reasons, but also in order to gain usable data in clinical trials. It is important to keep in mind that in some disciplines of medicine a discernible increase leading to clinically meaningful treatment benefit may be very small, thus introducing any unnecessary bias into measurements is generally avoided. This should not prevent using BYOD in clinical trials, however measurement equivalence needs to be demonstrated comprehensively. Essentially, measurement equivalence between BYOD modes needs to be demonstrated in a way that would be accepted by the regulatory authorities and the scientific community. The industry is still debating on the suitable types of validation even for more common types of modification, such as paper to electronic transformation. The suitable type of validation for BYOD has not yet been established.

The author recommends a thorough, quantitative validation study to be conducted with a few much-used instruments. Ideally, it would be advisable to individually vary different factors, such as device type, screen size, operating system, and manipulation method (touchscreen / mouse). Due to practical hurdles, a more realistic option might be choosing three extreme examples, such as very small mobile phone, a medium tablet, and a large desktop computer. Like discussed above, an option would be to actually let the subjects use their own devices in the validation study, which would be closest to practice and reveal the actual range of devices to be expected. As a caveat, subjects cannot be effectively randomized. A cost-effective way might be to use these two approaches in tandem in a proof of principle study.

After the plausibility of BYOD model has been demonstrated quantitatively, more concrete predictions about the measurement equivalence in BYOD can be made. Hopefully, after a successful validation, a less exhaustive testing regime for measurement equivalence might be sufficient (cf. Coons et al. (2009)).

If BYOD were adopted in clinical trials, potential benefits might be substantial.

Since the devices which the subject owns already fit into their lives, conceivably BYOD might enhance compliance and decrease dropout. Moreover, BYOD has potential to decrease the set-up cost of eCOA for the sponsor company. The perceived cost of eCOA is one of the key reasons why most of PROs are still gathered in paper format, despite proven benefits with electronic administration, so adopting BYOD might be also beneficial for data quality.

Practical Considerations of BYOD in Clinical Trials

Above, we discussed the regulatory challenges of using BYOD in clinical trials. Let us imagine in this section that those aspects are all taken care of and take one step further to consider the practical implications of BYOD in a clinical trials. There is an abundance of practical issues, ranging from fundamental to trivial, which need to be addressed for each trial, which is planning to use BYOD. In this section we will go over some of the major questions and evaluate the possible choices from the point of view of a software company with keen interest in usability.

The primary challenge is to choose the technical solution that would suit all: legislators, sponsor companies and end users alike. The two most often considered solutions are native applications and web-based implementation. Both can be developed for and used with desktop and laptop computers as well as mobile devices (tablets and mobile phones).

Applications have the benefit of being available also off-line, which can be important, especially in mobile devices. The user does not need a regular Internet connection in order to be able to use an application. The data does need to be transferred occasionally to the server, but this can also be done in batches (for example the subject fills in a daily questionnaire and only transfers data once a week).

The downside of native apps is the wide range of operating systems (OS) for which the application would have to be tailored. For example, laptop and table computers (e.g. Windows, OS X, any flavor of Unix-like OS) as well as mobile devices like smartphones and tablet computers (e.g. Android, iOS, Windows, various less common OSs like Sailfish) utilize a multitude of OSs with their own ecosystems and platform-specific guidelines for applications. Covering all of these is a monstrous task.

Providing customer support for native applications can be as formidable a project as developing for them. User-made hacks and system updates can affect the functioning of a native application. This would make answering user questions much more demanding and arduous and most likely would necessitate several different back-line customer support specialists, one for each separate operating system. This would add significantly to project cost.

If embracing the native application route, it seems reasonable to limit the number of OSs which will be supported. In the third quarter of 2013, the two predominant mobile operating systems, iOS and Android, have more than 90% of the combined market share (IDC, 2013).

Cost is a big motivator for pharmaceutical industry, and when looking at instru-

ment implementation, a web-based instrument requires less developer work hours and is thus likely to end up being less expensive. Web-based instruments can be opened and viewed with any device with a web browser, and they only need to be created once. As a special advantage, despite numerous more or less common web browsers they all show web pages essentially the same without any special tailoring. There are some compatibility issues in specific bits of css-code and older versions of desktop and mobile browsers. These are well documented and can be circumvented by careful design and by knowing the most common (or almost obscure) browsers in the study population.

Using a web-based system requires the user to be connected to the Internet while filling in the instrument. This may be an issue in some parts of the world, whereas most industrialized countries have readily available Internet connections with reasonably good coverage both for computers and mobile devices. It might be feasible for the sponsor company to cover the costs of subjects' Internet connection for the duration of the study. This verges on ethically dubious financial compensation of subjects in clinical trials, a potential source of bias, and thus needs to be checked with relevant regulatory authorities.

Both app-based and web-based models enable secure information transfer between the subject's device and data servers. The actual technical implementation of this is beyond the scope of this thesis, but the author would like to point out that this too can be easier to accomplish with a single web-based implementation than with operating system specific applications.

Subject authentication needs to be considered as well. In order to have verified source for the data, the subject should enter user id and password at the beginning of every data entry session. This requirement is the same for single-device studies, but must also be implemented for BYOD, even if the subject is the sole user of that device.

The implementations should always be designed keeping the target population in mind. This will help decide for example how likely it is that subjects would have their own suitable devices, and which device types are predominant. This will then help to decide whether the implementation should be designed for mobile or desktop. In affluent countries, it is normal for a typical household to have several Internet capable devices, in which case the type of the PRO measure can also impact the choice of primary target device. Also, it is possible that there are populations where using BYOD is not applicable because of insufficient penetration of technology.

Owning an Internet-capable device cannot and should not be an inclusion criterion. Excluding people without suitable device might severely damage the ecological validity of the study by biasing the sampling towards those more affluent and often more healthy as well. Rather, there should be enough loan devices on each site to ensure that subjects can choose whether they want to use their own device or loan a hand-held device. The necessary number of devices for each site depends on the study population, especially how likely they are to own a computer or a mobile device. The actual proportion of people who do not own or don't want to use their device is impossible to predict before some practical experience has been gained.

To summarize, there can be different suitable ways to gather ePRO with BYOD.

While choosing technical solutions, the following should be kept in mind:

- the implementation should be designed for the target population,
- information security and user authentication both need to be carefully considered, and finally,
- subjects should not be excluded from a study for not owning or not wanting to use their own device.

Recommendations and Future Work

In this section, the author outlines her recommendations for the technical implementation as well as suggestions for future work on validation of BYOD instruments.

From a practical point of view, the author considers a web-based implementation most suitable for clinical trials. A web-based instrument is quicker (i.e. less expensive) to develop than a number of OS-specific applications. However, a small native app whose function is to notify the user to fill in the ePRO and direct them to the correct web page might constitute a nice addition for mobile device users. Though this would be by no means necessary, it might potentially diminish loss of data and increase data quality.

Also, it might be useful to limit the suitable devices to small and medium (mobile phone and tablet) or medium and large (tablet, laptop and desktop computer). This decision should be done keeping in mind the projected target population and the type of the ePRO being implemented. Daily diaries would better work with mobile devices whereas long, multi-item instruments, like the one described in this thesis, would be better suited for larger screens. Limiting the range of screen sizes in this way would significantly reduce the potential sources of variation between different devices, which, in turn, should increase the quality of the data.

In order to enhance the user experience, the layout should somehow be adjusted to the screen. This can be achieved in different ways, but the author recommends either a combination of fixed layout and fluid layout, or a fluid layout with upper and lower limits.

Most attention should be paid to the response options. Response options should be uniform in size whenever possible. It would be beneficial if the rest of the layout would conform to the size of the screen. This will give a more tailored look and feel, which modern Internet users are accustomed to.

In instrument development, the capabilities of different browsers should be considered carefully. In order to keep the instrument cohesive for all users, features, such as rounded corners, should not be used if they cannot be displayed in the same way in all most common browsers.

Based on the body of literature presented in Chapter 2, Background, and keeping in mind the regulatory and practical issues outlined in the previous sections, BYOD does seem like a plausible option for patient diaries. The results from cognitive interviews do not rule out the possibility that using BYOD would prove plausible also for formal and validated instruments. However, to further prove the point

to authorities and ensure the applicability of the qualitative results on a larger sample, quantitative evaluation is also needed. The author believes that a handful of quantitative large-sample studies with widely used instruments should be enough to give significant evidence of the (lack of) impact of device size, type and software on the psychometric properties of all instruments.

If these quantitative studies give good indication of measurement equivalence between different device types, it should be sufficient to treat moving from another electronic implementation to BYOD implementation as a minor modification in the ISPOR Task Force scale (Coons et al., 2009) of modifications. This would mean that in the future, cognitive interviewing with a small subset would be sufficient proof of measurement equivalence.

6 Summary

Discovering, manufacturing and selling drugs constitutes a huge industry with global spending on prescription medications of over US$ 900 BN (IMS Health, 2013). Substantial costs are associated with potential revenues: developing a new drug prior to market permission costs approximately US$ 400 million (DiMasi et al., 2003). Before introducing a new drug to the market, pharmaceutical companies need to prove that the new compound is both safe to use and that it targets the symptoms it is claiming to target. In order to prove that, new drugs go through clinical trials, where safety and efficacy data is gathered while the treatment is tested on people. The data is later rigorously analyzed both by the sponsor company (i.e. the pharmaceutical company developing the treatment) and regulatory authorities, most notably European Medicines Agency (EMA) and US Department of Health and Human Services, Food and Drug Administration (FDA).

Clinical trials have designated endpoints, measures of success for that particular trial. The endpoints can be objectively measurable, such as blood pressure or survival rate. In certain disciplines of medicine, measured variables of interest cannot be as easily evaluable by the investigator (e.g. quality of life of a cancer patient, number and severity of migraine attacks). If a clinical trial only focuses on clinician-reported outcomes, vitally important aspects to patients can be too easily overlooked. Scales and measures in which the patient is asked about phenomena that are not easily observable are called patient reported outcomes (PRO).

Patient reported outcomes may be gathered by using paper-based or electronic questionnaires. PRO capture in clinical trials is ever so slowly changing its course from paper-and-pen to electronic, but the progress is slow due to perceived cost of ePRO and attitudes of sponsor companies and regulatory authorities. Traditionally, the sponsor company has provided the subjects with a device, commonly a mobile phone, to record ePROs. The goal of this thesis has been to evaluate the plausibility of using bring your own device (BYOD) model in capturing patient reported outcomes in clinical trials. In practice, BYOD means that the subjects are allowed to use their own devices for filling in the designated PRO measures.

If the subjects are allowed to use their own devices for filling in for example subject diaries during the trial, it might enable wider adoption of electronic data capture in large late-phase trials. Using subjects' own devices lowers the overall cost of the trial for the sponsor company and hopefully increases compliance, since the reporting devices already fit into the lives of the subjects.

This thesis has had three interconnected research questions:

1. Would it be possible, in theory, to use BYOD in clinical trials to capture patient reported outcomes?

2. How should instruments be designed to ensure measurement equivalence across a range of devices?

3. What kinds of validation are necessary to prove measurement equivalence of a multi-platform implementation of a PRO instrument?

Over the course of the project the first question was further focused to relate to the regulatory aspects and the quality of the data gathered by BYOD. The answers to these questions are presented below:

1. There are two key issues related to using BYOD for capturing patient reported outcomes: Firstly, it needs to be considered if the BYOD model can be used in clinical trials, i.e. if the regulatory requirements relating to electronic data capture can be met with BYOD As was discussed in chapter 5, Discussion, section Plausibility of BYOD in Clinical Trials, it is entirely possible to capture source data with BYOD in a way which is in keeping with the relevant regulations. Secondly, it needs to be established that scores and measurements of formal instruments are reliable, valid and comparable across platforms. This needs further validation work, which is described later in this section.

2. In this thesis, the author offers a number of tentative best practices for designing for several devices simultaneously. The author considers the web-based implementation easiest to harmonize over different device types. The design should always take into account the specific features of the target population. It is important that response options should be kept uniform in size regardless of device size or proportions whenever possible. It is a good idea for the layout to adapt to screens of different size and proportions, as long as there are upper and lower limits in place. Also, features which cannot be used in all devices or on all browsers should be avoided.

The project described in this thesis consists of developing a cross platform implementation of the Amsterdam IADL Questionnaire® and validating it with cognitive interviewing. The implementation was done using CRF Health's proprietary TrialStudio™ development tool and it was built on the TrialMax Web™ eCOA solution. The design of the implementation has been done mostly in accordance with the best practices described above.

The implementation was tested with 6 Finnish subjects so that each filled in a subset of the questions both on a tablet computer and on a desktop computer by using a randomized crossover design. The subjects all professed fluency in English. Cognitive interviewing was done entirely in English.

3. The results of the cognitive interviews did not provide sufficient evidence to judge the impact that the device might have had on subjects' interpretation of items in the questionnaire. Since cross platform studies, and especially studies using BYOD, are not very common or widely published, further validation is required. The author recommends carrying out statistical validation studies on well-known instruments that are widely used. With more evidence from bigger studies, more concrete suggestions can be made about the effect of BYOD on psychometric properties of PRO instruments. If the results of quantitative studies do not give evidence to the contrary, it would seem suitable to consider a well-executed BYOD implementation as a minor level of modification and thereby comparable for example to paper-to-electronic transformation. This

would mean that in the future, qualitative measures would be sufficient for proving measurement equivalence.

The current technology enables capturing source data with BYOD in a way that is in keeping with the relevant regulations. The quality, especially cross-device comparability, of the captured data is still debatable. After publication of the results from the statistical studies, it will be more clear how good measurement equivalence BYOD can provide. A sensible place to begin might be late-phase studies where the target population is fairly young. Given the price of drug development, it can take time before any sponsor company is willing to take a risk with their clinical trials.

References

21 C.F.R. pt. 312.62 Investigational New Drug Application, 2013.

Paul C. Beatty and Gordon B. Willis. Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2):287 – 311, 2007. ISSN 0033362X.

PC Beatty. Cognitive interviewing: the use of cognitive interviews to evaluate ePRO instruments. In Bill Byrom and Brian Tiplady, editors, *EPRO: Electronic Solutions for Patient-Reported Data*, pages 23–48. Gower Publishing, Ltd., 2010.

Gian Franco Buccheri, Domenico Ferrigno, Antonio Curcio, Ferruccio Vola, and Alberto Rosso. Continuation of chemotherapy versus supportive care alone in patients with inoperable non-small cell lung cancer and stable disease after two or three cycles of macc. results of a randomized prospective study. *Cancer*, 63(3): 428–432, 1989.

Bill Byrom and Brian Tiplady, editors. *EPRO: Electronic Solutions for Patient-Reported Data*. Gower Publishing, Ltd., 2010.

Tian-hui Chen, Lu Li, Joerg M Sigle, Ya-ping Du, Hong-mei Wang, and Jun Lei. Crossover randomized controlled trial of the electronic version of the chinese sf-36. *Journal of Zhejiang University Science B*, 8(8):604–608, 2007.

Debbie Collins. Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research*, 12(3):229–238, 2003.

S.J. Coons, C.J. Gwaltney, R.D. Hays, J.J. Lundy, J.A. Sloan, D.A. Revicki, W.R. Lenderking, D. Cella, and E. Basch. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value in Health*, 12(4):419–429, 2009.

George Demiris, Stanley M Finkelstein, and Stuart M Speedie. Considerations for the design of a web-based clinical monitoring and educational system for elderly patients. *Journal of the American Medical Informatics Association*, 8(5):468–472, 2001.

Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185, 2003.

Lynda C Doward and Stephen P McKenna. Defining patient-reported outcomes. *Value in Health*, 7:S4–S8, 2004.

Lynda C Doward, Ari Gnanasakthy, and Mary G Baker. Patient reported outcomes: looking beyond the label claim. *Health and Quality of Life Outcomes*, 8(1):89, 2010.

Sonyo Eremenco and Dennis A Revicki. Regulation and Compliance: Scientific and Technical Regulatory Issues Associated with Electronic Capture of Patient-Reported Outcome Data. In Bill Byrom and Brian Tiplady, editors, *EPRO: Electronic Solutions for Patient-Reported Data*, pages 79–106. Gower Publishing, Ltd., 2010.

European Medicines Agency. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products, July 2005.
`http://www.ema.europa.eu/pdfs/human/ewp/13939104en.pdf`.

Eurostat. More than 60% of individuals in the EU28 use the internet daily. Press Release, December 2013.
`http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122013-BP/EN/4-18122013-BP-EN.PDF`.

Peter Fayers and David Machin. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes*. Wiley, 2007.

PM Fayers, NM Bleehen, DJ Girling, and RJ Stephens. Assessment of quality of life in small-cell lung cancer using a daily diary card developed by the medical research council lung cancer working party. *British Journal of Cancer*, 64(2):299, 1991.

Forbes. The Cost Of Creating A New Drug Now $5 Billion, Pushing Big Pharma To Change, 2013.
`http://www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine/`.

Ari Gnanasakthy, Margaret Mordin, Marci Clark, Carla DeMuro, Sheri Fehnel, and Catherine Copley-Merriman. A review of patient-reported outcome labels in the united states: 2006 to 2010. *Value in Health*, 15(3):437–442, 2012.

Ari Gnanasakthy, Sandra Lewis, Marci Clark, Margaret Mordin, Carla DeMuro, V Strand, D Fiorentino, CC Hu, RM Day, RM Stevens, et al. Potential of patient-reported outcomes as nonprimary endpoints in clinical trials. *Health and Quality of Life Outcomes*, 11(1):83, 2013.

Seth B Goldsmith. The status of health status indicators. *Health Services Reports*, 87(3):212, 1972.

Gordon H Guyatt, David H Feeny, and Donald L Patrick. Measuring Health-Related Quality of Life. *Annals of Internal Medicine*, 118(8):622–629, 1993.

Chad J Gwaltney, Alan L Shields, and Saul Shiffman. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. *Value in Health*, 11(2):322–333, 2008.

IDC. Android Pushes Past 80% Market Share While Windows Phone Shipments Leap 156.0% Year Over Year in the Third Quarter, According to IDC. Press Release, November 2013.
http://www.idc.com/getdoc.jsp?containerId=prUS24442013/.

IMS Health. Global Pharmaceutical Sales, 2003-2012, 2013.
http://www.imshealth.com/deployedfiles/imshealth/Global/Content/Corporate/Press%20Room/Total_World_Pharma_Market_Topline_metrics_2012.pdf.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Guideline for Good Clinical Practice, 1996.
http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf.

B Jones, P Jarvis, JA Lewis, and AF Ebbutt. Trials to assess equivalence: the importance of rigorous methods. *BMJ: British Medical Journal*, 313(7048):36, 1996.

I. McDowell. *Measuring health : a guide to rating scales and questionnaires*. Oxford University Press, USA, 2006.

Damian J McEntegart. Equivalence Testing: Validation and Supporting Evidence When Using Modified PRO Instruments. In Bill Byrom and Brian Tiplady, editors, *EPRO: Electronic Solutions for Patient-Reported Data*, pages 185–212. Gower Publishing, Ltd., 2010.

Kenneth O McGraw and SP Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

Lee J Meehl and Paul E Cronbach. Construct validity in psychological tests. *Methods and Techniques in Business Research*, 52:156, 1955.

Jennifer E Moyer. Managing Mobile Devices in Hospitals: A Literature Review of BYOD Policies and Usage. *Journal of Hospital Librarianship*, 13(3):197–208, 2013.

Pirjo Räsänen, Eija Roine, Harri Sintonen, Virpi Semberg-Konttinen, Olli-Pekka Ryynänen, and Risto Roine. Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. *International Journal of Technology Assessment in Health Care*, 22(02):235–241, 2006.

Dennis Revicki, Ron D Hays, David Cella, and Jeff Sloan. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, 61(2):102–109, 2008.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

S.A.M. Sikkes, ESM De Lange-de Klerk, YAL Pijnenburg, and P. Scheltens. A systematic review of Instrumental Activities of Daily Living scales in dementia: room for improvement. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(1): 7–12, 2009.

S.A.M. Sikkes, E.S.M. de Lange-de Klerk, Y.A.L. Pijnenburg, D.L. Knol, P. Scheltens, and B.M.J. Uitdehaag. The development of a new IADL informant-based questionnaire: The Amsterdam IADL questionnaire. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 6(4):S115–S116, 2010.

Tom W Smith. Little things matter: A sampler of how differences in questionnaire format can affect survey responses. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 1046–1051, 1995.

Arthur A Stone, Saul Shiffman, Joseph E Schwartz, Joan E Broderick, and Michael R Hufford. Patient compliance with paper and electronic diaries. *Controlled Clinical Trials*, 24(2):182 – 199, 2003.

David L Streiner and Geoffrey R Norman. *Health measurement scales: a practical guide to their development and use*. Oxford University Press, USA, 2008.

Suomen virallinen tilasto. Väestön tieto- ja viestintätekniikan käyttö , 2013. `http://www.stat.fi/til/sutivi/2013/sutivi_2013_2013-11-07_tau_003_fi.html`.

The NPD Group. Internet connected devices surpass half a billion in u.s. homes, according to the npd group. Press Release, March 2013. `https://www.npd.com/wps/portal/npd/us/news/press-releases/internet-connected-devices-surpass-half-a-billion-in-u-s-homes-according-to-the-npd-group/`.

Gordon Thomson. BYOD: enabling the chaos. *Network Security*, 2012(2):5 – 8, 2012.

United States Census Bureau. Computer and Internet Use in the United States: 2011. U.S. Department of Commerce, May 2013.

US Department of Health and Human Services Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims, 2009. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf`.

US Department of Health and Human Services Food and Drug Administration. Guidance for industry: Electronic source data in clinical investigations, 2013a. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM328691.pdf`.

US Department of Health and Human Services Food and Drug Administration. Guidance for industry and food and drug administration staff: Mobile medical applications, 2013b.
`http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM263366.pdf`.

SD Walter, M Eliasziw, and A Donner. Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1):101–110, 1998.

S Whybrow, JR Stephen, and RJ Stubbs. The evaluation of an electronic visual analogue scale system for appetite and mood. *European Journal of Clinical Nutrition*, 60(4):558–560, 2005.

Gordon B Willis. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage, 2005.

Richard J Willke, Laurie B Burke, and Pennifer Erickson. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Controlled Clinical Trials*, 25(6):535–552, 2004.

World Health Organization and others. Preamble to the constitution of the WHO as adopted by the International Health Conference, New York, 19–22 June 1946. *World Health Organization Geneva*, 1948.