Norwegian University of Life Sciences
Faculty of Veterinary Medicine and Biosciences
Department of Chemistry, Biotechnology and
Food Science (IKBM)

Master Thesis 2015
60 credits

A statistical analysis of the treatment effects of Traditional Chinese Medicine (TCM) in various health problems

Veronika Lindberg

Master thesis in Applied Statistics (60 credits)

A statistical analysis of the treatment effects of Traditional Chinese Medicine (TCM) in various health problems

Author
Veronika Lindberg

Main Supervisor (NMBU)
Thore Egeland (1)

Main Supervisor
Johannes Baak (2)

Co-Supervisor (NMBU)
Trygve Almøy (1)

(1) NMBU **Norwegian University of Life Science**

(2) DR. MED. JAN BAAK AS
Behandling med Vestlig og Tradisjonell Kinesisk Medisin

PREFACE

This master thesis in applied statistics has been conducted at the Norwegian University of Life Science in the Oslo area in cooperation with the private clinic Dr. Med. Jan Baak AS in Tananger, Rogaland, Norway. The project was initiated in 2013 by Jan Baak. The data used in the study was collected during 2014 and 2015 as part of the study.

I want to thank Jan Baak for allowing me to take part in this interesting and sometimes challenging project. I learned a lot since I was allowed to participate in all the phases of the project, from the first outline of the project, the application for necessary approvals, design of questionnaires, performing of the survey, and finally the major part of the project, which was the statistical analyses of the data. I want to thank all collaborators for all encouragement and guidance within a subject that was in some cases new and unfamiliar to me, which made it possible for me to write this interdisciplinary master thesis.

I want to thank Thore Egeland for help and good advice during all the phases of the project, and especially during the last part of the project, when I experienced challenges with making statistical models for a small and unbalanced dataset. I also want to thank Trygve Almøy for inspiration and help with multivariate data analyses, and the entire Biostatistics group at IKBM for a positive and stimulating working environment.

Finally I want to thank my husband Aksel and my family for support and practical help at home during the project period.

ABSTRACT (ENGLISH)

We have in a pilot study retrospectively and prospectively investigated the treatment effects of Traditional Chinese Medicine in a general medical care practice. The health condition of the retrospective group (n=41 patients) was investigated one year after the last treatment. The health condition for the prospective group (n=7 patients) was followed from before start of treatment and until 9 months after the start of the treatment. The response rate was 16% in the retrospective and 23% in the prospective group (which is low, but not uncommon in Norwegian questionnaire surveys).

The patients were treated for a broad range of health problems, from everyday ailments to long lasting and very serious degenerating and malignant diseases. The mean duration of the health problems before start of treatment was 6 years for the retrospective group and 12 years for the prospective group.

The burden of the health problems was measured on a 10 point Visual Analogue Scale, and Health Related Quality of Life was measured by the RAND SF-36 questionnaire (Version 1.0). The measurements by the VAS Scale and SF-36 differ. For VAS, only the variables of interest are measured, while with SF-36, also the features that are not affected by the therapy are measured. Therefore, the VAS is more sensitive to changes in individuals, while the SF-36 is more often used for monitoring of changes in groups over time, and also for comparing outcomes from different studies.

The effect of the treatment was measured as the change in VAS and HRQoL scores between baseline and first follow up 3 months later for the prospective group. The effect was measured as change in scores and effect size. According to a conventional definition, an effect size above 0.8 was regarded as a large effect size, between 0.79 and 0.5 as a

moderate effect size, between 0.49 and 0.2 as small effect size, and below 0.19 as no difference.

We found an improvement of the health situation, both with VAS and with SF-36. The improvement measured on a 10 point VAS scale was a change by 1.9 points from baseline to the first follow up. The effect size was 1.0 (large). The improvement of the SF-36 summary scores were 4.6 for the Mental Component Summary (MCS) and 3.7 for the Physical Component Summary (PCS) in the same time period. The effect size was 0.7 (moderate) for MCS, and 0.3 (small) for PCS.

The number of treatments that would be necessary to improve health measured by SF-36 under similar circumstances, was predicted to be 2 treatments to reach a small improvement (effect size 0.2), 4 treatments to reach a moderate improvement (effect size 0.5) and 7 treatments to reach a large improvement (effect size 0.8).

A change in 3 to 5 SF-36 scores, which is equal to a small effect size (0.2), is regarded as the Minimum Clinically Meaningful Difference for SF-36. We found in the prospective group an average improvement of 4.1 for the SF-36 features, which is a good improvement. This again indicates that the patients have experienced a meaningful improvement of health.

As this was a pilot study with a small study group, the study should be repeated with a larger study group, and preferably with a control group, to confirm or reject the findings.

ABSTRACT (NORWEGIAN)

Vi har i en pilotstudie retrospektivt og prospektivt undersøkt behandlingseffekten av Tradisjonell Kinesisk Medisin i en medisinsk allmennpraksis. Helsetilstanden til den retrospektive gruppen (n = 41 pasienter) ble undersøkt ett år etter siste behandling. Helsetilstanden for den prospektive gruppen (n = 7 pasienter) ble fulgt fra før behandlingsstart og inntil 9 måneder etter starten av behandlingen. Svarprosenten var 16% i retrospektiv og 23% i den prospektive gruppen (som er lavt, men ikke uvanlig i norske spørreundersøkelser).

Pasientene ble behandlet for et bredt spekter av helseproblemer, fra hverdagslige plager til langvarige og svært alvorlige degenererende og ondartede sykdommer. Den gjennomsnittlige varigheten av helseproblemene før behandlingsstart var 6 år for den retrospektive gruppen og 12 år for den prospektive gruppen.

Byrden av helseproblemene ble målt på en 10 punkts Visual Analogue Scale, og Helserelatert livskvalitet ble målt ved RAND SF-36 spørreskjema (versjon 1.0). Målingene ved VAS og SF-36 er forskjellige. For VAS er det bare de variablene som er av interesse som blir målt, mens med SF-36 blir også de funksjonene som ikke er berørt av behandlingen målt. Derfor er VAS mer følsom for endringer i individer, mens SF-36 blir oftere anvendt for overvåkning av endringer i grupper over tid, og også for å sammenligne resultatene fra forskjellige undersøkelser.

Effekten av behandlingen ble målt som endring i VAS og HRQoL score mellom baseline og første oppfølging tre måneder senere for den prospektive gruppen. Effekten ble målt som endring i score og effektstørrelse. I henhold til en konvensjonell definisjon, ble en effekt størrelse over 0.8 ansett som en stor effektstørrelse, mellom 0.79 og 0.5 som en

moderat effekt størrelse, mellom 0.49 og 0.2 som en liten effekt størrelse, og under 0.19 som ingen forskjell.

Vi fant en bedring av helsesituasjonen, både med VAS og med SF-36. Forbedringen målt på en 10 poeng VAS skala var en endring på 1.9 poeng fra baseline til første oppfølging. Effektstørrelsen var 1.0 (stor). Forbedring av SF-36 summerte score var 4.6 for Mental Component Summary (MCS) og 3.7 for Physical Component Summary (PCS) i samme periode. Effektstørrelsen var 0.7 (moderat) for MCS, og 0.3 (liten) for PCS.

Antall behandlinger som ville være nødvendig for å forbedre helse målt ved SF-36 under lignende omstendigheter, ble predikert å være to behandlinger for å nå en liten forbedring (effekt størrelse 0.2), 4 behandlinger for å oppnå en moderat forbedring (effektstørrelse 0.5) og 7 behandlinger for å nå en stor forbedring (effektstørrelse 0.8).

En endring i 3 til 5 score, noe som tilsvarer en liten effektstørrelse (0.2), regnes som Minimum klinisk relevant forskjell for SF-36. Vi fant i den prospektive gruppen en gjennomsnittlig forbedring på 4.1 for SF-36 funksjoner, som er en god forbedring. Dette indikerer igjen at pasientene har opplevd en meningsfull forbedring av helse.

Ettersom dette var en pilotstudie med et lite antall deltakere i studien, bør studien gjentas med en større studie gruppe, og helst med en kontrollgruppe, for å bekrefte eller forkaste funnene.

ABBREVIATIONS

AM	Alternative Medicine
CAM	Complementary and Alternative Medicine
ChQoL	Chinese Health Related Quality of Life survey instrument
HIE	The Health Insurance Experiment
HRQoL	Health Related Quality of Life
MOS	Medical Outcomes Study
NAFKAM	The National Research Center in Complementary and Alternative Medicine
NSD	Norsk Samfunnsvitenskapelig Datatjeneste AS, ethical approval instance for the study
QoL	Quality of Life
PC1	Principal Component 1
PCA	Principal Component Analysis
PCR	Principal Component Regression
SD	Standard deviation
SF-36	Rand 36-Item Short Form Health survey instrument
SF-36 BP	Bodily Pain
SF-36 GH	General health
SF-36 MCS	Mental Component Summary
SF-36 MH	Emotional well-being/Mental Health
SF-36 PCS	Physical Component Summary
SF-36 PF	Physical functioning
SF-36 RE	Role functioning/emotional
SF-36 RP	Role functioning/physical
SF-36 SF	Social functioning
SF-36 VT	Energy/fatigue/Vitality
TCM	Traditional Chinese Medicine
Var	Variance
VAS	Visual Analogue Scale
WHO	World Health Organization
WHOQoL	WHO Quality of Life survey instrument
WHOQoL-BREF	Short form of WHOQoL
WM	Western Medicine

Table of Contents

Chapter 1: Introduction 4
1.1 Definition of Traditional Chinese Medicine 4
1.2 Origin of TCM 4
1.3 TCM therapies 4
1.4 TCM in China 4
1.5 TCM outside China 5
1.6 TCM in Norway 6
1.7 Philosophy behind Traditional Medicine 6
1.8 Research on Traditional Medicine 7
1.9 Project summary 8
1.10 Organization of the thesis 9
Chapter 2: Aim and objectives of the study 9
2.1 Aim 9
2.2 Objectives 10
Chapter 3: Literature review 11
3.1 WHO recommendations on Research on Traditional Medicine 11
3.2 Research on Traditional Medicine in Norway 12
3.3 Research on Efficacy and Safety of Traditional Chinese Medicine 13
3.4 Assessment of Health Related Quality of Life instruments 15
3.5 History of the SF-36 survey instrument 17
Chapter 4: Methods 18
4.1 Study design 18
4.1.1 Pilot study 18
4.1.2 Study setting 19
4.1.3 Project schedule 19
4.1.4 Selection of participants 19
4.1.5 Treatment 20
4.1.6 Data collection 20
4.1.7 Handling of collected data 21
4.1.8 Missing data within a survey form 21
4.1.9 Missing surveys forms 21
4.1.10 Reliability and Validity 21
4.2 VAS psychometric response scale 22
4.3 Likert psychometric response scale 22
4.4 Short Form - 36 patient health questionnaire 23
4.4.1 SF-36 Health categories 23
4.4.2 Normative data from the general Norwegian population 24
4.4.3 Transforming of SF-36 raw scores to Norm based scores 25
4.4.4 SF-36 raw scores, T-scores and z-scores 26
4.4.5 Construction of SF-36 summary scores 28
4.5 Effect size 29
4.5.1 Effect size for the retrospective group 29
4.5.2 Effect size for the prospective group 30
4.6 Study outcome 31
4.7 Statistical analysis 31
4.7.1 HRQoL as the response variable 31

4.7.2 Linear regression models 32
4.7.3 Model selection criteria 33
4.7.4 Principal Component Analysis 33
4.7.5 Mixed Model for handling of combined fixed and random effect terms 34
4.8 Software used in the project 36
Chapter 5: Results for the Retrospective group 36
5.1 Participant rate 36
5.2 Missing data 37
5.3 Health problems 37
5.4 Treatments 38
5.5 Response variables 38
5.5.1 Correlation between response variables 38
5.6 Explanatory variables 39
5.7 Univariate statistical analyses 40
5.7.1 Self-rating of health, SF-36 question 1 40
5.7.2 Self-rating of health transition, SF-36 question 2 41
5.7.3 Visual Analogue Scale mean scores 42
5.7.4 Mean scores for the SF-36 Health categories 42
5.7.5 Calculation of Effect size for SF-36 mean scores 43
5.7.6 Testing hypothesis about SF-36 mean scores 45
5.7.7 Power calculation 46
5.8 Bivariate statistical analyses 48
5.8.1 SF-36 mean scores compared to VAS mean scores 48
5.9 Multivariate statistical analysis 51
5.9.1 Principal Component Analysis on the response variables 51
5.9.2 Principal Component Regression 54
5.9.3 Dependencies within the dataset handled by the Mixed Model 55
5.9.4 Prediction of SF-36 Health category scores 56
5.9.5 Comparing the PCR model and the Mixed Model 58
Chapter 6: Results for the Prospective group 59
6.1 Participant rate 59
6.2 Missing data 59
6.3 Number of submitted survey forms over time 59
6.4 Health problems 60
6.5 Treatment 60
6.6 Response variables 60
6.7 Explanatory variables 60
6.8 Univariate statistical analyses 61
6.8.1 Self-rating of health, SF-36 question 1 61
6.8.2 Self-rating of health transition, SF-36 question 2 62
6.9 Bivariate statistical analyses 63
6.9.1 Change in VAS scores by Time and Person 63
6.9.2 Change in SF-36 scores by Time and Person 64
6.9.3 Change in group scores by Time 66
6.9.4 Testing hypothesis about change in Health problems by Time 69
6.9.5 Testing hypothesis about Change in Health problems by Number of treatments 71
6.9.6 Prediction of number of treatments necessary to reduce Health problems measured on a VAS scale 72
6.9.7 Testing hypothesis about Change in HRQoL by Number of treatments 73

6.9.8 Prediction of number of treatments necessary to improve HRQoL ... 75
6.10 Mixed model analysis ... 75
6.10.1 Model selection ... 75
6.10.2 Observed SF-36 summary scores ... 76
6.10.3 SPSS model ... 76
6.10.4 Prediction with random effects set to 0 ... 77
6.10.5 Prediction with random effects set to average value ... 78
Chapter 7: Discussion ... 80
7.1 Regression to the mean ... 81
7.2 Biased sample ... 81
7.3 Comparison against the general Norwegian population ... 82
7.4 Clinically Meaningful Differences in HRQoL ... 82
7.5 Reliability and Validity ... 82
7.6 Missing answers and missing survey forms ... 83
7.7 HRQoL as the response variable ... 84
7.8 Health problems ... 84
7.9 Limitations and strengths of the chosen study design ... 85
Chapter 8: Conclusion ... 85

List of tables ... 86

List of figures ... 87

References ... 89

Chapter 1: INTRODUCTION

1.1 Definition of Traditional Chinese Medicine

The term “Traditional Medicine” is often used interchangeably with the terms “Alternative Medicine” or “Complementary Medicine” [1]. In this document, the term “Traditional Chinese Medicine” (TCM) is used in the understanding of medicine as evolved in China over more than 2000 years.

1.2 Origin of TCM

The discipline of Traditional Chinese Medicine is old. The first known detailed descriptions and classification of diagnosis and treatments is more than 2000 years old (the work is known as “Questions and answers by the Yellow emperor”) [2]. Since then, TCM has developed over the next millenniums in China, with regular expansions of theories and therapeutic principles.

1.3 TCM therapies

TCM consists of 5 different therapeutic principles: herbal medicine, acupuncture, Tuina massage, diet science and Qi Gong exercise and therapy.

1.4 TCM in China

In the 20th century, China became increasingly influenced by Western Medicine (WM). The Traditional Medicine lost terrain, until it was reinstated by the Chinese Communist party in the middle of the century. In 1956, TCM was declared a culture inheritance by the Chinese Communist party at it’s annual congress and a massive revival took place [3]. Traditional practitioners got status in the official health care system, and a process to integrate Traditional Medicine with scientific-based Medicine started. The official medical system in China now offers treatments both with Western Medicine and with Traditional Medicine.F According to a national survey in China, *the number of TCM visits was 907 million in 2009, which accounts for 18% of all medical visits to surveyed institutions; the number of TCM inpatients was 13.6 million, or 16% of the total in all hospitals surveyed* [4].

1.5 TCM outside China

Acupuncture began to spread to Europe in the second half of the 17th century [2, 5]. From the early 1900's there was a slightly increasing interest for TCM in Western countries, and the interest really took off after increased contact between China and Western countries in the 1970's. The story about a New York Times reporter who was treated by acupuncture after an operation in Beijing in 1971 was apparently the first story about acupuncture treatment who reached the mass of English speaking citizens in the North America [6].

After President Nixon's visit to China in 1972, he ordered the director of the National Institute of Health to thoroughly study acupuncture. Many studies followed, and in 1997, acupuncture was officially approved as a medical action by the US Food and Drug Administration Modernization Act (FDAMA) [7].

With the massive emigration of Chinese citizens after the 1989 Tiananmen Square revolution in Beijing and the transition of Hong Kong from the UK to China in 1997, an enormous increase of TCM occurred in North America, Malaysia, Singapore, Australia and New Zealand. Many prospective randomized clinical trials studies since then have shown the efficacy of acupuncture and to a lesser degree also herbal medicine [8].

The interest in and use of Traditional Chinese Medicine, TCM, is increasing in the western world, including Europe [9, 10]. Leading American hospitals such as MD Anderson Cancer Center, Houston, Texas, the Johns Hopkins Hospital, Baltimore, Maryland, the Sloan Kettering Cancer Hospital, New York and many others have large departments for "Integrative Medicine/Oncology". Many medical faculties in North America offer some form of TCM education. In Europe, the Technical University of Munich, which is one of Germany's leading universities, started in 2013 the first European Master program in TCM for experienced western medical doctors. The master program has for three years attracted 16-28 students [11].

1.6 TCM in Norway

Treatment within the official health care system in Norway is reimbursed, while visits to Alternative Therapies must be paid by the patients themselves. Despite that the treatments are more expensive for the patients, 30-40% of the adult population in Norway used some form of Alternative Therapies in 2014 [12]. And although TCM is not official included in the public Health Care system in Norway, acupuncture was offered in 27% of Norwegian hospitals in 2001 [13].

1.7 Philosophy behind Traditional Medicine

From ancient times, health and diseases have been explained in a supernatural way, by good and evil gods, spirits, energies or powers. This is in contrast to Traditional Medicine and Modern Western Medicine where health and diseases are explained in a natural way. Hippocrates (400 BC) is regarded as the founder of Traditional Medicine in Europe. He stated that "Illness has a natural cause" and "Diagnosis and treatment should be based upon experience and reason" [14].

The development from ancient to Traditional Medicine and further to modern Western Medicine is illustrated in Figure 1.

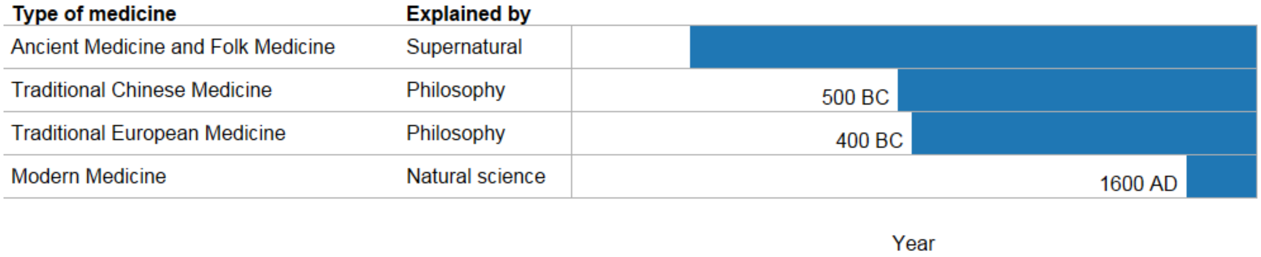


Figure 1. Approximately historical timeline of Medicine.

The main difference between Traditional Medicine in Europe and China is the different philosophies used to explain health and diseases. In Europe, philosophy has mainly been founded on the understanding that the fundamental component of the universe is matter. In China, philosophy has mainly been founded on the understanding that the fundamental component of the universe is energy. Health and diseases can then be explained by flow or disturbance of the flow of energy.

In modern Western Medicine, health and diseases are no longer explained by philosophy, but by natural science. Treatments are aimed to be evidence based instead of merely empirically based. Cell biology is important in explaining diseases and treatments. This approach fits well for transition from Traditional to Modern Medicine in Europe. On the other hand, natural science is more than Newtonian based science. In a frame where the fundamental component of the universe is explained by energy, maybe parts of Traditional Chinese Medicine could be explained by natural science as well [9, 15]. Nevertheless, there are basic differences between WM, which is mostly analytical and quantitative, and TCM which is mostly holistic and qualitative.

1.8 Research on Traditional Medicine

Treatment with various forms of Traditional Medicine is offered both inside and outside the national health care systems worldwide. World Health Organization, WHO, has provided guidelines for how research and evaluation of Traditional Medicine should be carried out [16]. In Norway, The National Research Center in Complementary and Alternative Medicine, NAFKAM, was designated as a WHO Collaborating Centre for Traditional Medicine in 2008 [17]. NAFKAM is located at the University of Tromsø and is funded by the Norwegian Ministry of Health and Social Affairs. WHO and their collaborators aim to ensure that Traditional Medicine is used properly. All types of therapies should be evidence based, and Western scientific methods are promoted to assess the efficacy and safety of both Traditional Herbal Medicine and Traditional Procedure Based Therapies [4, 18]. Since 1997, the use of Complementary and Alternative Medicine has been documented in several reports both by NAFKAM and others in Norway [12].

WHO recommends Health Related Quality of Life (HRQoL) survey instruments to be used in research of Traditional Medicine, because such survey instruments will capture both positive and adverse effects of a treatment [16]. In Germany, as well as in the USA and Canada, acupuncture is approved for certain diseases by the private insurance companies [19].

1.9 Project summary

We have in a pilot study investigated the treatment effects in a TCM general care practice. The patients were treated by Jan Baak, who is an experienced physician who is educated in both Western and Traditional Chinese Medicine. 41 patients were studied retrospectively, and 7 patients were studied prospectively.

Patients who had finished their treatment at the time the current study started, were invited to participate in the retrospective part of the study. They were invited to fill out a standardized questionnaire after the treatment was finished. New patients registered after the start of the current study were invited to participate in the prospective part of the study.

They were invited to fill out standardized questionnaires before the first treatment and every fourth week after the first treatment. The responses from all respondents were registered in an anonymous and de-identified database and analysed by gender, age, symptoms and treatment.

The prospective group was followed over 9 months. The treatment effects of TCM was measured on two different scales, a 10 point Visual Analogue Scale and Health Related Quality of Life using the RAND Short Form-36.

1.10 Organization of the thesis

A brief summary of the statistical methods used in the project are given in Chapter 2: Aim and objectives of the study. The next chapter covers both review on literature about research on TCM and review of literature about HRQoL instruments. The Methods chapter covers a closer description of the methods used in the study. Both the methods and the results, which are presented in the two following chapters, are described at a level that should be possible for non-statisticians to follow.

Chapter 2: AIM AND OBJECTIVES OF THE STUDY

2.1 Aim

The aim of the study was to document the effectiveness of medical acupuncture in general care according to Traditional Chinese Medicine principles. The effectiveness was measured by use of scientifically established Western scientific methods. We assessed whether the health condition of the patient group was better (had improved), was worse or remained unchanged after the treatment period.

When health and health changes can be measured in an appropriate way, statistical analysis can be used as a tool to select best possible treatments for different conditions.

Better use of resources will reduce costs and efforts, and improve patient satisfaction and overall quality, as illustrated in Figure 2.

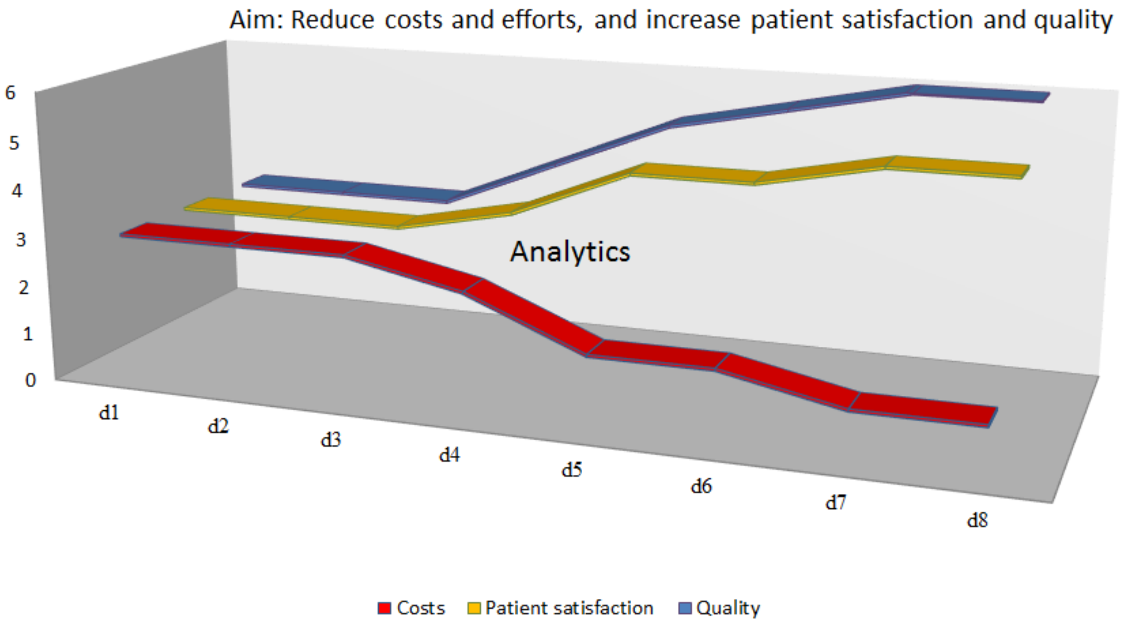


Figure 2. Illustration of how statistical analysis can be used as a tool to reduce costs and improve quality.

2.2 Objectives

The hypothesis for the study was:

- There is an improvement of the health situation after the treatment period, measured both by VAS and by HRQoL, and the improvement increases by number of treatments received.

Chapter 3: LITERATURE REVIEW

3.1 WHO recommendations on Research on Traditional Medicine

Before the first known detailed descriptions and classification of diagnosis and treatments ("Questions and answers by the Yellow emperor") [2], Traditional Medicine topics such as herbal medicine, acupuncture, physical exercise and diet habits had been passed on from one generation to the next for many thousands of years. Over the last 2000 years, many new written sources have been published. Traditional Chinese medicine has therefore greatly developed over the past 2000 years. It is often stated that the long term use of both herbal medicine and procedure based therapies are prove of both safety and efficacy, and the accumulated experience greatly exceeds the insight one can get by limited formal scientific studies. However, that does not take away that further studies are important. The extensive use of bloodletting in Europe in the past is an example of the necessity of closer investigation of long-term used Traditional Therapies [20].

WHO published in 2003 a list of diseases for which there is enough scientific proof to use TCM medical interventions. In the new forthcoming International Coding of Disease #11, TCM diagnoses are formally included. The WHO states that any form of medical interventions, also Traditional Chinese Medicine interventions, need to be investigated whenever possible, because some of the treatments may be efficacious, some are probably not, and some can be found to be harmful even if they have been used for a long time.

WHO has provided guidelines for how research and evaluation of Traditional Medicine should be carried out [16]. Randomized controlled clinical trial is regarded as the best possible study design, but other study designs, such as observational studies are also regarded as valuable.

Any medical intervention (also by TCM) may have both positive and negative impact on health. Health Related Quality of Life (HRQoL) survey instruments can be used in evaluation of Traditional Medicine, either together with biological measurements, together with other psychometric measurements or alone, because such survey instruments will capture both positive and adverse effects of a treatment [16].

Further recommendations from the WHO on research and use of Traditional Medicine are given in the WHO Traditional Medicine Strategy documents [4, 18]. Education and training of practitioners is promoted to ensure that therapies are used safely. Closer integration with national health care systems is promoted to ensure that useful interventions may be offered in a safe and cost-effective manner, and prevent the use of harmful or useless therapies

3.2 Research on Traditional Medicine in Norway

Until 2004, the use of folk medicine and Traditional Medicine was regulated in Norway through the "Medical Quackery Act" of 1936. The act was restrictive, and only physicians and dentists were allowed to prescribe drugs, and to perform surgical intervention or give injections or anesthesia. In 2004, the "Medical Quackery Act" was replaced with the less restrictive "Act on alternative treatment of diseases".

Although the law was restrictive, acupuncture with needles was introduced in Norway around 1970. In 1997 a committee was appointed by the Norwegian governmental Ministry of Health and Social Affairs to report on various aspects of alternative medicine. The committee, led by Professor Jarle Aarbakke, concluded that acupuncture was documented effective for some medical conditions, and probably effective for others [9]. After their recommendation, a center for research on alternative medicine was established in Tromsø in 2000. The center was organized as an independent unit under The Faculty of

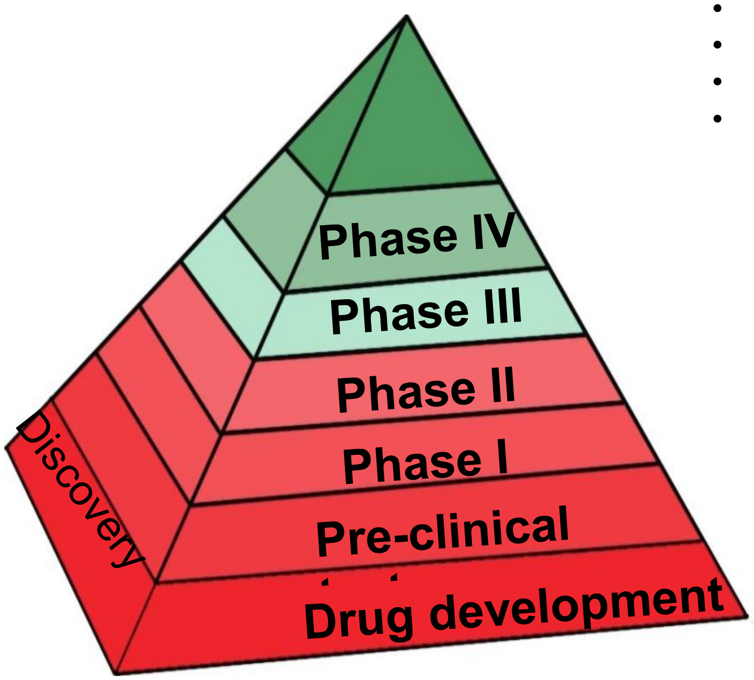
Medicine, and named The National Research Center in Complementary and Alternative Medicine, NAFKAM.

The center has hosted several international conferences on acupuncture, and in 2008 the center became a WHO Collaborating Centre for Traditional Medicine [17]. The report from the Aarbakke Committee has been follow up by several reports both by NAFKAM and others [12, 21].

Possible side effects of herbal medicine and dietary supplements are registered in the same way as possible side effects of commercial medicine in Norway, and information on known and potential side effects are made available to the public by regularly updates on the NAFKAM website.

3.3 Research on Efficacy and Safety of Traditional Chinese Medicine

Efficacy of a treatment refers to the capacity to save lives and improve health condition in human subjects, and the safety refers to the ability to do so without doing more harm than good. In modern Western Medicine, evaluation of new therapies are regulated by a well-defined set of steps, from laboratory tests, tests on animals, and small scale and large scale tests on human subjects. The steps are illustrated in Figure 3.



- Post marketing surveillance
- Full scale evaluation
- Initial clinical investigation
- Clinical pharmacology and toxicity

Altman D.G. Practical statistics for medical research. Chappman & Hall/CRC 1999
http://clinicaltrials.gov

Figure 3 The different phases of drug development [22].

The formal approach is in principle very suitable when the mechanism of the disease or treatment is well understood and can be explained by natural science, and also for TCM where the treatment mechanism is not so well understood. When the therapies are already in use, the steps of research, as described over, can be reversed. Only when the efficacy of a therapy can be documented, further effort should be made to understand the working mechanism of the treatment [10, 23].

Research on mechanism and components is very important in development of new drugs in the pharmacological industry. In TCM, however, the treatment is usually composed of various elements, and the treatment aims to improve the overall health, not only the remove the current expression of symptoms. Research on the system effect is therefore more appropriate than research on the component effect in TCM [9, 10]. Nevertheless, systems biology can be useful to get insight in the functioning and interaction of different various elements in TCM treatment.

The system effect of TCM treatments can be captured by Health Related Quality of Life surveys as is often done for western medical treatments. HRQoL measures the patient reported function and well-being in a “holistic” way, where not only physical functioning is covered, but also daily activities, vitality, social and emotional aspects. The questions in HRQoL surveys may be considered as an extension to the dialogue with the patient, only standardized and scored in a manner such that the patient’s feedback more easily can be used as a health outcome suitable for scientific research [23].

When HRQoL is used to measure the change in health status over time, both positive and negative effects are captured. The ability to catch up adverse effects is important when assessing the safety of a treatment.

Acupuncture is regarded as a safe treatment when given by trained practitioners [9, 24]. Both for acupuncture and other traditional treatments, in addition to possible adverse effects, the main threat is that necessary medical treatment is delayed if practitioners do not refer the patients to an appropriate physician when necessary. It is therefore important that practitioners of traditional medicine have basic medical knowledge and cooperate with the official health care system [25].

3.4 Assessment of Health Related Quality of Life instruments

Since 1970s, self-reported Health Related Quality of Life (HRQoL) has increasingly been used as a health outcome indicator [10, 26]. WHOQoL is a survey instrument with 100 questions developed and recommended by WHO. WHOQoL is not as widely used in health surveys as shorter survey instruments, because it is a challenge to recruit voluntary participants who are willing to spend more than 10-20 minutes to complete a questionnaire. A shorter form of the questionnaire with 26 questions is also available from WHO, but because the subset of questions are more focused on overall QoL than on health, WHOQoL-BREF is not the best choice when the effect of a treatment shall be evaluated [27].

For evaluation of TCM treatment, the Chinese ChQoL questionnaire with 50 questions is expected to be the best instrument to capture health changes [28]. The ChQoL questionnaire is developed based on the TCM understanding of health, and therefore covers aspects of health not covered by Western questionnaires, like a person's ability to adapt to climate and season changes. This may be valuable information for the TCM practitioner, and captures a wider aspect of the health improvements as understood by TCM. But because the TCM theory is commonly not well understood by European patients, most European patients will not be able to complete the questionnaire without instructions [29].

Treatment with TCM aims to improve the overall health condition of the patient, and not only reduce specific symptoms of diseases. This indicates that any HRQoL survey instrument could be used to measure the effect of the TCM treatment. The widely used SF-36 questionnaire does not measure the identical health categories as the ChQoL, but as it is found to correlate to ChQoL, and measures similar facets, it can be used instead of ChQoL to measure the effect of TCM treatment when the TCM theory is not so well understood [28].

Thus, among the HRQoL instruments considered, SF-36 is found to be the best instrument to measure the effect of TCM in general care. The benefits are that the survey instrument is validated for a broad range of medical conditions, cultures and languages. There are not too many questions, and the questions are easy to understand for patients. The greatest advantage of using SF-36 is that the outcome of the study can be compared to the outcome of other studies. As the outcome for TCM can be compared to treatments given by modern Western Medicine, SF-36 can be used as an instrument to distinguish between useful and useless interventions in TCM.

3.5 History of the SF-36 survey instrument

SF-36 is a standardized Health Related Quality of Life (HRQoL) survey instrument that is widely used both in Norway and internationally, and validated for different diseases, cultures and languages [30]. The survey instrument aims to measure the general health status, including physical, mental and social functioning. The instrument is the most used of the general health-status measures [31].

SF-36 is suitable both to compare groups and to measure changes in the same individual over time. Answers to questions are translated into scores by procedures described in SF-36 scoring instructions; with 0 being the lowest value and 100 the highest value. High values represent good function, health and quality of life [32]. There exists published norm data for the Norwegian population [33]. The effect of treatment can therefore be assessed against both the norm data and the baseline data.

The SF-36 short-form survey instrument with 36 questions was designed in the Medical Outcomes Study, MOS. This study was a continuation of the Health Insurance Experiment, HIE, a multi-year project where a range of scales were developed to measure health and health changes. The development of the assessed instruments are illustrated in Figure 4.

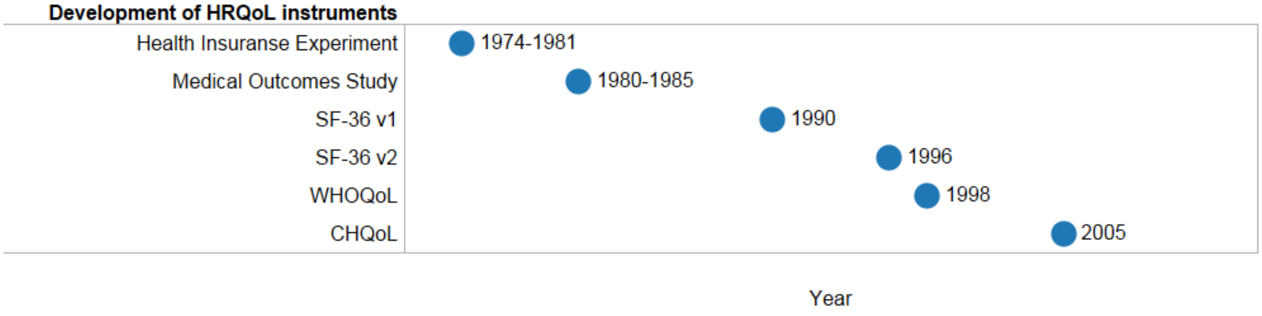


Figure 4. Development of some HRQoL instruments.

Version 2 of the SF-36 instrument covers several improvements in wording and scaling of the questions. The calculated scores will be more accurate in the second version, but the first version can still be used, and the results are comparable [34]. RAND SF-36 (Version 1.0), which is available online free of charge [35], has been used in the study.

Chapter 4: METHODS

4.1 Study design

4.1.1 Pilot study

In this pilot study we have investigated the treatment effects in a TCM general care practice. The participants in the study have filled out self-reported questionnaires, where they were rating the burden of their health problems on a VAS scale and filled out a standardized health survey questionnaire form to measure the general health condition. Changes in VAS and HRQoL scores are illustrated in Figure 5.

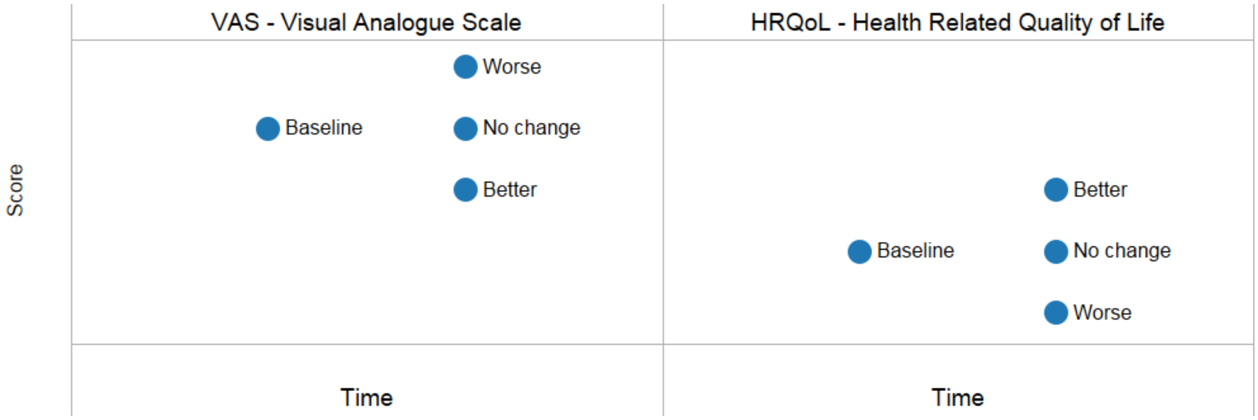


Figure 5. VAS and HRQoL measurements.

4.1.2 Study setting

The study has been a cooperation project between the Norwegian University of Life Science in the Oslo area, and the private clinic Dr. Med. Jan Baak AS in Tananger, Rogaland, Norway. The private clinic is a general care practice with approximately 400 patients. Ethical approval for the project was obtained from "Personvernombudet for forskning", Norsk Samfunnsvitenskapelig Datatjeneste AS (NSD).

4.1.3 Project schedule

The project was initiated the last quarter of 2013. For the prospective part of the study, patients were followed in the period from 01.01.2014 to 31.12.2014. For the retrospective part of the study, data from patients from 2012 and 2013 were collected the third quarter of 2014.

4.1.4 Selection of participants

The target group for the study was patients who voluntarily visited the clinic in the given period of time.

The inclusion criteria for the study were:

1. The patient was intellectually and mentally capable to provide independent written consent for participation in the study.
2. The patient was ambulant.
3. For patients under 16 years, both the patient and parents/guardians should give written consent for participation.

The exclusion criteria for the study were:

1. Pregnant women were not included.

Patients treated between 1.1.2012 and 31.12.2013 were invited to participate in the retrospective part of the study.

New patients during 2014 were invited to participate in the prospective part of the study.

4.1.5 Treatment

Treatment with TCM in Europe is mostly used in addition to, and not instead of Modern Western Medicine. The patient group in the study consists both of patients who got treatment with TCM alone, and of patients who got combined treatments. Information about which treatment the patients have received was collected from the patient records at the clinic.

The therapies used in TCM include both Herbal Medicine and Procedure Based Therapies. In the study group, classical acupuncture has been given in combination with Herbal Medicine and general health advices.

4.1.6 Data collection

New patients during 2014 were verbally informed about the study, and received written information about the study as well together with the first questionnaire. The patient group visiting the clinic come from different countries in Europe. To ensure that the information was understood by the patients, both the information brochure and the questionnaire were given in their preferred language: Norwegian, English, German or Dutch. The participants in the retrospective group received one questionnaire before treatment started, and follow-up questionnaires every fourth week in the following months.

Former patients from 2012 and 2013 were contacted by email and asked if they wanted to participate in the study. Information brochure, consent form and questionnaire were sent by post to those who wanted to participate.

The answers were returned in postage-paid envelopes. Every form was marked with a user code. The user code was used to connect the responses to information about given treatment from the medical records at the clinic.

4.1.7 Handling of collected data

De-identified data from the survey forms and from patient records from the clinic were added to a project database. The database has been used to perform different analyses and generating of reports. The project database will be erased when the project ends in 2015, as required by the permissions obtained for the study.

4.1.8 Missing data within a survey form

Most statistical methods assume that data sets are complete. In the scoring instructions for RAND SF-36, handling of missing answers and exclusion criteria are given. If more than 7 answers are missing in a form, the form shall be excluded. If less than 7 answers are missing in a form, the score for each missing answer shall be s estimated as the mean of the score of the other answers in the same health category.

4.1.9 Missing surveys forms

The Mixed Model which has been used for the prospective study can handle unbalanced design. This means that all the forms that are submitted can be included in the analysis. We will assume that data are missed by random for the analysis. However, this assumption will be discussed further in the Discussion section.

4.1.10 Reliability and Validity

A measure is said to have a high reliability if it produces similar results under consistent conditions. Both VAS and HRQoL instruments are found to be reliable in self-administered surveys [36].

A measure is said to be valid if it measures what it is meant to measure. Both VAS and HRQoL instruments are compared to other clinical measurements, and are found to be valid measurements of health concepts [36].

Reliability and Validity of the outcomes from the study will be discussed further in the Discussion section.

4.2 VAS psychometric response scale

In general practice, a wide variety of medical conditions are treated. It is therefore difficult or impossible to use disease specific measurements to measure the burden of the diseases. Hence, we had to use a general instrument for reporting of health problems. Visual Analogue Scale, VAS, is such a general instrument. The patients were asked to grade their health problems on a scale from 0 to 10 on a self-reported form, where 0 mean best possible condition and 10 mean worst possible condition. In addition to their specific health problems they were asked to grade their experience of pain and lack of energy in the same manner.

For the prospective study, it was expected that both the health problems and experience of pain and lack of energy would have decreased during the treatment period.

For the retrospective group, a low score would indicate that the treatment effect persisted after the treatment was finished.

4.3 Likert psychometric response scale

The Likert scale is named after its inventor, psychologist Rensis Likert. The respondents shall grade their attitudes to a series of statements on a symmetric agree-disagree scale. Each statement is referred to as a Likert item. Number of options for each item may vary. Possible options or categories for an item might be: 1-Not at all, 2-Slightly, 3-Moderately, 4-Quite a bit, 5-Extremely. The answer for a statement is scored on a numeric or reverse numeric scale. The series of the statements which are related are then grouped, and a Likert scale is calculated as a sum or a mean value for each group.

The direction of a Likert scale is often reversed compared to VAS. When used to measure Health and Quality of Life, a high score means a good condition, and a low score means a poor condition. As for VAS, it is difficult to interpret the absolute value of a Likert scale. Even when the options are symmetric and ordered, a value of 50 on a 100 point Likert Scale does not necessarily mean double as good as 25.

4.4 Short Form - 36 patient health questionnaire

The general health condition can be reported by the patients themselves or by the practitioner. In the retrospective group, the patients had finished the treatment, and in the prospective group, the patients should be followed up after the treatment was completed. The general health condition was therefore measured on self-reported survey forms. The SF-36 survey instrument was suitable in both the retrospective and the prospective part of the study.

4.4.1 SF-36 Health categories

A great effort was made during development of the SF-36 questionnaire to make scales that were balanced and easy to interpret [34, 37, 38].

There are 36 questions or Likert Items in the SF-36 questionnaire. The distance between the different options for each item is assumed to be equal. The values for each item can then be interpreted as interval-level data instead of ordinal data.

Each of the answers were first transformed to Item percentile scores, with a range from 0 to 100, where 100 is best. The answers were then grouped into 8 Health categories, and a mean value was calculated for each category. One of the questions is about change in health status, and because this is not related to any of the other questions, this question stands alone and is omitted from the calculation of the categories. The categories and number of questions forming each category are listed in Table 1.

Table 1. Construction of SF-36 Health categories and Summary categories.

8 Health categories	Number of questions	Summary categories	Number of questions
Physical functioning (PF)	10	Physical Component Summary (PCS)	21
Role functioning/physical (RP)	4		
Bodily Pain (BP)	2		
General health (GH)	5		
Energy/fatigue/Vitality (VT)	4	Mental Component Summary (MCS)	14
Social functioning (SF)	2		
Role functioning/emotional (RE)	3		
Emotional well-being/Mental Health (MH)	5		

4.4.2 Normative data from the general Norwegian population

Normative data from the general Norwegian population from 1998 was used to transform the raw scores to Norm based scores. SF-36 scores broken down by gender and 10 year age groups were extracted from Table III in the published article with the norm data [33]. The table consist of scores for the 8 health categories calculated from answers from 2323 respondents (66 % response rate, males and females between 19-80 years). The summary categories, PCS and MCS, were not published. The extracted dataset is shown in Table 2.

Table 2. Normative data for the general Norwegian population.
Each cell gives mean, standard deviation and number of persons for the health category by gender and age group.

	age_group / gender													
	29		30		40		50		60		70		ALL	
group	M	F	M	F	M	F	M	F	M	F	M	F	M	F
PF physical_ functioning	94,7 12,40 n=231	94 10,70 n=273	94,1 11,20 n=244	91,5 13,40 n=234	91,9 12,30 n=216	88,7 17,20 n=220	87,2 17,40 n=171	85,6 16,60 n=175	84,3 16,90 n=118	70,5 23,30 n=142	75 19,80 n=103	56,1 27,80 n=108	89,8 15,50 n=1083	84,8 20,80 n=1152
RP physical_role_ functioning	88,5 25,20 n=234	85,2 28,90 n=266	86,9 27,30 n=245	83,6 31,00 n=229	86,4 28,70 n=214	83 32,90 n=220	78 35,90 n=171	77,6 36,20 n=174	68,1 43,80 n=124	55,3 43,30 n=133	52,5 43,80 n=97	37 43,00 n=100	80,5 33,60 n=1085	75,4 37,70 n=1122
RE emotional_ role_functioning	84,4 29,70 n=233	78,9 32,30 n=265	86,9 27,20 n=245	81,4 33,80 n=226	89,2 26,00 n=212	84,1 30,70 n=219	87,5 27,90 n=168	84,3 30,90 n=173	78,6 31,90 n=120	74,5 38,50 n=130	69,7 37,60 n=94	59,5 44,20 n=97	84,5 29,70 n=1072	79,1 34,60 n=1110
BP bodily_pain	83,4 20,70 n=234	79,7 23,00 n=272	79,7 24,60 n=249	76,5 24,50 n=234	78,9 25,50 n=217	74,4 26,30 n=223	73,2 25,50 n=180	73,8 27,10 n=179	70,6 25,40 n=129	62,6 27,80 n=149	69,4 27,40 n=106	59,5 29,00 n=115	77,2 25,00 n=1115	73 26,60 n=1172
SF social_role_ functioning	88 20,00 n=235	85,3 19,30 n=273	89,7 17,70 n=250	84,5 22,80 n=236	87,6 20,90 n=220	85,7 24,70 n=225	86,5 24,10 n=181	86 21,30 n=181	89,3 20,20 n=131	81,5 22,70 n=152	82,3 23,80 n=110	74,1 28,70 n=117	87,6 20,90 n=1127	83,7 23,10 n=1184
MH mental_health	77,9 15,90 n=234	76,4 15,40 n=271	80 15,20 n=250	77,3 15,80 n=233	80,6 15,80 n=219	77,9 18,40 n=223	79,7 16,00 n=177	79,5 17,30 n=174	81,2 15,80 n=127	77,9 17,80 n=144	82,7 16,90 n=101	76,7 17,80 n=102	80 15,80 n=1108	77,6 17,00 n=1141
VT vitality	61,1 16,90 n=234	56,6 19,90 n=271	63,7 17,50 n=249	55,8 20,00 n=232	65,4 21,90 n=219	58,5 21,50 n=225	62,4 21,60 n=178	62 21,00 n=176	64,7 21,60 n=129	55,4 22,80 n=145	61,9 21,80 n=104	50,6 22,90 n=108	63,2 19,90 n=1113	56,9 21,20 n=1157
GH general_ health_ perceptions	83 16,20 n=229	82,5 18,00 n=269	81,1 19,20 n=244	80,5 19,80 n=233	79,3 21,20 n=214	79,3 22,90 n=218	74,1 22,50 n=173	74,7 22,40 n=162	68 25,10 n=117	63,1 25,10 n=137	67,5 22,60 n=95	62,5 22,10 n=92	77,4 21,30 n=1072	76,3 22,50 n=1111

Norm based scores provide a basis for interpreting HRQoL scores relative to the reference group, the general Norwegian population. Norm based scores are easier to interpret than the raw scores, because the health categories get the same mean value [39].

4.4.3 Transforming of SF-36 raw scores to Norm based scores

The mean values from the general Norwegian population were extracted from Table 2. The raw SF-36 scores were first transformed to Norm based z-scores, and then transformed to Norm based T-scores.

The formulas used are:

$$z_{ij} = \frac{x_{ij} - \mu_{jk_i}}{\sigma_{jk_i}}$$

$$t_{ij} = 50 + 10 * z_{ij}$$

where $i = 1, \dots, n$ correspond to person.

$j = 1, \dots, 8,$ is health category.

k_i is age and gender group for person i.

x_{ij} is the raw score for person i and health category j.

μ_{jk_i} is the population mean for health category j and age and gender group k_i.

σ_{jk_i} is the population SD for health category j and age and gender group k_i.

z_{ij} is the Norm based z-score for person i and health category j.

t_{ij} is the Norm based T-score for person i and health category j.

Some examples of raw scores transformed to z-scores and T-scores are given in Table 3. An average population standard deviation of 25 is used in the last example in the table, where gender and age is unknown. The last example is also visualized in Figure 6.

Table 3. Examples of transformation of raw SF-36 scores to Norm based scores.

Health Category	Gender	Age	Pop. mean	Pop. SD	Raw score		Norm based z-score	Norm based T-score
Bodily Pain	Male	20	83.40	20.70	Best	100	$\frac{100 - 83.40}{20.70} = 0.80$	$50 + 10 * 0.80$ $= 58.0$
Bodily Pain	Male	20	83.40	20.70	Mean	83.4	$\frac{83.40 - 83.40}{20.70} = 0$	$50 + 10 * 0 = 50$
Bodily Pain	Male	20	83.40	20.70	Worst	0	$\frac{0 - 83.40}{20.70} = -4.03$	$50 + 10 * (-4.03)$ $= 9.7$
Bodily Pain	Male	80	69.40	27.40	Best	100	$\frac{100 - 69.40}{27.40} = 1.12$	$50 + 10 * 1.12$ $= 61.2$
Bodily Pain	Male	80	69.40	27.40	Mean	64.9	$\frac{69.4 - 69.40}{27.40} = 0$	$50 + 10 * 0 = 50$
Bodily Pain	Male	80	69.40	27.40	Worst	0	$\frac{0 - 69.40}{27.40} = -2.53$	$50 + 10 * (-2.53)$ $= 24.7$
Bodily Pain	Unknown	Unknown	75.10	25.00		22.5	$\frac{22.5 - 75.1}{25} = -2.10$	$50 + 10 * (-2.10)$ $= 29.0$

4.4.4 SF-36 raw scores, T-scores and z-scores

SF-36 raw scores are measured in a range from 0 to 100, where 0 is worst and 100 is best. SF-36 z-scores are standardized scores with mean = 0 and SD = 1. SF-36 T-scores are z-scores which are scaled by 10 and shifted by 50. The T-score scale with mean 50 and SD 10 is widely used in psychometrics [39].

The raw SF-36 scores are measured in a scale that is easy to understand. A high score means good health, and a low score means poor health. When the dataset is transformed to Norm based z-scores, a negative score means health below the Norm. This negative health score is not easily understood by laymen, and Norm based scores are therefore usually reported as T-scores to the public. It is easier to communicate that a score below 50 means health below the Norm, as illustrated in Figure 6.

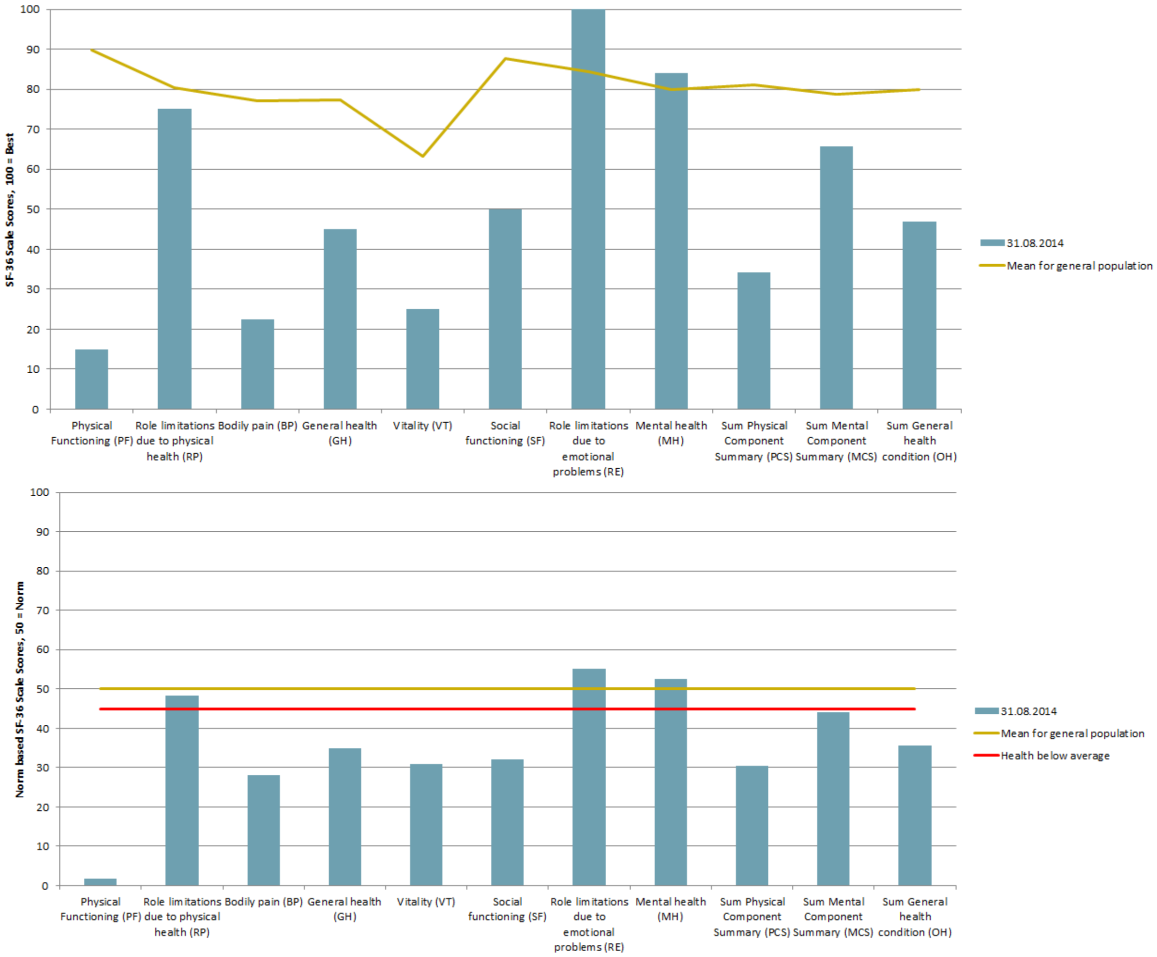


Figure 6. Visualization of SF-36 raw scores and SF-36 Norm based T-scores.

Raw SF-36 scores are shown as blue bars in the top pane, and norm based T-scores are shown as blue bars in the bottom pane. The mean for the general Norwegian population is shown as a green line. The red line indicates values below the Norm.

4.4.5 Construction of SF-36 summary scores

The summary scores, PCS and MCS, can be calculated by different formulas. The different formulas give different weights to the 8 health categories. The impact of different calculation rules is evaluated in several reports [38, 40].

The original weights derived from the US Norm data from 1998 are widely used internationally when country specific weights are not published. Briefly explained, the US Norm weights were calculated as factor scoring coefficients. The so called orthogonal series of weights were calculated by ignoring correlation between the summary scores. The so called oblique series of weights accounted for the correlation between the summary scores [41].

Country specific weights are not published for the Norwegian norm data. Instead of using the US Norm weights, the 8 health categories were given equal weights in the present study. The overall summary category (the mean of Physical and Mental health) does not have any practical interpretation, but was included in the study only as a reference value. Equal weights were given to each health category when the summary scores were calculated. The equal weights are illustrated in Table 4.

Table 4. Weights given to each SF-36 category.

8 health categories	Weights for Physical Component Summary, PCS	Weights for Mental Component Summary, MCS	Weights for overall summary category, HRQoL
Physical Functioning (PF)	0.25	0	0.125
Role-Physical (RP)	0.25	0	0.125
Bodily Pain (BP)	0.25	0	0.125
General Health (GH)	0.25	0	0.125
Vitality (VT)	0	0.25	0.125
Social Functioning (SF)	0	0.25	0.125
Role-Emotional (RE)	0	0.25	0.125
Emotional Well-Being (MH)	0	0.25	0.125

Physical Component Summary was calculated by

$$PCS = \frac{1}{4}(PF + RP + BP + GH)$$

Mental Component Summary was calculated by:

$$MCS = \frac{1}{4}(VT + SF + RE + MH)$$

HRQoL was calculated by:

$$HRQoL = \frac{1}{2}(PCS + MCS)$$

4.5 Effect size

The algorithms used to calculate effect size depends upon the design of the study. The effect size, or the standardized mean difference ES between two groups is defined by $ES = \frac{\bar{x}_1 - \bar{x}_2}{SD}$ where $\bar{x}_1$ and $\bar{x}_2$ are the group means and SD is the pooled standard deviation or the standard deviation for the reference group. For this study, SD was chosen to be the standard deviation for the reference group. When the standard deviation for the control group is used in the denominator, the calculated effect size can be called *Glass g* [42].

After conventional definition, an effect size above

A convention for the evaluation of effect size is given in Cohen [43] as

0.00-0.19 = No difference in group means,

0.20-0.49 = Small effect size,

0.50-0.79 = Moderate effect size,

0.80 and above = Large effect size.

4.5.1 Effect size for the retrospective group

For the retrospective part of the study, the mean of the observed scores were compared to the general Norwegian population. The formula for the effect size was then:

$$ES = \frac{\bar{x}_1 - \bar{x}_2}{SD} = \frac{mean\ of\ group\ scores - mean\ for\ the\ general\ population}{SD(the\ general\ population)}$$

When this approach is used for an example with group mean = 100, population mean = 83.4, and population SD = 20.7, the effect size can be calculated as:

$$ES = \frac{100 - 83.40}{20.70} = 0.80$$

When the scores are transformed to z-scores, the effect size can be calculated as

$$ES = \frac{0.80 - 0}{1} = 0.80$$

And when the scores are transformed to T-scores, the effect size can be calculated as

$$ES = \frac{58.0 - 50}{10} = 0.80$$

This illustrates that regardless of which scale the scores are measured in, the effect size will be the same. Effect size is therefore a measurement that is easy to interpret.

4.5.2 Effect size for the prospective group

For the prospective part of the study, the baseline observations were used as the reference group. The formula for the effect size was then:

$$ES = \frac{\bar{x}_1 - \bar{x}_2}{SD} = \frac{mean\ of\ follow\ up\ scores - mean\ of\ baseline\ scores}{SD(baseline\ scores)}$$

For example, when the mean of SF-36 follow up scores = -1.191, mean of baseline scores = -1.601, and SD of baseline scores = 0.80 the effect size can be calculated as:

$$ES = \frac{-1.191 + 1.601}{0.80} = 0.513$$

Another example, when the mean of VAS follow up scores = 4.783, mean of baseline scores = 6.700, and SD of baseline scores = 1.88 the effect size can be calculated as:

$$ES = \frac{4.783 - 6.700}{1.88} = -1.020$$

Both examples above represents an improvement, but because the VAS and SF-36 scores are measured on scales with different magnitudes, an improvement is measured as a positive effect size for SF-36 scores and a negative effect size for VAS scores. To avoid confusion when the results are presented, the sign of the effect size for VAS scores was changed, such that a positive effect size means an improvement in both cases.

4.6 Study outcome

As mentioned previously, the health condition of the patients was measured on two different scales, a 10 point Visual Analogue Scale and Health Related Quality of Life using the RAND SF-36 (Version 1.0) questionnaire. The effect of the treatment was measured as the change in VAS and HRQoL scores between baseline and the follow up 3 months later for the prospective group. The effect was measured as change in scores and effect size.

4.7 Statistical analysis

4.7.1 HRQoL as the response variable

HRQoL can be used as 1 single response variable, 8 independent response variables, or 1 multi-level response variable in statistical models.

A single response variable can be constructed as a linear combination of the 8 health categories, as described in 4.4.5 “Construction of SF-36 summary scores”.

In this study, 1 multi-level response variable was constructed by reformatting the dataset from wide to long format. Instead of 8 variables on a row for each observation, there were 8 rows of the new constructed variable for each observation. This is illustrated in Table 5 and Table 6.

Table 5. The response variables in a table in wide format.

Person	Time	PF	RP	BP	GH	VT	SF	RE	MH	More variables….
1.	1	…..	…..	…..	…..	…..	…..	…..	…..	xxx
1.	2	…..	…..	…..	…..	…..	…..	…..	…..	xxx
2.	1	…..	…..	…..	…..	…..	…..	…..	…..	xxx

Table 6. The response variables were rearranged to a multi-level variable in a table in long format.

Person	Time	Health.category	Score	More variables….
1.	1	PF	…..	xxx
1.	1	RP	…..	xxx
1.	1	BP	…..	xxx
1.	1	GH	…..	xxx
1.	1	VT	…..	xxx
1.	1	SF	…..	xxx
1.	1	RE	…..	xxx
1.	1	MH	…..	xxx
1.	2	PF	…..	xxx
1.	2	PF	…..	xxx
1.	2	RP	…..	xxx
1.	2	BP	…..	xxx
1.	2	GH	…..	xxx
1.	2	VT	…..	xxx
1.	2	SF	…..	xxx
1.	2	RE	…..	xxx
1.	2	MH	…..	xxx
2.	1	RP	…..	xxx
2.	1	BP	…..	xxx
2.	1	GH	…..	xxx
2.	1	VT	…..	xxx
2.	1	SF	…..	xxx
2.	1	RE	…..	xxx
2.	1	MH	…..	xxx

4.7.2 Linear regression models

Simple linear regression models were fitted to explore relationship between pair of variables.

Multiple linear regression models were fitted to explore relationship between one response variable and multiple regressors. In some of the cases, the explanatory variables were of interest for the analysis. In other cases, the explanatory variables were included as nuisance variables. The nuisance variables accounted for much of the variation in the dataset, but were not of interest for the analysis. Some of the explanatory variables were also included to avoid confounding.

4.7.3 Model selection criteria

In the cases where there were possible to fit a model with different number of explanatory variables, the Akaike Information Criterion, AIC, was used to select to best model. This criterion is designed to find the model that best fits the data while punishing models with many parameters, i.e., we try to avoid overfitting. AIC is calculated by the formula

$$AIC = 2k - 2\ln(L)$$

where k is number of parameters of the model and L is the likelihood function for the model. The model with lowest AIC is considered to be the best model.

4.7.4 Principal Component Analysis

Principal Component Analysis can be used to explore the variation in the dataset in a neat way. In PCA, a set of new variables are created as linear combinations of the original variables. The coordinate system is rotated in a way such that as much as possible of the variation in the dataset is kept in the first of the new created variables, Principal Component 1 (PC1). All of the next components are constructed to be orthogonal to the previous components and explain as much as possible of the remaining variability. The number of components created is equal to or less than number of the original variables.

The variation in the dataset is sensitive to the measurement scales of the data. If PCA is performed on the raw dataset, the analysis is performed on the covariance matrix. Scaling and centring of the variables prior to the analysis was performed to remove the problems following from variables measured on different scales. PCA on the standardized variables is the same as to perform the PCA analysis based on the correlation matrix, as was done in this thesis.

4.7.5 Mixed Model for handling of combined fixed and random effect terms

When the health status is measured for the same individual at different times, the measurements will not be independent. Standard statistical methods assumes that the measurements are independent, and they need to be extended when used for dependent data. The Mixed Model can handle the dependencies within the dataset when repeated measurements are made on the same individual. The Mixed Model can also handle unbalanced design, where observations are missing at various measurement points. The model also accounts for individual differences, as different intercepts (different baseline levels) and different slopes (different change over time) can be fitted for each individual.

Multiple regression terms can be added to the mixed effects model. The model with fixed and random factors can be described by the equation

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is a vector of the observed values,

$\boldsymbol{Z}$ is the design matrix for the random effects,

$\boldsymbol{u}$ is a vector of random effects which are allowed to vary between subjects,

$\boldsymbol{X}$ is the design matrix for the fixed effects,

$\boldsymbol{\beta}$ is a vector of fixed effects, and $\boldsymbol{\varepsilon}$ is the error term.

Assumptions for the model:

The random effects and the error term have a multivariate normal distribution,

$$\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{\varepsilon}\end{pmatrix} \sim N(\mu, \Sigma)$$

where

$$E\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \mathbf{0}$$
$$Var\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{R}\end{bmatrix}$$

It follows that:

$$E(\boldsymbol{y}) = \boldsymbol{X\beta}$$

$$Var(\boldsymbol{y}) = \boldsymbol{ZGZ'} + \boldsymbol{R}$$

An example of a model with random intercept (different baseline level) and random slope (different change over time) for different persons is:

$$y_{ij} = \beta_1 + \beta_2 Time_{ij} + b_{1i} + b_{2i} Time_{ij} + \varepsilon_{ij}$$

(Notation from *Applied longitudinal analysis, by* Fitzmaurice, G.M [44])

The regression line for the fixed effect terms gives an estimate of the expected value for the population: $y_{ij} = \beta_1 + \beta_2 Time_{ij}$

The random effect terms adds the deviation by person to the estimate of the expected value for the population: $y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})Time_{ij}$

The marginal model adds the mean of the deviation by person to the estimate of the expected value for the population: $y_{ij} = (\beta_1 + \overline{b_1}) + (\beta_2 + \overline{b_2})Time_{ij}$

Different approaches can be used to predict the outcome for future observations.

1. The random effects can be set to 0, so that only the fixed effect terms in the model are used for prediction.

2. Alternatively, the marginal model may be used for prediction, so that the average of the random effect terms are added to the fixed effect terms.

4.8 Software used in the project

The data has been manually collected and inserted into a Microsoft Office Excel workbook. The input forms and scoring algorithms have been programmed in Visual basic for Microsoft Office 2013. Further visual analyses and statistical modelling are performed in Tableau (v. 9.0), R (v. 3.1.3) and SPSS (v. 23).

Liner models have been fitted in R using *stats::lm*, and mixed models by *lme4::lmer*.

The Mixed model for the prospective group was modelled in SPSS by the Mixed command. The marginal model was modelled by the EMMEANS command.

Chapter 5: RESULTS FOR THE RETROSPECTIVE GROUP

5.1 Participant rate

Between 1.1.2012 and 31.12.2013 treatment of 309 medical cases were registered in the patient records. Out of the 309 cases, 42 were cases where a person had returned to the clinic to get a new treatment within the project period, which gives 267 registered patients in the project period. Out of the 267 patients, 13 were excluded because they did not fill the study criteria, and 1 patient was excluded because of missing contact information. All the remaining 253 patients were contacted and invited to participate in the study. 42 persons returned completed survey forms. 1 person decided to withdraw the delivered answer. 41 answers of 253 possible give a participant rate of 16 %.

Number of participants by gender are listed in Table 7.

Table 7. Number of participants by gender for the retrospective group.

Gender	Number of participants (n=41)	Percent
Female	27	66
Male	14	34

Number of participants by 10 year age groups are listed in Table 8.

Table 8. Number of participants by age group for the retrospective group.

Age group	Number of participants (n=41)	Percent
< 30	4	10
30-39	8	20
40-49	5	12
50-59	13	31
60-69	6	15
> 69	5	12

A participant rate of 16 % is a low participant rate. The Research Council of Norway published an article in the magazine "Forskning" in 2013 on low participation rate in surveys. Their conclusion was that most people are tired of surveys, but that it is still randomly who is responding and who is not responding. The selection can thus be representative although the participant rate is low [45].

5.2 Missing data

There were 1 missing answer in 1 of the survey forms for the retrospective group. No survey forms were excluded because of missing data.

5.3 Health problems

The 41 participants in the study group were treated for a broad range of health problems, from severe diseases to more everyday ailments. 7 out of the 41 person (18%) were treated for health problems that had lasted for less than 3 months. 34 persons (82%) were treated for health problems that had lasted for more than 3 months. On an average, the health problems had lasted in 6 years before the treatment period started. The persons were treated for 1 to 6 health problems each, with an average of 2 health problems.

5.4 Treatments

On an average, the participants received 4 treatments each. All of the persons were treated by classical acupuncture. Some of the patients got treatment with TCM in addition to treatment with Modern Western Medicine.

To measure the component effect of TCM, a specific treatment for a specified condition had to be isolated. This was not possible in the current study, and the effect of the TCM treatment is therefore measured as the system effect of being in a TCM treatment.

5.5 Response variables

HRQoL variables derived from the SF-36 survey form were used as response variables in the study.

5.5.1 Correlation between response variables

The correlation between the response variables are visualized in Figure 7.

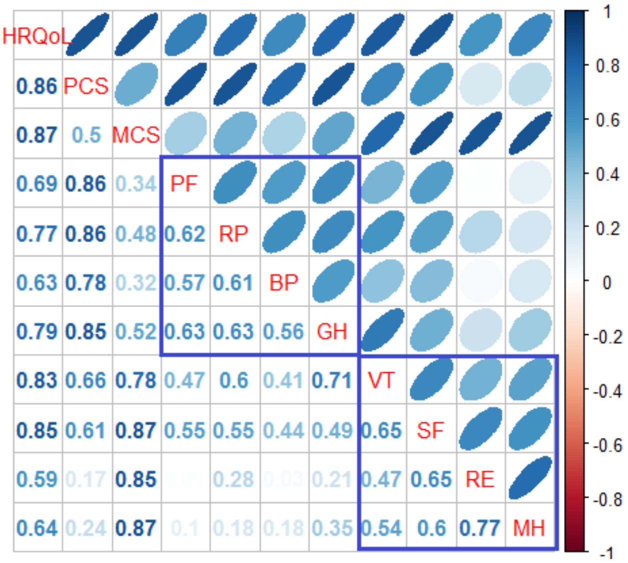


Figure 7. Correlation between the response variables for the retrospective group.

Empty cells have correlation close to 0.

HRQoL = Health Related Quality of Life, PCS = Physical Component Summary, MCS = Mental Component Summary, PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health, VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health

As expected, the categories which are summarized to Physical Component Summary: Physical functioning (PF), Role functioning/physical (RP), Bodily Pain (BP) and General health (GH), are highly correlated, and the categories which are summarized to Mental Component Summary: Energy/fatigue/Vitality (VT), Social functioning (SF), Role functioning/emotional (RE) and Emotional well-being/Mental Health (MH), are highly correlated. The summary variables (PCS, MCS, HRQoL) are of course highly correlated to the categories which makes up the sums.

5.6 Explanatory variables

Explanatory variables are described in Table 9. Gender, age, duration of the health problems and number of treatments were extracted from the patient records. The other variables were extracted from the survey forms.

Table 9. Explanatory variables for the retrospective group.

ID	Variable	Type	Range	Description
1.	Person	Number	1:41	Id for person
2.	Time	Number	1	Equal for all – 1 year after treatment
3.	Age	Number	16:80	
4.	Age group	Factor	6 levels	10 year age groups
5.	Gender	Factor	2 levels	Male/Female
6.	Problem mean	Number	1:10	Mean of Problem 1-6 measured on a VAS scale from 0-10, 10=worst
7.	Pain	Number	1:10	Pain measured on a VAS scale from 0-10, 10=worst
8.	Lack of energy	Number	1:10	Fatigue measured on a VAS scale from 0-10, 10=worst
9.	Number of complaints	Number	1:6	Number of health problems
10.	Chronic Acute	Factor	2 levels	Chronic or acute illness
11.	Duration	Number		Months ill before start of treatment
12.	Q1.answer	Factor	5 levels	Self-rating of health, Excellent-Poor
13.	Q2.answer	Factor	5 levels	Self-rating of health transition, Much better-Much worse that 1 year ago

5.7 Univariate statistical analyses

5.7.1 Self-rating of health, SF-36 question 1

In the first question in the SF-36 questionnaire the participants were asked to grade their own health in 5 levels. The results are presented in Table 10.

Table 10. Self-rating of health for the retrospective group.

In general, would you say your health is	Answer (n=41)	Percent
Excellent	2	4.9
Very good	12	29.3
Good	15	36.6
Fair	8	19.5
Poor	4	9.8

In the NAFKAM report from December 2014 the same item was graded for the general Norwegian population (n=1001). In that report, people who seek alternative therapies graded their own health about equal as those who do not seek alternative therapies.

In the Figure 8, the study group is compared to the general population and to the group who seek alternative therapies (CAM users) from the NAFKAM report. By inspecting the data visually, we can see that there are more in the study group who grade their own health as poor than in the general Norwegian population.

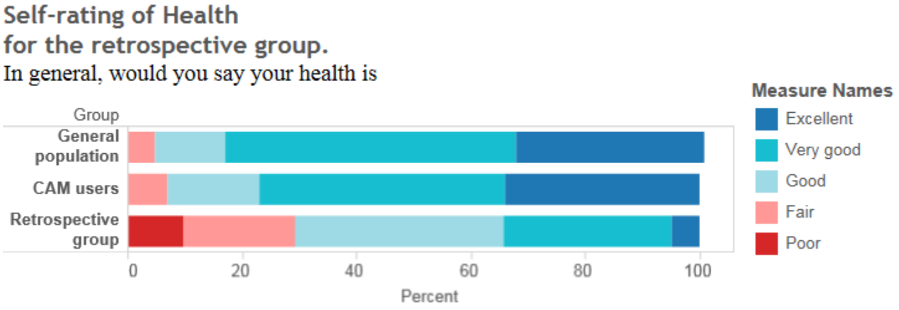


Figure 8. Self-rating of health, Retrospective group against general Norwegian population and CAM users in Norway.

5.7.2 Self-rating of health transition, SF-36 question 2

In the second question in the SF-36 questionnaire the participants were asked to grade their own health transition the last year in 5 levels. The results are presented in Figure 9.

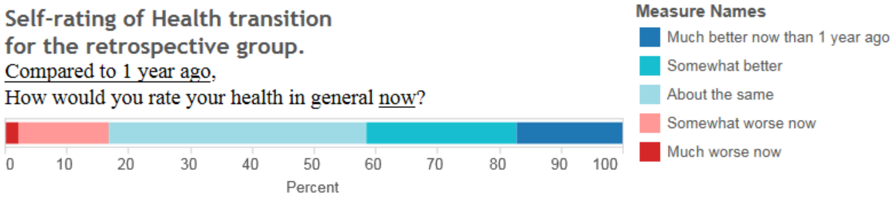


Figure 9. Visualization of Self-rated health transition for the retrospective group.

In a summary, 41,5% reported that they have got somewhat or much better health. 41.5% reported that their health was about the same as one year ago. 17% reported that they had got worse or somewhat worse health the last year. The wording of the question does not capture if the worsening or improvement was because of the treatment or for other reasons.

In the NAFKAM report from December 2014, 2% reported that they had got poorer health after they had received alternative treatments. The number is not directly comparable to SF-36 self-related health transition, because the wording of the questions are slightly different. The SF-36 item will capture both patients who have experienced adverse effects of the treatment, and patients who have got poorer health because of other reasons. Although adverse effects are not explicitly captured by the question, we cannot rule out that some of the patients have experienced adverse effects of the treatment.

We did not have any baseline measures for the retrospective group, and hence the effect of the treatment could not be measured by health transition. Instead of analysing the health transition, SF-36 mean scores after finished treatment were compared against Norm data for the general Norwegian population.

5.7.3 Visual Analogue Scale mean scores

The participants were asked to grade their health problems, their experience of pain and lack of energy on a 10-points scale. The participants had graded between 1 to 6 health problems each, and a mean value was calculated for each person. The means for the whole group are presented in Table 11 and Figure 10.

Table 11. Mean VAS scores for the retrospective group.

Scores Variable	Mean	SD	n
Mean of Health problems	3.1	2.9	41
Lack of energy	3.7	3.2	41
Pain	2.8	2.9	41

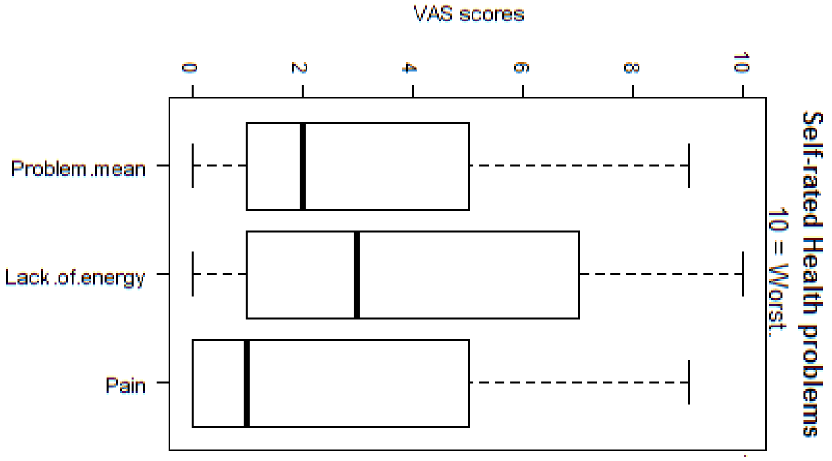


Figure 10. Mean VAS scores for the retrospective group.

5.7.4 Mean scores for the SF-36 Health categories

The observed mean scores, measured in T-scores and z-scores, and the difference in from the mean for the General Norwegian population are listed in Table 12.

On an average, SD for the normative data equals to 25. When the mean scores shall be translated back to difference in z-scores, the difference can be calculated as z-score*25=difference in SF-36 scores. For HRQoL this is $(-0.376) * 25 = -9.4$ scores.

On an average, the observed HRQoL is 9.4 scores below the mean for the general Norwegian population.

Table 12. SF-36 Norm based scores for the retrospective group.

Health category	Norm based scores, Group means		Difference from the Norm in raw scores
	T-scores (Norm = 50, SD = 10)	z-scores (Norm = 0, SD = 1)	difference = z-scores * 25
PF - Physical Functioning	47.49	-0.251	-6.3
RP - Role Physical	44.18	-0.582	-14.6
BP - Bodily Pain	50.59	0.059	1.5
GH - General Health	43.18	-0.682	-17.1
VT - Vitality	45.37	-0.463	-11.6
SF - Social Functioning	43.09	-0.691	-17.3
RE - Role Emotional	49.07	-0.093	-2.3
MH - Mental Health	46.96	-0.304	-7.6
PCS – Physical Component ¼(PF+RP+ BP+GH)	46.36	-0.364	-9.1
MCS – Mental Component Summary, ¼(VT+SF+RE+MH)	46.12	-0.388	-9.7
HRQoL – Health Related Quality of Life, ½(PCS+MCS)	46.24	-0.376	-9.4

5.7.5 Calculation of Effect size for SF-36 mean scores

As previously defined, the effect size for the SF-36 mean scores were calculated by:

$$ES = \frac{\bar{x}_1 - \bar{x}_2}{SD} = \frac{mean\ of\ group\ scores - mean\ for\ the\ general\ population}{SD(the\ general\ population)}$$

For the variable HRQoL, the effect size was calculated as: $ES = {}^{(46.24 - 50)}/_{10} = -0.376$

According to the definition of effect size in Cohen [43], the calculated effect size for HRQoL was within the range for small effects. This could be interpreted such that it was a small difference between the observed mean scores and the mean scores for the general Norwegian population.

In Table 12 on page 43, the Norm based T-scores were converted to norm based z-scores. In this case, where the general population was used as the reference group, the norm based z-scores were equal to the effect size, because the values for the reference group were standardized z-scores with mean = 0 and SD = 1. In Figure 11 on page 44, the norm based T-scores were compared visually to the mean for the general population. The T-scores for BP (Bodily Pain) and RE (Role Emotional) were within the range from 48 to 52 (which is equal to +/- 0.2 in z-scores and equal to +/- 5.0 in T-scores), and could be considered to be within the range for the mean of general Norwegian population.

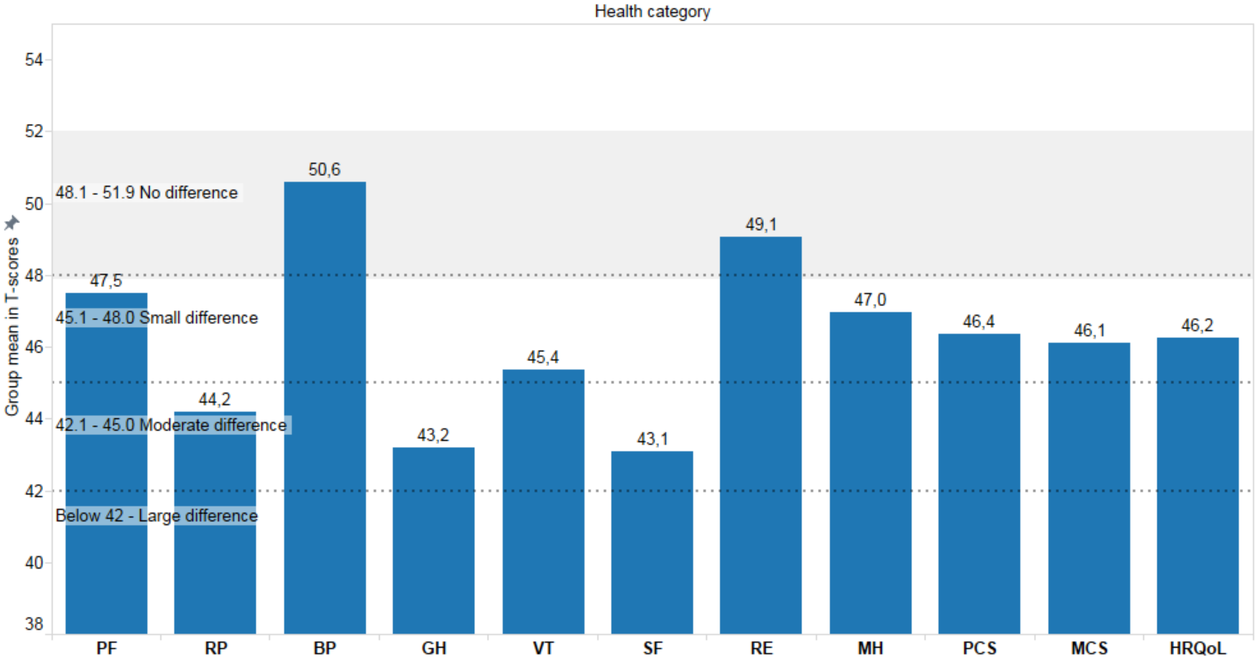


Figure 11. Visualization of SF-36 Norm based scores for the retrospective group. Higher is better, Norm = 50.

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health,
PCS = Physical Component Summary, MCS = Mental Component Summary, HRQoL = Health Related Quality of Life.

Difference in Norm based T-scores when the reference group is the general population:
0.0-1.9 = No difference, 2.0-4.9 = Small difference, 5.0-7.9 = Moderate difference, 8.0 and above = Large difference

5.7.6 Testing hypothesis about SF-36 mean scores

We wanted to test if the mean of the SF-36 health categories, HRQoL, was equal to the mean for the general Norwegian population or not. A two-sided test was performed on the difference. Usually a t-test is performed to test difference between group means, but as the standard deviation for the population was assumed known, we could perform a z-test for the hypothesis.

$$H_0: The\ expected\ difference\ between\ HRQoL\ and\ the\ general\ population\ is\ 0$$

$$d = E(\bar{x} - \mu_0) = 0$$

$$H_A: The\ difference\ between\ HRQoL\ and\ the\ general\ population\ is\ not\ equal\ to\ 0$$

$$d \neq\ 0$$

The summarized data for the test sample and the Norm data are given in Table 13.

Table 13. Summarized data for the general Norwegian population and the retrospective group.

	Mean value	SD	n
General Norwegian population	μ_0 = 50	σ = 10	(2311)
Retrospective group	$\bar{x}$= 46.24	(s = 9.2)	41

The test statistics was calculated by the formula:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

This is the same as the effect size calculated in the previous section, (5.7.5, Calculation of Effect size for SF-36 mean scores), multiplied by the square root of the sample size:

$$z = effect\ size * \sqrt{n} = (-0.376) * \sqrt{41} = -2.4$$

The significance level for the test was chosen to be 0.05. The critical value for the test was therefore $z_{0.05/_2} = 1.960$

The calculated test statistics, $|z| = 2.4$, was greater than $z_{0.25} = 1.960$. Hence the null hypothesis could be rejected with the significance level 0.05, and *we can state that Health measured by HRQoL differs significally from the mean for the general Norwegian population for the retrospective study group*.

This result is consistent with the findings in the Health Insurance Experiment, where persons with chronic diseases were found to score lower on HRQoL than the general population.

5.7.7 Power calculation

The significance level, or the probability of doing a type I error, for the hypotheses test was chosen to be 0.05. The power of the test, i.e., probability of rejecting H_0 if $\mu = \mu_1$ is

$$P\left(\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > z_{0.05/2}\right)$$

$$= P\left(Z > z_{0.05/2} - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right) + P\left(Z < -z_{0.05/2} - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)$$

$$= 1 - P\left(Z \leq 1.96 - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right) + P\left(Z \leq -1.96 - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)$$

R code was used to find the probabilities, and with the observed values, we found that the power of the test is

$$= 1 - pnorm(\,1.96 + 2.4) + pnorm(-1.96 + 2.4)$$

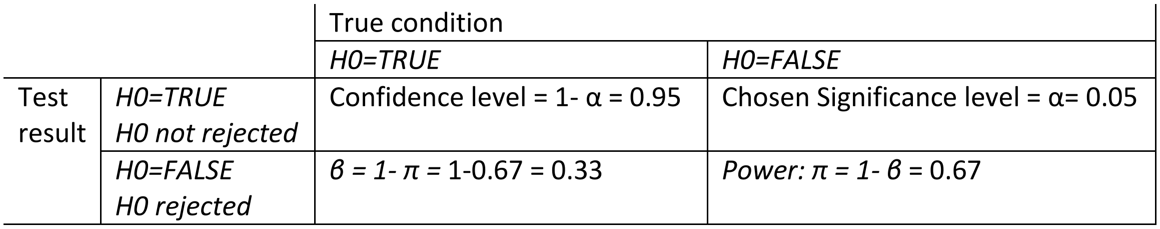
$$= 1 - 0.9999935 + 0.6700314$$

$$= 0.6700379$$

The power of the performed z-test was thus calculated to be 0.67.

The values for significance level and power for the test are illustrated in

Figure 12.



		True condition	
		H0=TRUE	*H0=FALSE*
Test result	*H0=TRUE* *H0 not rejected*	Confidence level = 1- α = 0.95	Chosen Significance level = α= 0.05
	H0=FALSE *H0 rejected*	*β = 1- π =* 1-0.67 = 0.33	*Power: π = 1- β =* 0.67

Figure 12. Values for significance level and power for the test illustrated for the retrospective group.

The power curve is plotted in Figure 13.

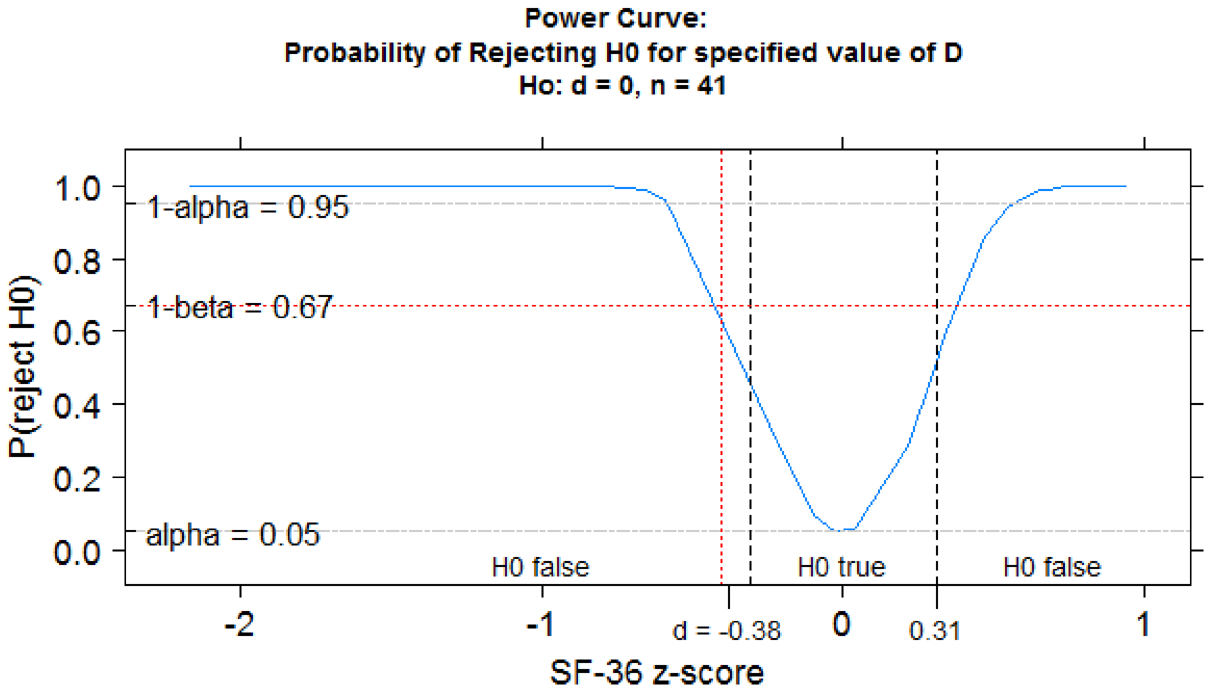


Figure 13. Power curve for the hypothesis test for the retrospective group.

The calculated power is below 0.8, which usually is considered sufficient. A rough estimate of number of participants necessary to reach the necessary power, is 56 persons under the same circumstances.

5.8 Bivariate statistical analyses

5.8.1 SF-36 mean scores compared to VAS mean scores

If SF-36 and VAS are measuring the same concept, the values should be correlated. SF-36 and VAS were measured in reversed scales, and we were expecting a negative correlation between the SF-36 scores and the VAS scores. As shown in Figure 14 and Figure 15 on page 49, the strongest correlation is between “Fatigue” (“Lack of energy”) and the SF-36 scores and the SF-36 summary scores. This indicates that “Lack of energy” measured on a VAS scale captures much of the same information as the SF-36 scores.

In Traditional Chinese Medicine, illness is explained by disturbance of the flow of energy in the energy channels of the body. It is therefore interesting that the data shows such a strong relationship between “Lack of energy” and HRQoL (r = -0.9). The strong relationship is illustrated by a simple regression model of “HRQoL” on “Lack of energy” in Figure 16 on page 50 (Multiple R-squared = r^2=0.8).

The correlation between Pain measured on a VAS scale and SF-36 Bodily Pain is illustrated in Figure 14 on page 49 (r = -0.6). The relationship is also illustrated by a simple regression model of Pain measured by VAS on SF-36 Bodily Pain in Figure 17 on page 50 (Multiple R-squared = r^2= 0.4). The SF-36 Bodily Pain variable was constructed out of two questions, and capture both the experience of pain and how much pain influences work activities. Pain measured on a VAS scale captures only the experience of pain. A moderate correlation between the two variables seems therefore to be reasonable.

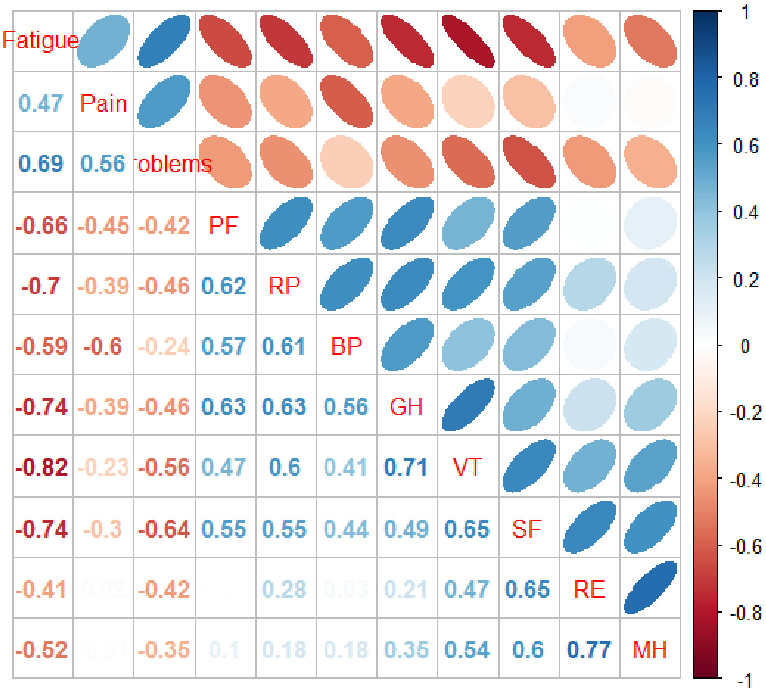


Figure 14. Correlation matrix for SF-36 and VAS scores for the retrospective group.

Empty cells have correlation close to 0.

SF-36 variables: PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health, VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health.

VAS variables: Fatigue=Lack of energy, "Pain", Problems=Mean of health problems.

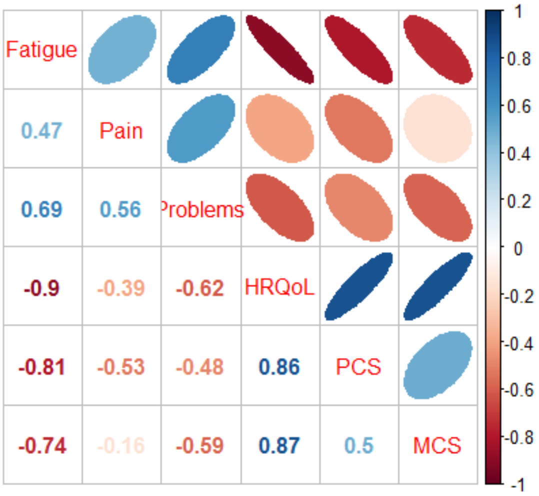


Figure 15. Correlation matrix for SF-36 summary scores and VAS scores for the retrospective group.

HRQoL = Health Related Quality of Life, PCS = Physical Component Summary, MCS = Mental Component Summary.

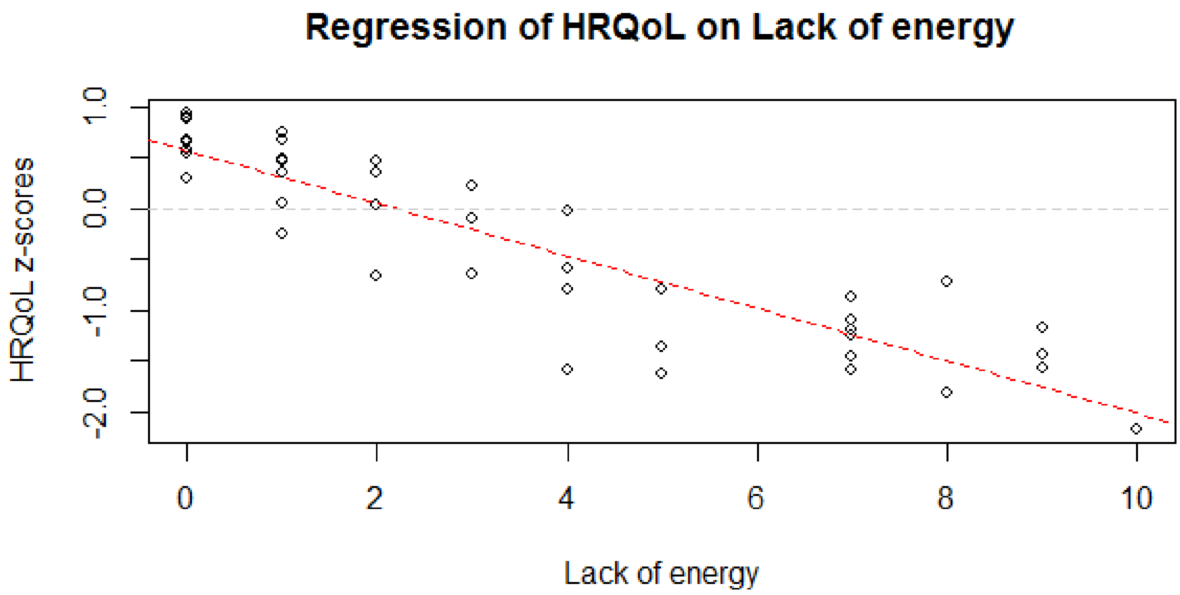


Figure 16. Relationship between SF-36 HRQoL z-scores and Lack of energy for the retrospective group.

(Multiple R-squared = 0.8) z-scores: higher is better, VAS: 0=Best, 10=Worst.

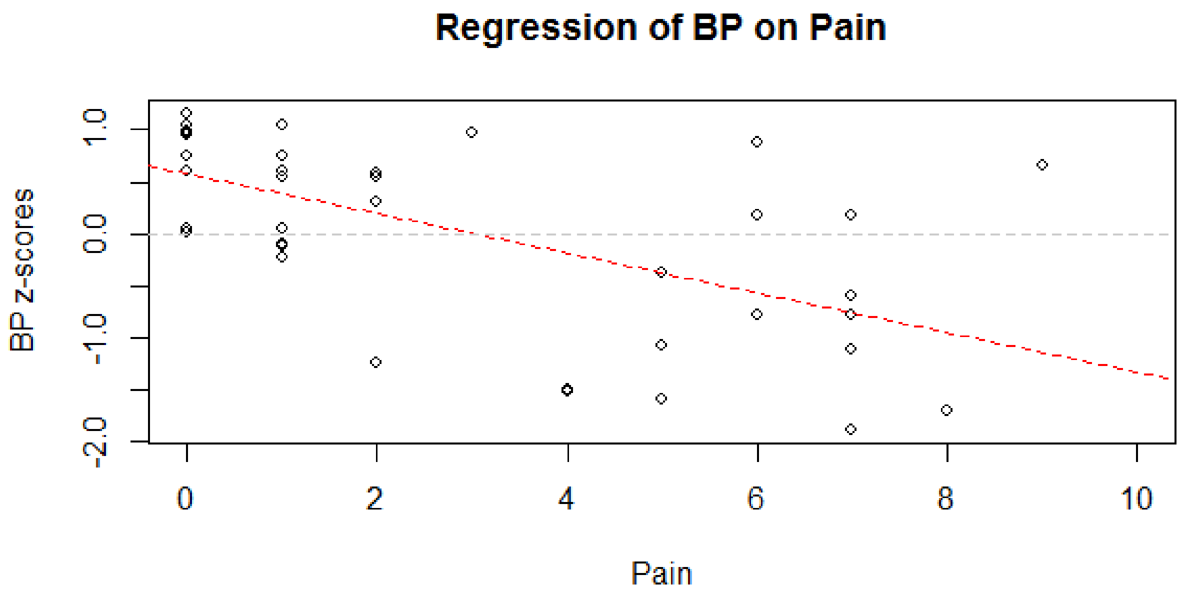


Figure 17. Relationship between SF-36 BP z-scores and Pain for the retrospective group.

(Multiple R-squared = 0.4) z-scores: higher is better, VAS: 0=Best, 10=Worst.

5.9 Multivariate statistical analysis

5.9.1 Principal Component Analysis on the response variables

PCA was performed on the response variables to reduce the number of response variables in the model. The variables were scaled and centered, such that the analysis was performed on the correlation matrix. Out of the 8 response variables there were created 8 Principal Components. The importance of the principal components are listed in Table 14.

Table 14. The importance of the Principal Components for the retrospective group.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.0705	1.3175	0.7344	0.6886	0.6585	0.5008	0.4113	0.3321
Proportion of Variance	0.5359	0.2170	0.0674	0.0593	0.0542	0.0314	0.0212	0.0138
Cumulative Proportion	0.5359	0.7528	0.8202	0.8795	0.9337	0.9651	0.9862	1.0000

The variance explained by the components are plotted in Figure 18.

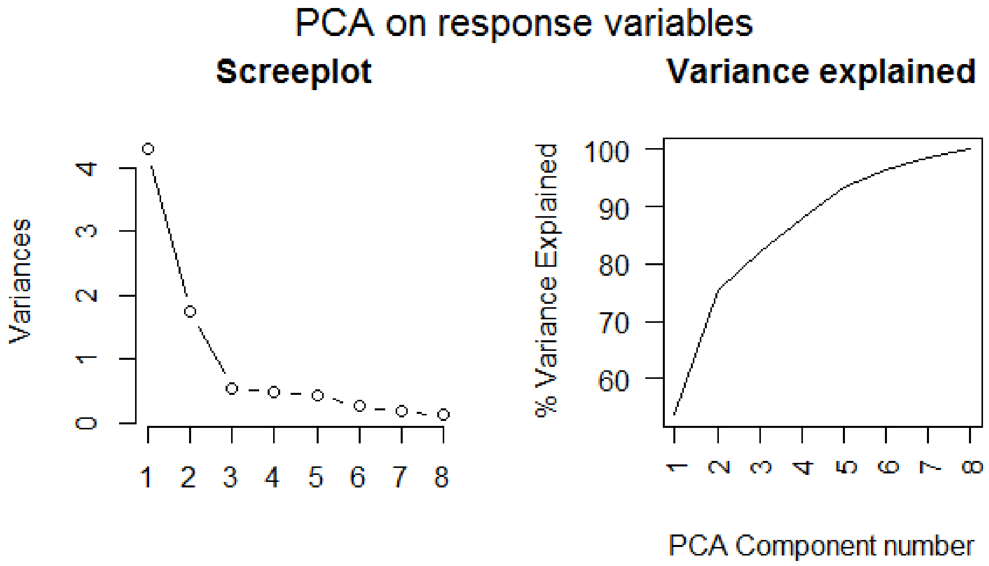


Figure 18. Variance explained by the PCA components for the retrospective group.

A biplot of the two first components is visualized in Figure 19.

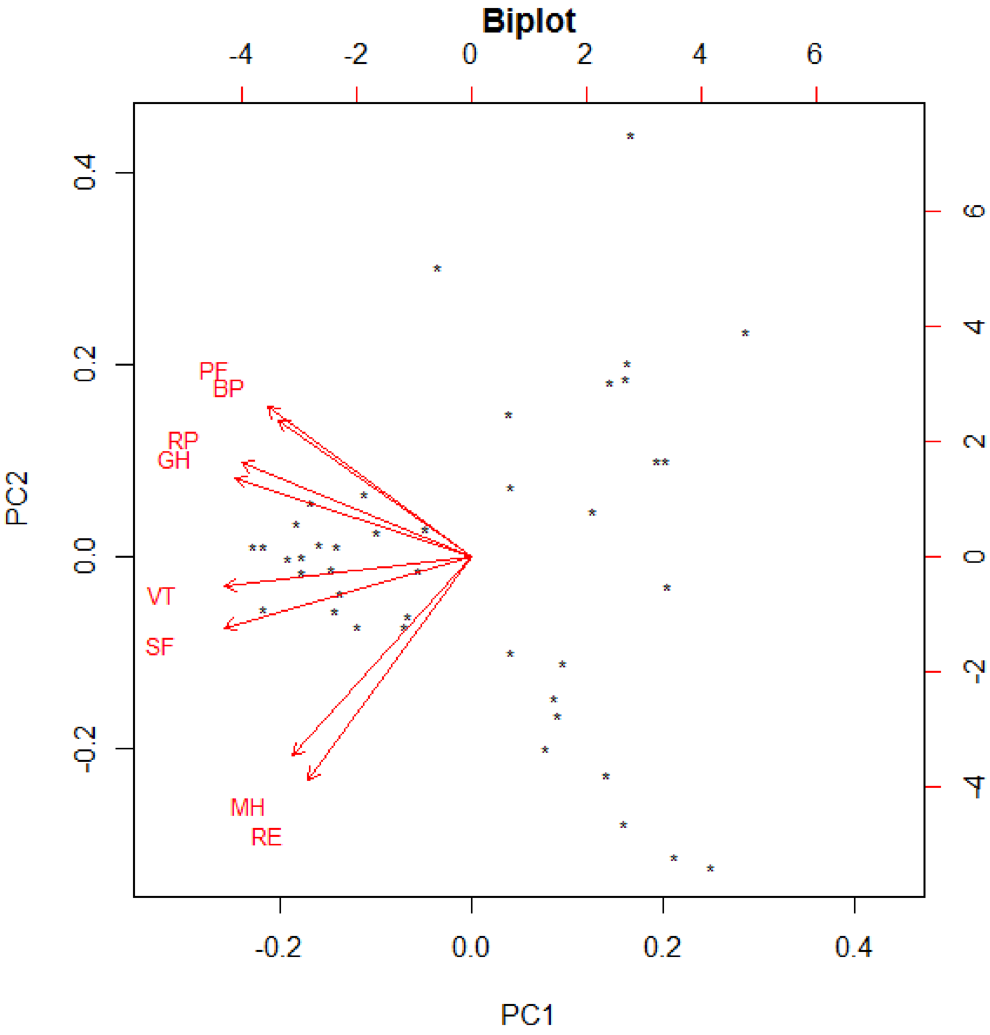


Figure 19. Biplot of the first two components for the retrospective group.

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health, VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health.

The component loadings, or the weights of the linear combination of the original variables, are listed in Table 15.

Table 15. Component loadings for the retrospective group.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
PF	-0.3367	0.3839	0.1748	0.4687	0.4725	0.1854	-0.3329	-0.3472
RP	-0.3787	0.2411	0.1183	0.1315	-0.7263	0.2384	-0.3037	0.3010
BP	-0.3192	0.3508	0.3825	-0.7349	0.0299	-0.1778	0.0723	-0.2257
GH	-0.3893	0.2012	-0.5664	-0.1165	0.1408	0.4430	0.4952	0.1109
VT	-0.4066	-0.0765	-0.5200	0.0765	-0.1217	-0.6805	-0.1748	-0.2103
SF	-0.4067	-0.1831	0.4247	0.2951	0.1787	-0.3142	0.4371	0.4603
RE	-0.2691	-0.5738	0.1858	0.0720	-0.2623	0.2694	0.2205	-0.6072
MH	-0.2945	-0.5122	-0.0561	-0.3335	0.3361	0.2157	-0.5256	0.3175

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health.

The loading for each of the variables in PC1 were in the range -0.3 to -0.4. This can be interpreted such that it was reasonable to give the variables equal weights in the construction of the overall mean, see section 5.7.4, Mean scores for the SF-36 Health categories.

The signs differ for the variables which are summarized into the PCS and MCS summary categories. This can be interpreted such that PC2 can be used to discriminant between Physical and Mental health.

If we should make a model where the principal component explained 90 % of the variation of the response variables, we had to select the 5 first principal components.

We wanted to reduce the number of response variables from 8 to a 1, and selected the first principal component as input to the regression model. PC1 explained 54 % of the variation of the response variables.

5.9.2 Principal Component Regression

We added PC1 to the dataset and performed multiple regression with PC1 as the response variable. In that way PC1 were substituting the 8 health category response variables. In the preceding chapter, the summary variable HRQoL was created by giving each of the 8 response variables weights based on established scoring rules [32]. By replacing HRQoL with PC1, the 8 response variables were given weights (loadings) based on the principal component analyses.

A linear regression model was designed with PC1 as the response variable and all the explanatory variables as regressors. Stepwise backward selection with AIC criteria was used to find the most significant explanatory variables. Outliers were detected, and the model selection process was repeated without those outliers. The explanatory variables with significance level below 0.05 were “Lack of energy” and “Number of health problems” in both cases, suggesting that the selected model was robust with respect to outliers.

The linear equation for the final model was

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where x_1 is the variable $Lack\ of\ energy$

and x_2 is the variable $Number\ of\ health\ problems$

and Y is Principal Component 1.

The regression coefficients were inserted to the model to predict HRQoL:

$$\widehat{HRQoL} = -PC1$$

$$= 2.2 - 0.6 * Lack\ of\ energy - 0.1 * Number\ of\ health\ problems$$

The predicted values for different cases of the model are listed in

.

Table 16. Predicted norm based SF-36 z-scores based on the PCA model.

Case			Predicted values of HRQoL based on the PCA model	
	Lack of energy	Number of Health problems	z-scores	T-scores
Best possible case	0	1	2.1	71
Mean case	3.5	2	-0.1	49
Worst possible case	10	6	-4.4	6

5.9.3 Dependencies within the dataset handled by the Mixed Model

It was expected that the levels of the 8 Health category response variables were dependent within person. A mixed model was designed to handle the dependencies within the dataset as described in section 4.7.5, Mixed Model for handling of combined fixed and random effect terms.

A new multi-level variable was constructed out of the 8 response variables. The dataset was reshaped from wide to long format with 8 rows for each person, one row for each category with the corresponding score value. The factors SF-36 Health category and Subject (Person) were fully crossed in the dataset, and the factors were modeled as random factors in the mixed model. The random factors with mean values and confidence intervals are visualized in Figure 20.

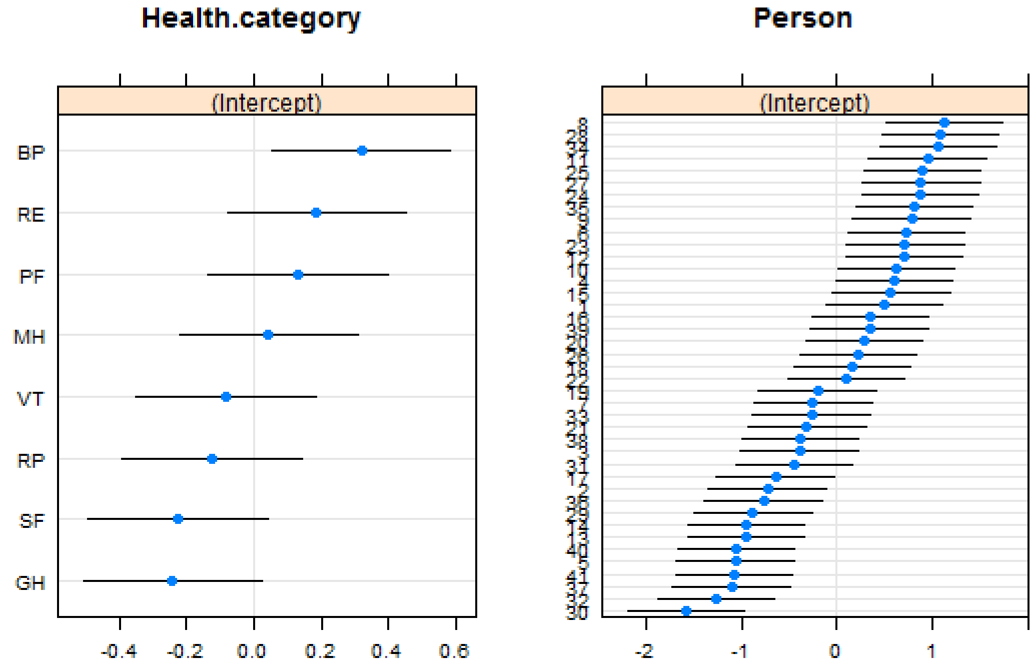


Figure 20. Predicted random effects of SF-36 Health categories and Person for the retrospective group.

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health.

5.9.4 Prediction of SF-36 Health category scores

The prediction of the SF-36 Health category scores with and without accounting for dependencies in the dataset are given in Table 17. The difference in the prediction for the two models was in the range +/- 0.1 z-scores, or +/- 1 T-scores.

Table 17. Predicted values for the SF-36 health categories, comparing lm and lmer methods for the retrospective group.

Health category	Intercept model fitted with *lme4::lmer* *Score ~ 1 + (1|Health.category) + (1|Subject)*	Intercept model fitted with *stats::lm* *Score ~1 + Health.category*	Difference in prediction of z-scores
	Predicted level accounting for dependencies within the dataset	Predicted level ignoring dependencies within the dataset (= estimates from)	
PF	-0.286	-0.251	0.0
RP	-0.524	-0.582	0.1
BP	-0.063	0.059	-0.1
GH	-0.597	-0.682	0.1
VT	-0.439	-0.463	0.0
SF	-0.603	-0.691	0.1
RE	-0.172	-0.093	-0.1
MH	-0.324	-0.304	0.0

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health.

All the possible explanatory variables were added to the model, and stepwise backward selection with AIC criteria was used to find the most significant explanatory variables. The explanatory variables with significance level below 0.05 were "Lack of energy" and "Number of health problems". This gives the same parameters as in the regression model described in 5.9.2.

The regression coefficients were inserted to the model to make a prediction of HRQoL:

$$\widehat{HRQoL} = 0.7 - 0.3 * Lack\ of\ energy - 0.1 * \ Number\ of\ health\ problems$$

The predicted values for different cases of the model are listed in Table 18.

Table 18. Predicted norm based SF-36 z-scores based on the Mixed model.

Case			Predicted values of HRQoL based on fixed effects of the Mixed model	
	Lack of energy	Number of Health problems	z-scores	T-scores
Best possible case	0	1	0.6	56
Mean case	3.5	2	-0.55	44.5
Worst possible case	10	6	-2.9	21

5.9.5 Comparing the PCR model and the Mixed Model

The PCR model and the Mixed Model were used to predict the mean level of the SF-36 Health categories. The predicted values for the models at different levels for the variable "Lack of energy" are visualized in Figure 21. The number of health problems is set to 2 in the visualization. Both models suggests that a patient who is scoring 4 or above on a 10 point VAS for "Lack of energy" has health below the mean for the general Norwegian population.

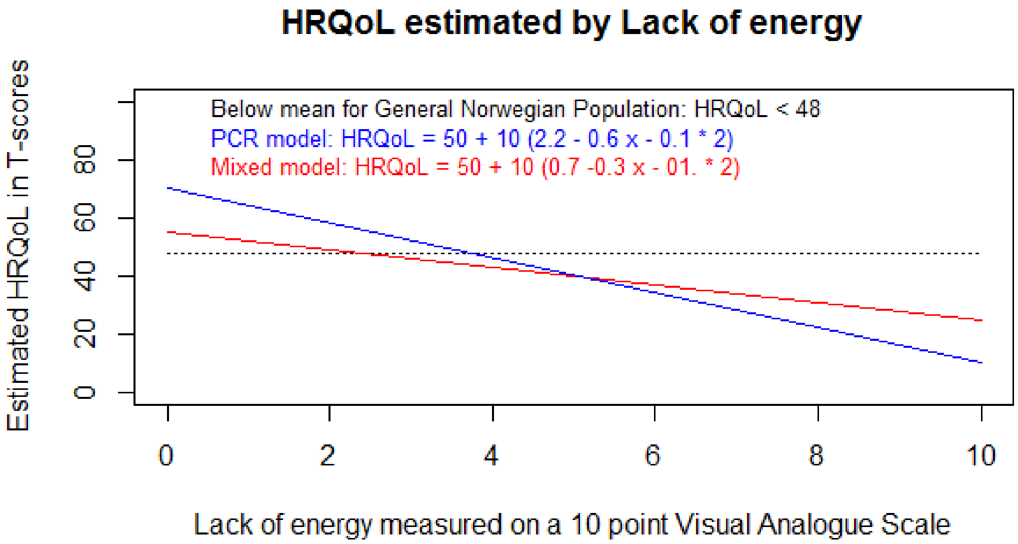


Figure 21. Comparing the PCR and the Mixed models for the retrospective group.

T-scores: higher is better, VAS: 0=Best, 10=Worst.

Chapter 6: RESULTS FOR THE PROSPECTIVE GROUP

6.1 Participant rate

Between 1.1.2014 and 31.12.2014 30 patients were asked to participate in the prospective group. 7 respondents out of 30 possible gives a participant rate of 23%.

The number of participants by gender are listed in Table 19.

Table 19. Number of participants by gender for the prospective group.

Gender	Number of participants (n=7)	Percent
Female	6	86
Male	1	14

The number of participants by 10 year age groups are listed in Table 20.

Table 20. Number of participants by age group for the prospective group.

Age group	Number of participants (n=7)	Percent
< 30	4	57
30-39	0	0
40-49	1	14
50-59	0	0
60-69	2	29
> 69	0	0

6.2 Missing data

There were no missing answers in the survey forms for the prospective group. No survey forms were excluded because of missing data.

6.3 Number of submitted survey forms over time

The schedule for the study was that the participants should fill out one survey form before start of treatment, and one survey form every fourth week in the following months. 7 persons (100 %) filled out the survey form before start of treatment, and 6 persons (86 %) submitted the first follow-up form during a period of 3 months. 4 persons submitted 2 or 3 more follow-up forms at various times. The submitted forms were grouped at 3, 6 and 9 months after start of treatment.

6.4 Health problems

All 7 persons had chronic health problems. The health problems had lasted from 5 months to 25 years before the treatment period started. The average duration of the health problems before the treatment started were 12 years. The persons were treated for 2 to 12 health problems each, with an average of 6 health problems.

6.5 Treatment

On an average, they received 6 treatments each. 4 of the persons were treated by classical acupuncture, and 3 of the persons got combined treatment with classical acupuncture and herbs.

6.6 Response variables

HRQoL variables derived from the SF-36 survey form were used as response variables in the study. Change in Health problems, Pain and Lack of energy measured on a Visual Analogue Scale were also investigated.

6.7 Explanatory variables

Explanatory variables are described in Table 21.

Gender, age, duration of the health problems and number of treatments were extracted from the patient records. The other variables were extracted from the survey form.

Table 21. Explanatory variables for the retrospective group.

ID	Variable	Type	Range	Description
1.	Person	Number	1:41	Id for person
2.	Form number	Number	1:4	Form number for each person
3.	Time	Number	0:9	0-9 months after first treatment
4.	Age	Number	16:80	
5.	Age group	Factor	6 levels	10 year age groups
6.	Gender	Factor	2 levels	Male/Female
7.	Problem mean	Number	1:10	Mean of Problem 1-12, measured on a VAS scale from 0-10, 10=worst
8.	Pain	Number	1:10	Pain measured on a VAS scale from 0-10, 10=worst
9.	Lack of energy	Number	1:10	Fatigue measured on a VAS scale from 0-10, 10=worst
10.	Number of complaints	Number	1:6	Number of health problems
11.	Chronic Acute	Factor	2 levels	Chronic or acute illness
12.	Duration	Number		Months ill before start of treatment
13.	Q1.answer	Factor	5 levels	Self-rating of health, Excellent-Poor

6.8 Univariate statistical analyses

6.8.1 Self-rating of health, SF-36 question 1

Recall that in the first question in the SF-36 questionnaire the participants were asked to grade their own health in 5 levels. The results for question 1 before the treatment started are presented in Table 22, and the results for the same question at the first follow-up are presented in Table 23.

Table 22. Self-rating of health before start of treatment for the prospective group.

In general, would you say your health is	Answer (n=7)	Percent
Excellent	0	0
Very good	0	0
Good	2	28,6
Fair	2	28,6
Poor	3	42,9

Table 23. Self-rating of health at first follow-up (3 months after first treatment) for the prospective group.

In general, would you say your health is	Answer (n=7)	Percent
Excellent	0	0
Very good	0	0
Good	2	33,3
Fair	3	50
Poor	1	16,7

6.8.2 Self-rating of health transition, SF-36 question 2

The second question in the SF-36 survey is about health transition the last year. However, because the prospective group was followed a shorter period than 1 year, this question was not suitable to investigate the health transition for the prospective group. Instead, the first question was investigated at two different occasions: Before start of treatment, and at the first follow-up 3 months later. The difference in SF-36 scores for question 1 are presented in Table 24 and also visualized in Figure 22.

Table 24. Health transition after 3 months for the prospective group.

	Before start of treatment, T-scores			At first follow-up, 3 months after first treatment, T-scores			Difference in T-scores	Difference in z-scores	Effect size $\frac{Difference\ in\ T\ scores}{SD_1}$	
	Mean	SD_1	n_1	Mean	SD_2	n_2				
SF-36 question1	21,429	22,493	7	29,167	18,820	6	7,738	0,774	0,344	Small difference

Effect size 0.00-0.19 = No change, 0.20-0.49 = Small effect, 0.50-0.79 = Moderate effect, 0.80 and above = Large effect.

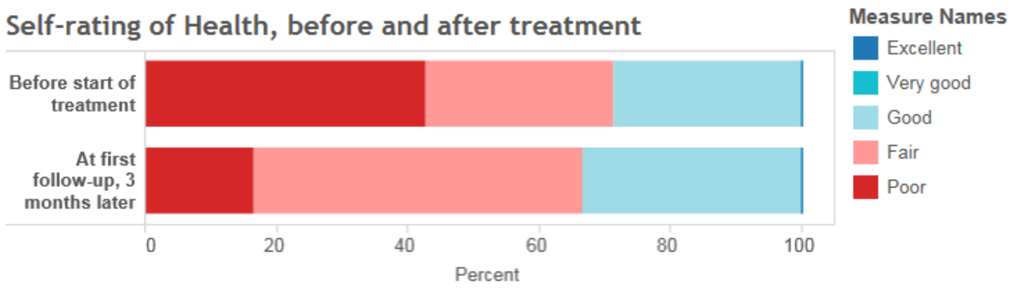


Figure 22. Visualizing of Health-transition for the prospective group.

6.9 Bivariate statistical analyses

6.9.1 Change in VAS scores by Time and Person

The participants were asked to grade their health problems, their experience of pain and lack of energy on a 10-points VAS scale. The participants had graded between 1 to 12 health problems each, and a mean value was calculated for each person. The change in VAS scores by Person is visualized in Figure 23. An improvement is visualized as a positive change.

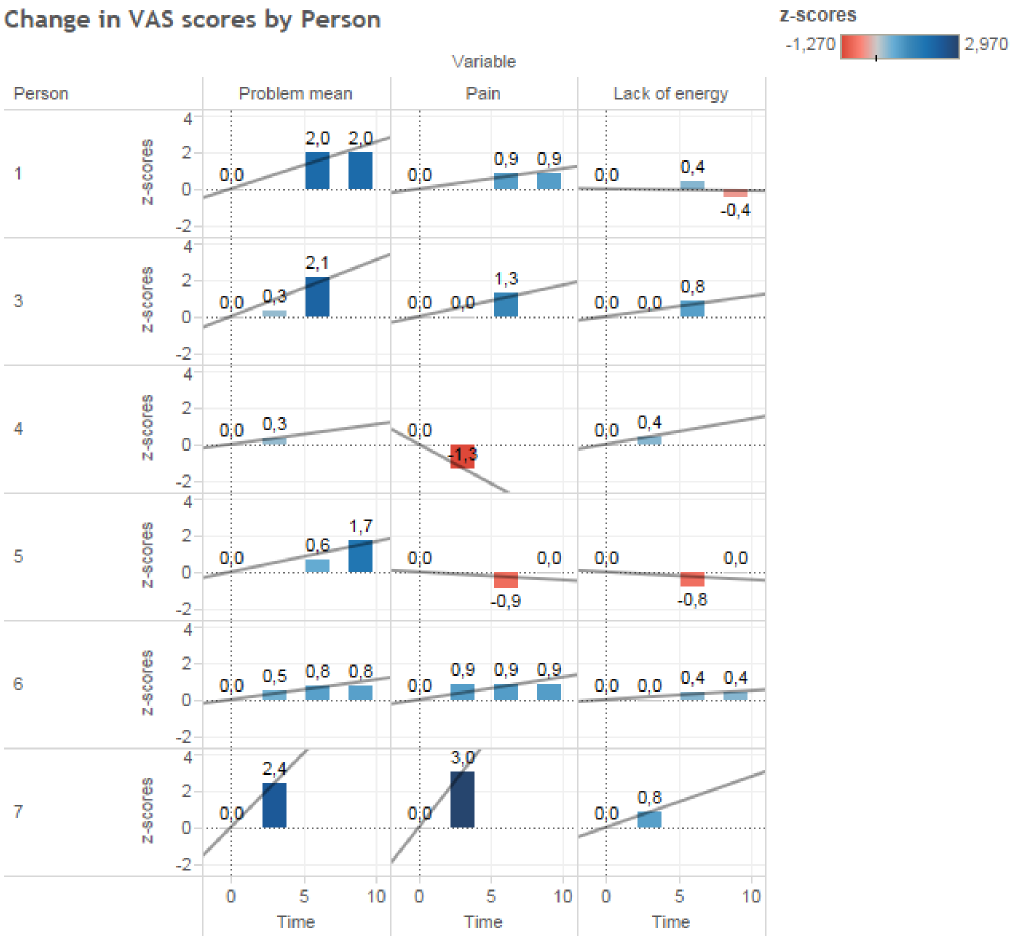


Figure 23. Visualization of change in VAS scores by Person for the prospective group. z-scores: higher is better.

"Time" is months after first treatment (0, 3, 6, 9 months).
Person 2 submitted only the first survey form, and no change in scores was observed.

6.9.2 Change in SF-36 scores by Time and Person

The change in SF-36 scores by Person is visualized in Figure 24, and the change in SF-36 summary scores by Person is visualized in Figure 25.

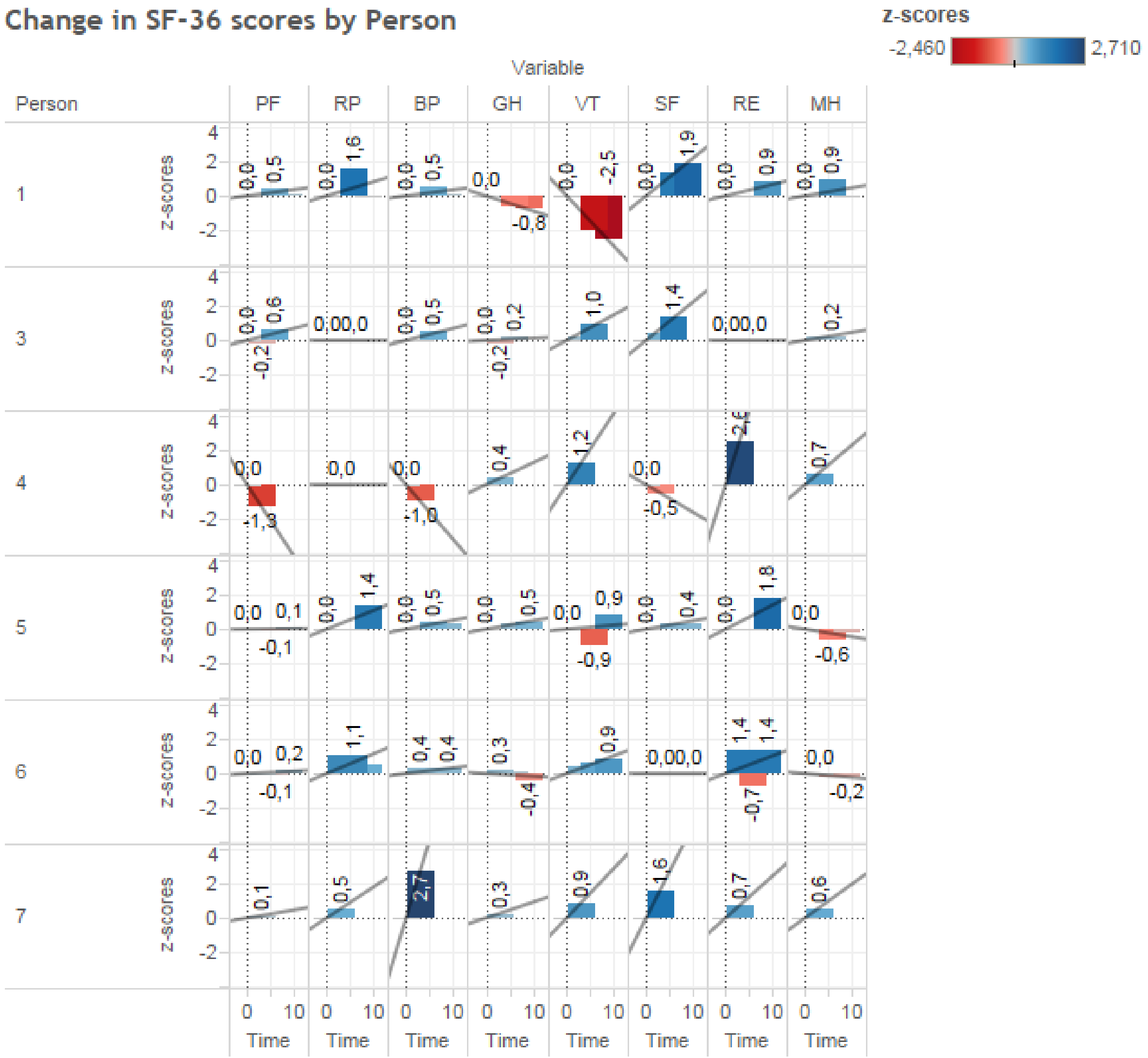


Figure 24. Visualization of change in SF-36 scores by Person for the prospective group. z-scores: higher is better.

"Time" is months after first treatment (0, 3, 6, 9 months).

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health.

Person 2 submitted only the first survey form, and no change in scores was observed.

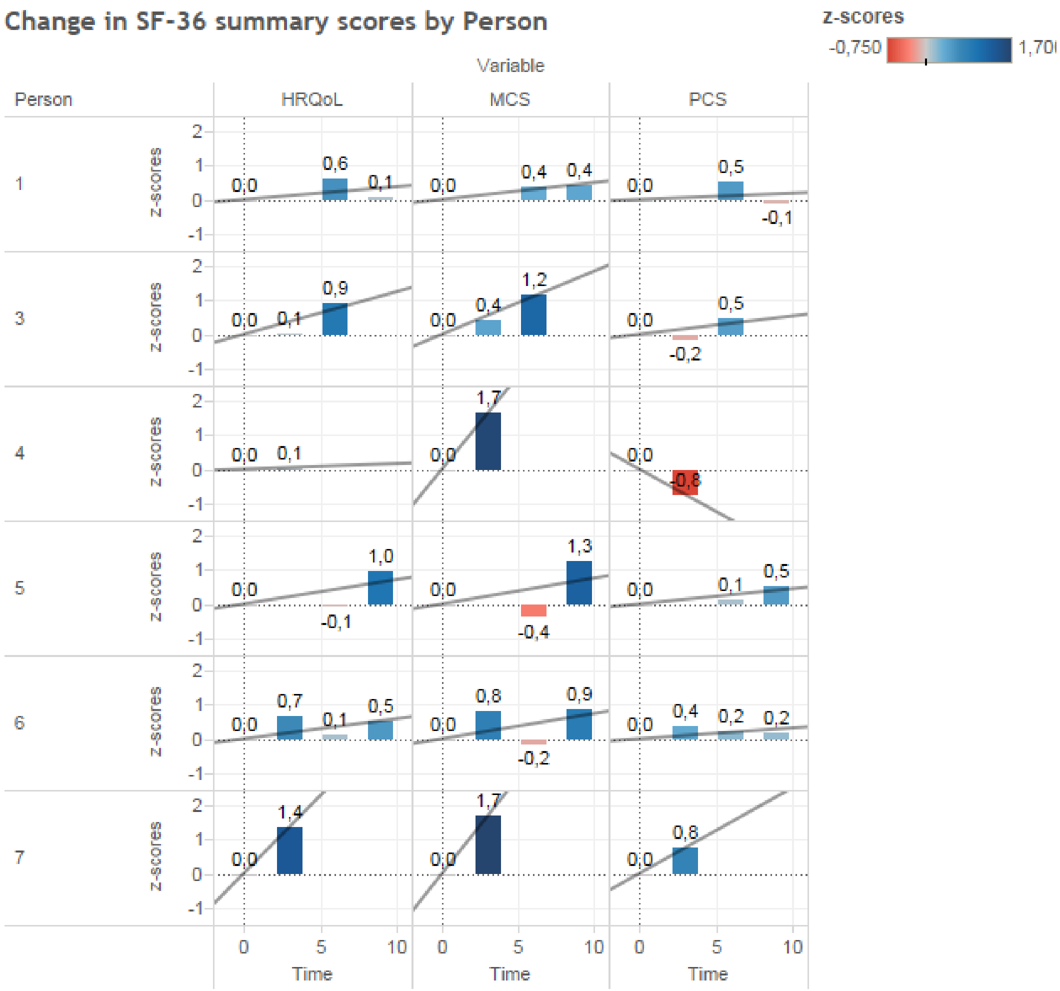


Figure 25. Visualization of change in SF-36 summary scores by Person for the prospective group. z-scores: higher is better.

"Time" is months after first treatment (0, 3, 6, 9 months).

*HRQoL = Health Related Quality of Life , PCS = Physical Component Summary, MCS = Mental Component Summary***.**

Person 2 submitted only the first survey form, and no change in scores was observed.

6.9.3 Change in group scores by Time

The change in scores in raw scores and effect size between baseline and the first follow-up for the prospective group are presented in Table 25 for the variables measured on a VAS scale. A large effect size (ES ≥ 0.80) was found for the mean of Health problems, and a small effect size (0.20-0.49) was found for Pain and Lack of energy. A positive effect size indicates an improvement.

Table 25. VAS scores at two occasions for the prospective group.

	Before start of treatment, Raw scores			At first follow-up, 3 months after first treatment, Raw scores			Difference in raw scores	Effect size $\frac{Difference\ in\ z\ scores}{SD_1}$	
Variable	Mean	SD_1	n_1	Mean	SD_2	n_2			
Mean of Health problems	6,700	1,880	7	4,783	1,127	6	1,917	1,020	Large effect size
Pain	6,286	2,360	7	5,167	2,137	6	1,119	0,474	Small effect size
Lack of energy	5,571	2,440	7	5,000	2,366	6	0,571	0,234	Small effect size

The change in z-scores, T-scores and effect size between baseline and the first follow-up for the prospective group are presented in Table 26 for the SF-36 variables.

Small to moderate effect sizes (0.20-0.79) were found for the SF-36 summary categories (HRQoL, PCS, MCS).

No difference was found for Physical Functioning, Vitality and Mental Health, and small to moderate effect sizes (0.2-0.79) were found for Role Physical, Bodily Pain, General Health, Social Functioning and Role Emotional.

Table 26. SF-36 scores at two occasions for the prospective group.

Variable	Before start of treatment, z-scores			At first follow-up, 3 months after first treatment, z-scores			Difference in z-scores	Difference in T-scores	Effect size $\frac{z\ scores\ diff}{SD_1}$	
	Mean	SD_1	n_1	Mean	SD_2	n_2				
PF	-2,889	3,000	7	-2,723	3,200	6	0,165	1,65	0,055	No change
RP	-1,876	1,080	7	-1,522	1,326	6	0,354	3,54	0,328	Small effect
BP	-1,816	1,030	7	-1,098	1,703	6	0,717	7,17	0,696	Moderate effect
GH	-1,699	1,440	7	-1,477	1,749	6	0,222	2,22	0,154	Small change
VT	-1,354	1,020	7	-1,242	0,756	6	0,113	1,13	0,110	No change
SF	-1,509	1,360	7	-0,667	0,883	6	0,842	8,42	0,619	Moderate effect
RE	-1,047	1,200	7	-0,362	1,127	6	0,685	6,85	0,571	Moderate effect
MH	-0,619	1,130	7	-0,435	0,983	6	0,184	1,84	0,162	No change
PCS	-2,070	1,390	7	-1,705	1,889	6	0,365	3,65	0,262	Small effect
MCS	-1,132	0,680	7	-0,676	0,670	6	0,456	4,56	0,670	Moderate effect
HRQoL	-1,601	0,800	7	-1,191	1,021	6	0,410	4,10	0,513	Moderate effect

Effect size 0.00-0.19 = No change, 0.20-0.49 = Small effect size, 0.50-0.79 = Moderate, 0.80 and above = Large.

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health,
PCS = Physical Component Summary, MCS = Mental Component Summary, HRQoL = Health Related Quality of Life.

The effect sizes are visualized in Figure 26.

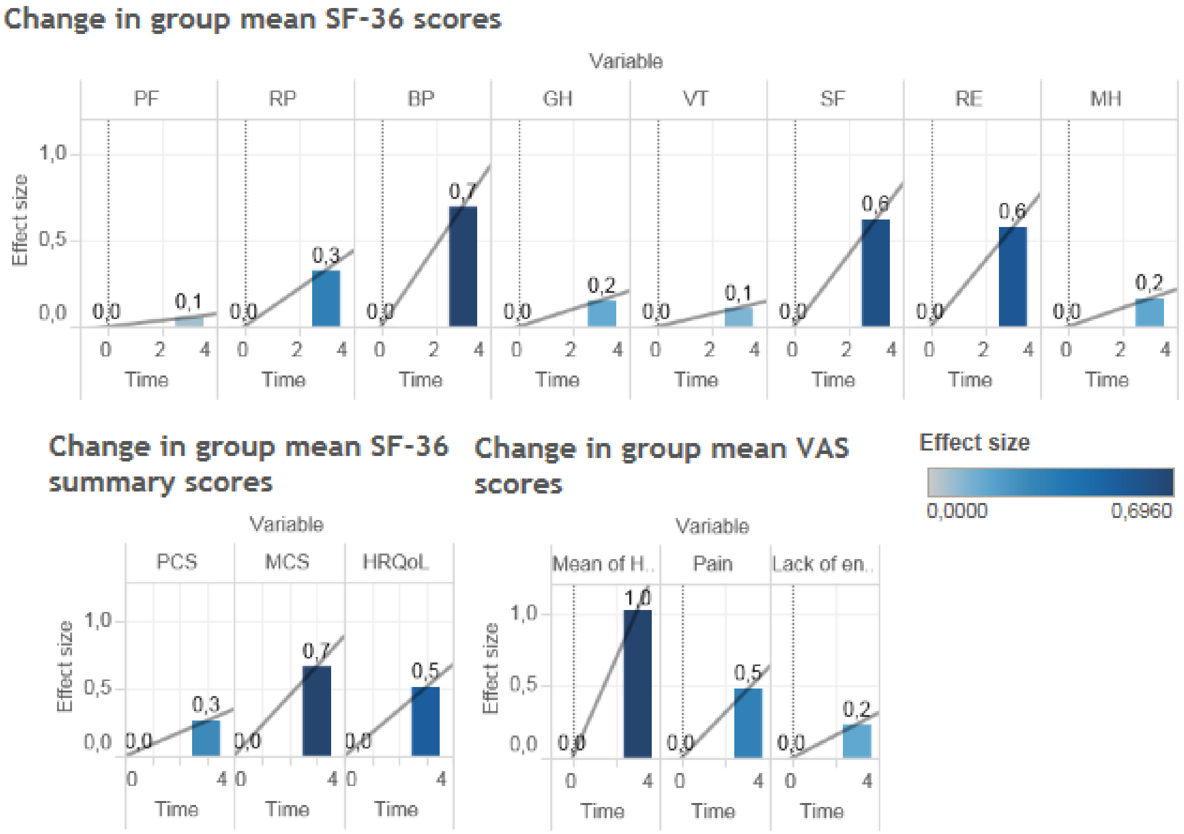


Figure 26. Visualization of effect size for SF-36 scores, SF-36 summary scores and VAS scores for the prospective group. Effect size: higher is better.

A positive effect size indicates an improvement.

"Time" is months after first treatment (0 and 3 months).

Effect size 0.00-0.19 = No change, 0.20-0.49 = Small effect size, 0.50-0.79 = Moderate, 0.80 and above = Large.

SF-36 variables: PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health,
VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health,
PCS = Physical Component Summary, MCS = Mental Component Summary, HRQoL = Health Related Quality of Life.

VAS variables: Lack of energy, Pain, Problems=Mean of health problems.

The changes in T-scores, are visualized in Figure 27.

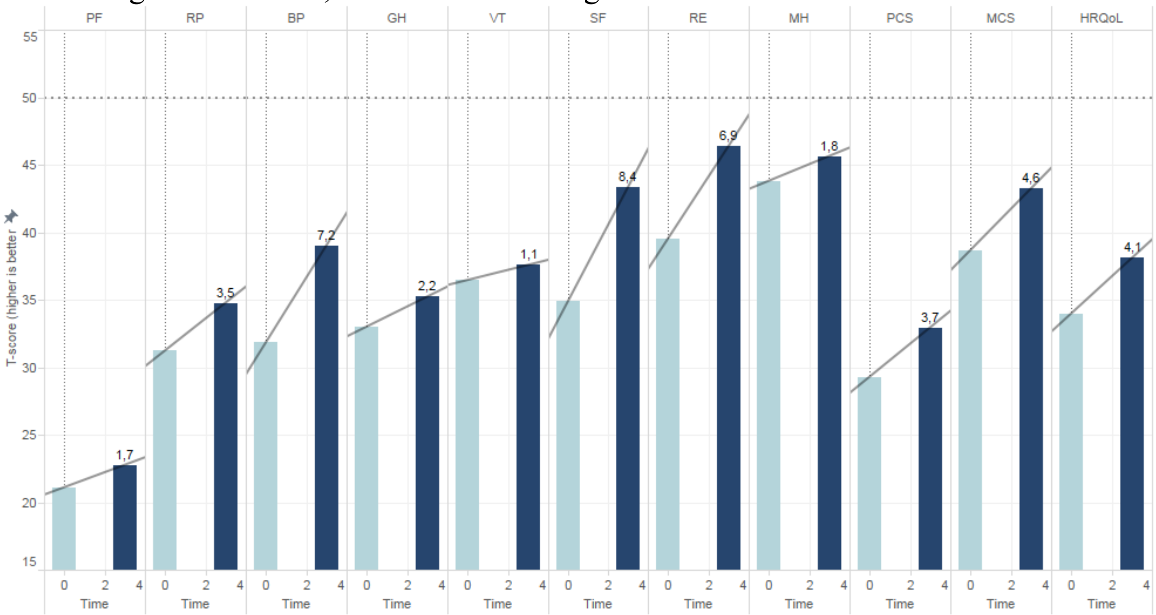


Figure 27. Visualization of change T-scores for SF-36 Norm based scores for the prospective group. T-scores: higher is better.

PF = Physical Functioning, RP = Role Physical, BP = Bodily Pain, GH = General Health, VT = Vitality, SF = Social Functioning, RE = Role Emotional, MH = Mental Health, PCS = Physical Component Summary, MCS = Mental Component Summary, HRQoL = Health Related Quality of Life.

6.9.4 Testing hypothesis about change in Health problems by Time

A one sided t-test was performed to test if the change in Health Problems, measured on a 10 point VAS scale, was greater than 0. The change in Health problems was measured as the difference between the baseline scores and the scores at the first follow-up, 3 months later.

$$H_0: The\ change\ in\ Health\ problems\ from\ baseline\ to\ follow\ up\ 1 = 0$$

$$d = E(\bar{x}_2 - \bar{x}_1) = 0$$

$$H_A: The\ change\ in\ Health\ problems\ from\ baseline\ to\ follow\ up\ 1 > 0$$

$$d > 0$$

The test statistics was calculated by the formula:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SD(\bar{x}_1)/\sqrt{n}}$$

This is the same as the effect size calculated in the previous section, (6.9.3, Change in group scores by Time), multiplied by the square root of the sample size.

$$t = effect\ size * \sqrt{n}$$

$$= 1.0 * \sqrt{6} = 2.45$$

The significance level for the test was chosen to be 0.05. The critical value for the test was $t_{0.05,\ 5} = 2.015$

The calculated test statistics, $|\mathrm{t}| = 2.45$, was greater than $t_{0.25,\ 5} = 2.015$. Hence the null hypothesis could be rejected with the significance level 0.05, and *we can state that the change in VAS scores from baseline to follow up 1 is higher than 0*.

Power calculation

The r package *pwr* was used to calculate the power of the t-test, and to give a rough estimate of the sample size needed to reach power of 0.8.

The power of the performed t-test was calculated to be 0.5. Usually, a power above 0.8 is considered to be necessary to correctly reject H_0 given that the true condition is that H_0 is false.

Under similar circumstances, number of participants in the study group should have been increased from 6 to 13, to reach the necessary power of 0.8. Then the probability of doing a type I error would have been reduced from 0.50 to 0.20.

6.9.5 Testing hypothesis about Change in Health problems by Number of treatments

A simple linear regression model was fitted to predict the change in Health problems by Number of treatments.

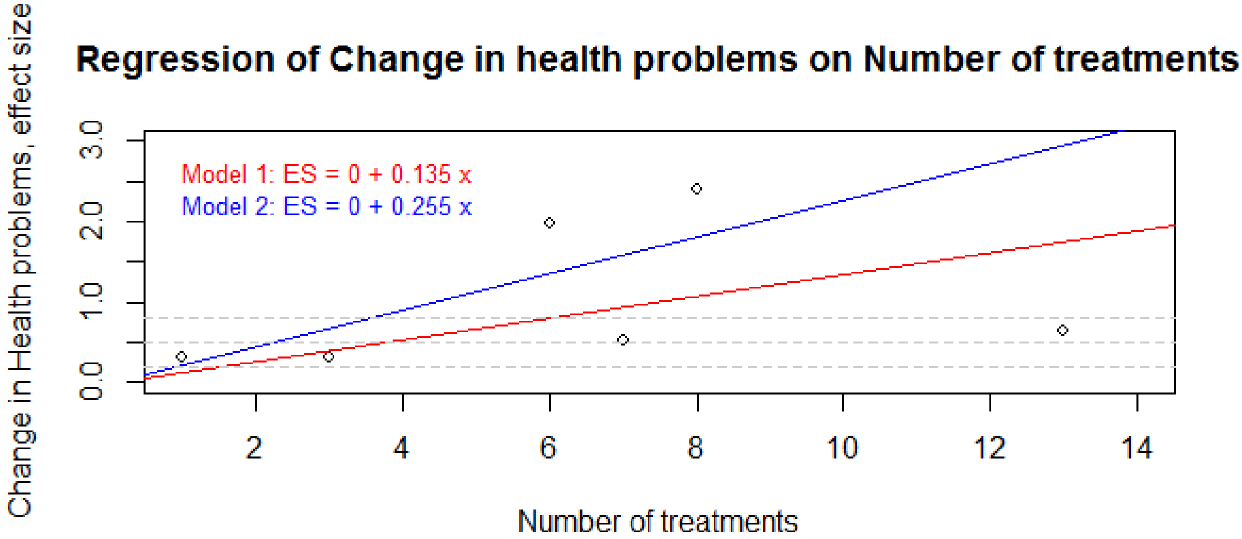


Figure 28. Relationship between Change in health problems and Number of treatments for the prospective group.

(Multiple R-squared = 0.6 for Model 1 and 0.8 for Model 2) Effect size: Higher is better.

Effect size 0.00-0.19 = No change, 0.20-0.49 = Small effect size, 0.50-0.79 = Moderate, 0.80 and above = Large..

The linear equation for the model was

$$Y = 0 + \beta_1 x_1$$

Where x_1 is the variable $Number\ of\ treatments$

and Y is the variable $Change\ in\ health\ problems$ measured as effect size after the formula

$$ES = \frac{Change\ in\ VAS\ scores}{SD\ (baseline\ VAS\ scores)}$$

The regression model was fitted with the entire dataset, and without the single detected outlier. The coefficient β_1, was 0.135 (Standard error = 0.053, df = 5) for the model fitted with the entire dataset and 0.226 (Standard = error 0.056, df = 4) for the model fitted without the outlier.

A one sided t-test was performed to test if the change in Health Problems by Number of treatments was greater than 0.

$$H_0\text{: } The\ change\ in\ Health\ problems\ by\ Number\ of\ treatments = 0$$

$$\beta_1 = 0$$

$$H_A\text{: } The\ change\ in\ Health\ problems\ by\ Number\ of\ treatments > 0$$

$$\beta_1 > 0$$

The test statistics was calculated by the formula:

$$t = \frac{\beta_1}{SE(\beta_1)}$$

The significance level for the test was chosen to be 0.05.

Model 1: critical value $t_{0.05,\ 5} = 2.015$, test statistics $t = \frac{0.135}{0.053} = 2.547$.

Model 2: critical value $t_{0.05,\ 4} = 2.132$, test statistics $t = \frac{0.226}{0.0556} = 4.065$.

The calculated test statistics was greater than the critical value for both models. Hence the null hypothesis could be rejected with the significance level 0.05, and *we can state that the change in VAS scores by number of treatments is higher than 0.*

6.9.6 Prediction of number of treatments necessary to reduce Health problems measured on a VAS scale

The health problems were measured on a 10 point VAS scale.

The estimated standard deviation was 1.88 for the baseline VAS scores (Table 25). A reduction in VAS scores by 1.5 or more corresponds to a large effect size ($\frac{1.5}{1.88} = 0.8$).

The necessary number of treatments to reach a large effect size could then be estimated by the models from the previous section:

$$Model\ 1\text{: } 0.8 = 0 + 0.135\, x_1$$

$$Model\ 2{:}\ 0.8 = 0 +\ 0.226\ x_1$$
$$x_1 = \frac{0.8}{\beta_1}$$

By Model 1, *6 treatments would be necessary to reduce the Health problems measured on a VAS scale by 1.5 points.*

By the more optimistic Model 2, *4 treatments would be necessary to reduce the Health problems measured on a VAS scale by 1.5 poins.*

6.9.7 Testing hypothesis about Change in HRQoL by Number of treatments

A simple linear regression model was fitted to predict the change in HRQoL by Number of treatments.

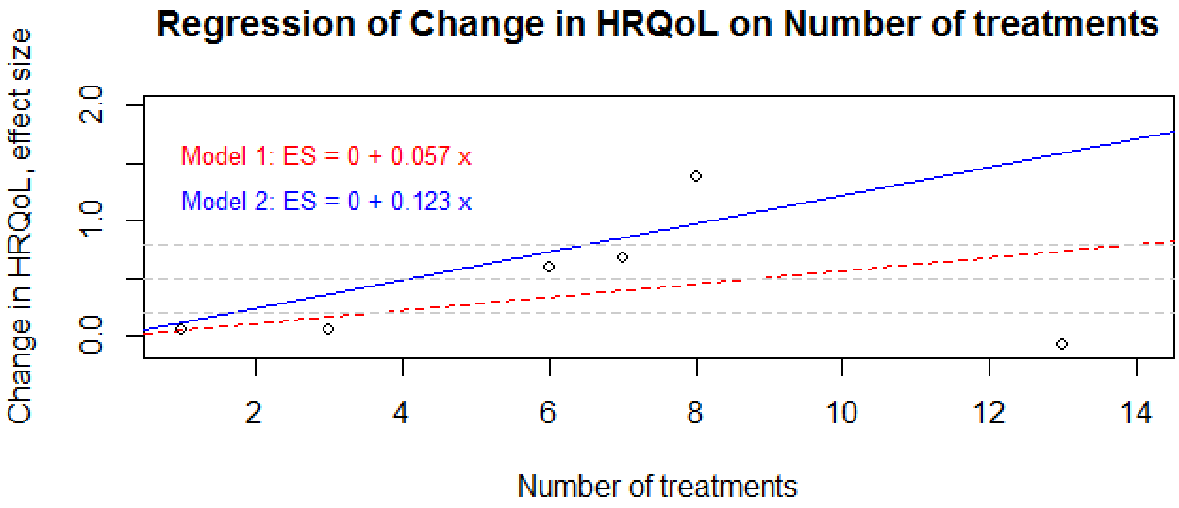


Figure 29. Relationship between Change in HRQoL and Number of treatments for the prospective group.

(Multiple R-squared = 0.4 for Model 1 and 0.9 for Model 2) Effect size: higher is better.

Effect size 0.00-0.19 = No change, 0.20-0.49 = Small effect size, 0.50-0.79 = Moderate, 0.80 and above = Large.

The linear equation for the model was

$$Y = 0 + \ \beta_1 \, x_1$$

where x_1 is the variable $Number\ of\ treatments$

and Y is the variable $Change\ in\ HRQoL$ measured as effect size after the formula

$$ES = \frac{Change\ in\ HRQoL\ z\ scores}{SD\ (baseline\ HRQoL\ z\ scores)}$$

The regression model was fitted with the entire dataset, and without the single detected outlier. The coefficient β_1, was 0.057 (standard error = 0.032, df = 5) for the model fitted with the entire dataset and 0.123 (standard error = 0.022, df = 4) for the model fitted without the outlier.

A one sided t-test was performed to test if the change in Health Problems by Number of treatments was greater than 0.

$$H_0{:}\, The\ change\ in\ HRQoL\ by\ Number\ of\ treatments = \ 0$$

$$\beta_1 = 0$$

$$H_A{:}\ The\ change\ in\ HHRQoL\ by\ Number\ of\ treatments > \ 0$$

$$\beta_1 > \ 0$$

The test statistics was calculated by the formula:

$$t = \frac{\beta_1}{SE(\,\beta_1)}$$

The significance level for the test was chosen to be 0.05.

Model 1: critical value $t_{0.05,\ 5} = 2.015$, test statistics $t = \frac{0.057}{0.032} = 1.781.$

Model 2: critical value $t_{0.05,\ 4} = 2.132$, test statistics $t = \frac{0.123}{0.022} = 5.591.$

The calculated test statistics was greater than the critical value for Model 2.

Hence the null hypothesis could be rejected with the significance level 0.05 for model 2, and *we can state that the change in HRQoL by number of treatments is greater than 0.*

6.9.8 Prediction of number of treatments necessary to improve HRQoL

The standard deviation was 0.80 for the baseline HRQoL z-scores. A change in HRQoL z-scores by 0.6 (which equals to a change in 6 T-scores) or more corresponds to a large effect size ($\frac{0.6}{0.8} = 0.8$).

The necessary number of treatments to reach a large effect size could then be estimated by Model 2 from the previous section:

$$Model\ 2: 0.8 = 0 + \ 0.123\ x_1$$
$$x_1 = \frac{0.8}{0.123} = 6.5$$

By Model 2*, 7 treatments would be necessary to increase HRQoL by 6 T-scores.*

By using the same calculations for moderate and small effect sizes,

4 treatments would be necessary to increase HRQoL by 4 T-scores (ES 0.5)

and 2 treatments would be necessary to increase HRQoL by 2 T-scores (ES 0.2) .

6.10 Mixed model analysis

6.10.1 Model selection

The intention was to fit a mixed model in R which accounted for the dependencies within the dataset. As we got few participants in the study group, it was not possible to estimate all the wanted variance components by the observed values. For practical reasons, SPSS was chosen rather than R for the analysis.

The SF-36 response could be fitted as a two level variable with the summary categories Physical Component Summary and Mental Component Summary instead of the

8 levels used for the retrospective group. The random effect of Person was kept in the model. This made it possible to make a model with random intercept and random slope by Person.

The time aspect was added by the variable Schedule, which grouped the submitted forms at 0, 3, 6 and 9 months after start of treatment. Two covariates of interest were added to the model: Lack of energy and Number of treatments. There were not enough observations to estimate possible interaction effects between the variables.

6.10.2 Observed SF-36 summary scores

Observed PCS and MCS scores are visualized in Figure 30.

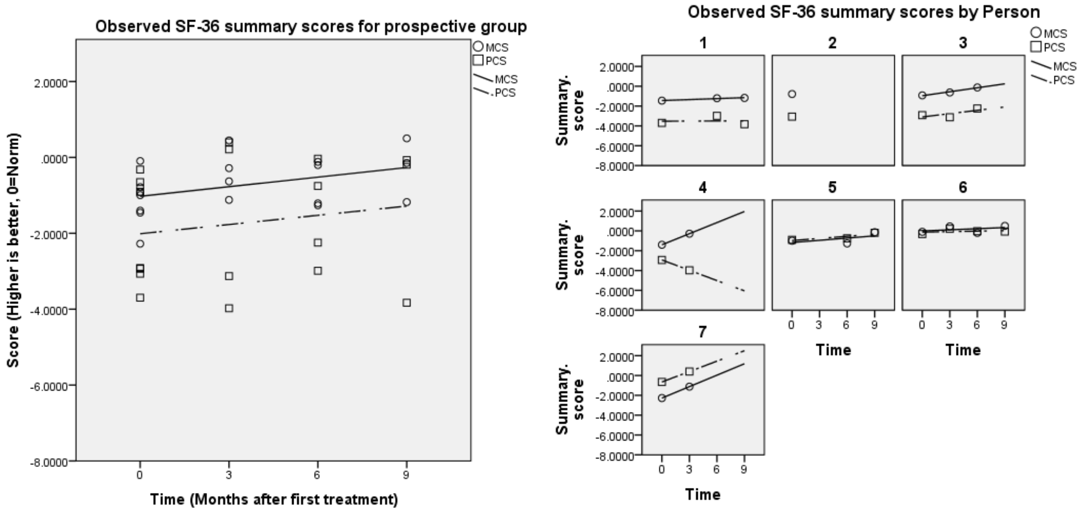


Figure 30. Observed SF-36 summary scores for the prospective group.

PCS = Physical Component Summary, MCS = Mental Component Summary.

6.10.3 SPSS model

The model was fitted using the following syntax:

```
MIXED
   Summary.score
  BY  Scedule Summary.category
  WITH Lack.of.energy Number.of.treatments
/METHOD = ML
```

```
/FIXED = INTERCEPT  Lack.of.energy  Number.of.treatments Scedule Summary.category |
SSTYPE(3)
/RANDOM = INTERCEPT  Scedule| SUBJECT(Person) COVTYPE(ID)
/REPEATED=  Scedule| SUBJECT(Person Summary.category) COVTYPE(DIAG)
/EMMEANS=TABLES( Scedule*Summary.category) WITH (Lack.of.energy=MEAN).
```

The fitted values by Person and Time are visualized in Figure 31. The regression lines have different intercepts and different slopes for different persons.

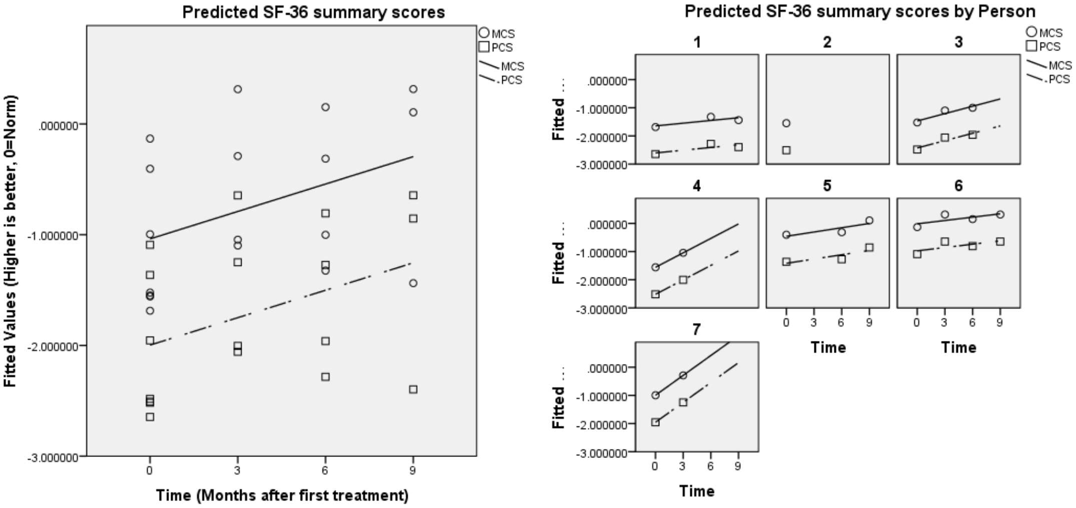


Figure 31. Predicted PCS and MCS z-scores by Person and Time for the prospective group.

PCS = Physical Component Summary, MCS = Mental Component Summary.

6.10.4 Prediction with random effects set to 0

The SPSS model can be used to predict the outcome for future observations. If the random effects are set to zero, and the fixed effects set to mean values, the scores can be estimated by:

$$Score = Main\ intercept + \beta_1 Lack\ of\ energy + \beta_2 Number\ of\ treatments + \beta_3 Scedule + \beta_4 Health\ summary\ category$$

Variable to estimate	Time	Main intercept	$\beta_1 Lack\ of\ energy$	$\beta_3 Number\ of\ treatments$	$\beta_3 Scedule$	$\beta_4 Health\ summary\ category$	Estimated score
PCS	Baseline	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0	0	-2,2
	Follow up 1	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0.5	0	-1,7
	Follow up 2	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0.3	0	-1,9
	Follow up 3	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0.4	0	-1,8
MCS	Baseline	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0	1.0	-1,2
	Follow up 1	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0.5	1.0	-0,7
	Follow up 2	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0.3	1.0	-0,9
	Follow up 3	-2.3	$-0.1 * 4.833$	$0.1 * 6.17$	0.4	1.0	-0,8

6.10.5 Prediction with random effects set to average value

The marginal model was used to estimate group means at the different scheduled times with random effects set to the average value. The predicted values are listed in Table 27 together with observed values and values predicted with the random effects set to 0.

Table 27. Predicted values of PCS and MCS z-scores by the marginal model, together with observed scores and predicted scores by the fixed effect model.

Type	Variable	Time	Group mean in z-scores	Group mean in T-scores
Observed	MCS	0	-1,1	38,68
		3	-0,4	46,04
		6	-0,7	43,02
		9	-0,3	47,27
	PCS	0	-2,1	29,30
		3	-1,6	33,78
		6	-1,5	34,95
		9	-1,4	36,36
Predicted by random effect terms set to 0	MCS	0	-1,2	38,34
		3	-0,7	43,34
		6	-0,9	41,34
		9	-0,8	42,34
	PCS	0	-2,2	28,34
		3	-1,7	33,34
		6	-1,9	31,34
		9	-1,8	32,34
Predicted by marginal model	MCS	0	-1,0	39,77
		3	-0,5	44,96
		6	-0,7	42,65
		9	-0,6	43,69
	PCS	0	-2,0	30,18
		3	-1,5	35,36
		6	-1,7	33,06
		9	-1,6	34,10

Covariates appearing in the model are evaluated at the following values: Lack.of.energy = 4,833,Number.of.treatments = 6,17.

The T-scores from Table 27 are plotted in Figure 32. The difference in predicted T-scores is in the range from 5 to 5.2 between baseline scores and first follow up.

The mean of the observed values are dropping at follow up 2, and increasing again at follow up 3. By inspecting the plot of the observed values in Figure 30 on page 76, the drop at follow up 2 might indicate that Person 4 and Person 7 (who left the study after follow up 1) influenced highly the result at follow up 1. The rest of the group showed a slower improvement, which is reflected in the new rise to the last follow up.

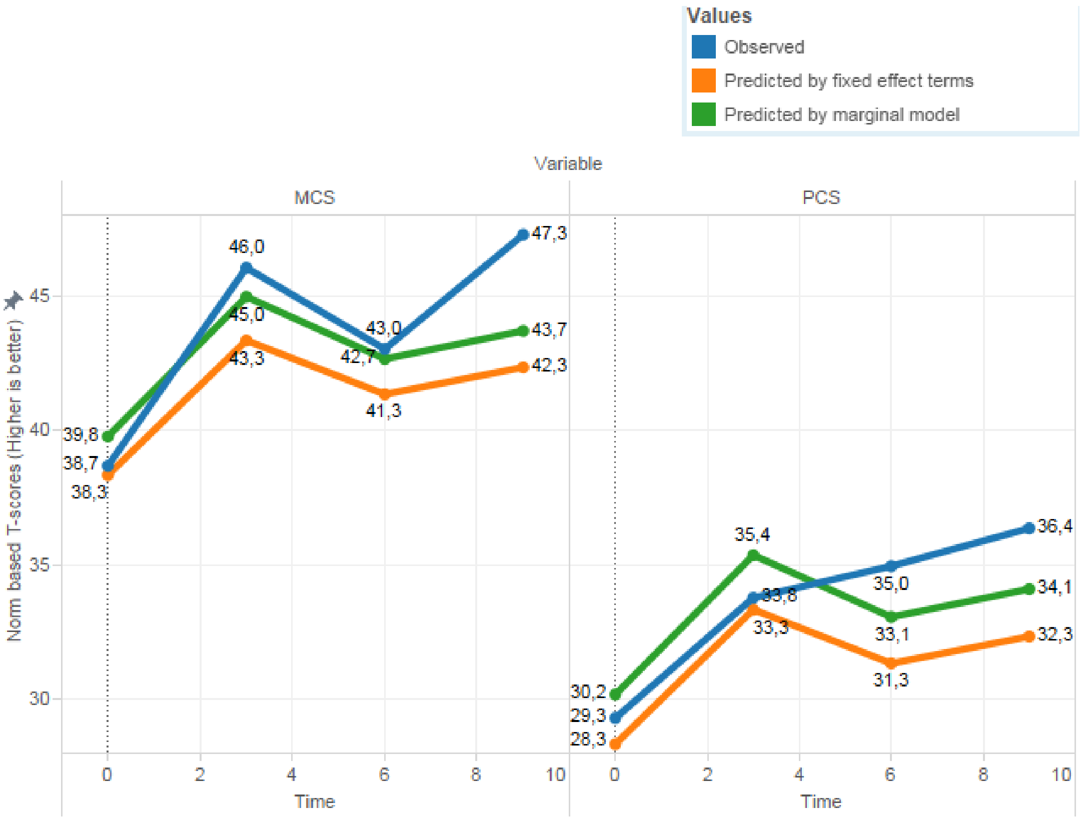


Figure 32. Visualization of (1) Observed values together with (2) values predicted with Random effects set to 0 and (3) values predicted with Random effects set to average.

Chapter 7: Discussion

The purpose of this study was to document the treatment effect of TCM in various health problems. The treatment effect was defined as the change in health status from before start of treatment and the first follow up 3 months later for the prospective group. A significant treatment effect of number of treatments was found both when health was measured by HRQoL and by VAS.

7.1 Regression to the mean

The patients in the prospective group were treated for health problems that had lasted for a long time. The health situation for chronically ill patients will fluctuate over time, and usually they will seek help when the situation is bad. The phenomena that the group scores will improve from the time they seek help, regardless if they get treatment or not (regression to the mean), is usually taken care of by comparison of the treatment group against a control group.

This study was implemented without a control group, and the improvement could therefore possibly be considered as a result of natural fluctuation over time. A small correlation between the response and the regressor would support the argument that most of the effect was due to the regression of the mean phenomena. The large correlation shown in 6.9.7, Testing hypothesis about Change in HRQoL by Number of treatments, supports the argument that most of the change was due to a real treatment effect. However, the study should be repeated with a control group to estimate how big impact this phenomena has on the measured effect.

On the other hand, many of the patients in the study group have tried many different treatments offered by the official health care system before they tried TCM treatment, and the improvement for those cannot so easily be explained by natural fluctuation.

7.2 Biased sample

82% of the participants in the retrospective group and all participants in the prospective group were treated for health problems that had lasted for more than 3 months. The average duration of health problems before start of treatment were 6 years and 12 years for the two groups. Thus, we have reason to believe that the study group was not a random sample (of patient with similar conditions) from the general Norwegian population.

7.3 Comparison against the general Norwegian population

If the patients groups had been random samples of the general Norwegian population, we could expect that the health measured by SF-36 was equal to the mean for the general Norwegian population after finished treatment. For the retrospective group, we found that HRQoL was 9 scores below the norm one year after finished treatment. For the prospective group, we found that HRQoL was 34 scores below the norm before start of treatment, and 38.1 below the norm 3 months later. Although HRQoL still was low, an improvement of 4.1 is a good improvement for the present group of persons with long lasting health problems.

7.4 Clinically Meaningful Differences in HRQoL

The change in HRQoL that is clinically meaningful is different in different contexts. Often a change in 3 to 5 scores is regarded as the minimum change that is worth considering [46, 47]. For the prospective group, an improvement of 3.7 was found for Physical Component Summary, and an improvement of 4.6 for Mental Component Summary. This indicates that the patients have experienced a meaningful improvement of health.

7.5 Reliability and Validity

It can be expected that factors other than the actual treatment affect health-related quality of life over time. A few additional questions were added to the follow-up questionnaire to detect changes in medication, hospitalization, injuries or other health changes that have arisen. It would then be possible to exclude survey forms where the results were highly influenced by other factors than the actual treatment situation. We did not find cases where the findings were highly influenced by information from the additional questions, and all the submitted forms were used in the study.

For the prospective group, the changes in health status could be captured both by the VAS variables and the SF-36 variables. This indicates that both VAS and SF-36 were reliable measures for health changes for the present group. As discussed previously, we have reason to believe that the study group was not a random sample of the general Norwegian population, and the results are therefore not necessarily valid for other groups.

7.6 Missing answers and missing survey forms

In the present study, only one answer was missing in one survey form. When several answers are missing, the algorithms used to substitute the missing values can influence the result, but this was not an issue in this study.

There were few persons in the prospective group. 7 persons submitted the survey form before start of treatment. 6 persons submitted the first follow up form, 4 persons submitted the third follow up form, and 2 persons submitted the fourth follow up form. The survey forms were submitted at various times, and none of the persons submitted follow up forms as often as intended.

Because of few survey forms in the third and fourth follow up, most of the analyses were performed based on the change from baseline to follow up 1. The random effect model can handle unbalanced design, and the model was fitted with all the submitted forms. However, because there were too few observations to estimate a full model, the analysis was performed on a reduced model.

We assumed that the missing survey forms were missed by random. The only pattern we could consider, was that the patient answered the survey forms as long as they were in treatment. A short period of time after they had finished the treatment, they did not return more follow up forms. Hence we could not estimate how long the treatment effect persisted after the treatment was finished.

7.7 HRQoL as the response variable

Results for the SF-36 survey is usually reported as 8 health categories and 2 summary components. The overall mean of the 8 health categories, or the mean of the 2 summary components, is of interest as a reference value, even when the mean of mental and physical health have no practical interpretation.

For the retrospective group, we constructed the summary components first as an average of 4 and 4 health categories, and HRQoL as an average of the summary components. Later, the 8 health categories were given different weights after the principal component analysis. The way the summary scores are constructed will of course make a difference on the result.

When the outcome from different studies shall be compared, the construction method of the summary scores are of interest. Therefore, in some studies, the summary scores are reconstructed from the 8 health categories before the results are compared [47].

7.8 Health problems

The study group was a heterogeneous group. Most of the patients were treated for multiple health problems each, and together the 48 patients were treated for more than 100 health problems. A few of the health problems were everyday ailments such as runny nose and other common infections. Most of the health problems were long lasting problems such as obesity, sleep problems, skin problems, back pains, hearing impairment, anxiety and depression. Many of the health problems were even more serious diseases, such as malignant diseases, rheumatologic and neurological diseases. We did not find it possible to identify homogenous subgroups after health problems, and the results were therefore not compared against other studies.

7.9 Limitations and strengths of the chosen study design

The present study was designed as a small pilot study with no control group. As mentioned before, a control group is necessary to avoid effects like regression to the mean. A placebo group with a shame treatment would also be necessary to estimate the system effect of "being in a treatment". The measured effect can therefore be said to be the sum of the true treatment effect, the placebo effect and the effect of regression to the mean. If the true component effect of the treatment should be revealed, a sufficient large homogenous group should be chosen, and the participants should be randomly assigned to the treatment or the control group.

On the other hand, the strength of the present study is that the effect is measured as the effect experienced by the patients in a real treatment situation.

Chapter 8: CONCLUSION

A significant effect of the TCM treatment was found in the study. The effect of the treatment was found to increase by the number of treatments received. The effect of the treatment was captured both by the VAS and the SF-36 instruments.

The interpretation of the findings is limited by the study design. Nevertheless, the results presented in this study can be valuable information for further research.

The mixed effect model was found to be very useful for the study, and can be used quite generally. The Principal component regression also proved to be useful.

LIST OF TABLES

Table 1. Construction of SF-36 Health categories and Summary categories. 24
Table 2. Normative data for the general Norwegian population. Each cell gives mean, standard deviation and number of persons for the health category by gender and age group. 25
Table 3. Examples of transformation of raw SF-36 scores to Norm based scores. 26
Table 4. Weights given to each SF-36 category. 28
Table 5. The response variables in a table in wide format.. 32
Table 6. The response variables were rearranged to a multi-level variable in a table in long format... 32
Table 7. Number of participants by gender for the retrospective group. 37
Table 8. Number of participants by age group for the retrospective group. 37
Table 9. Explanatory variables for the retrospective group. 39
Table 10. Self-rating of health for the retrospective group. 40
Table 11. Mean VAS scores for the retrospective group. 42
Table 12. SF-36 Norm based scores for the retrospective group. 43
Table 13. Summarized data for the general Norwegian population and the retrospective group. 45
Table 14. The importance of the Principal Components for the retrospective group. 51
Table 15. Component loadings for the retrospective group. 53
Table 16. Predicted norm based SF-36 z-scores based on the PCA model. 55
Table 17. Predicted values for the SF-36 health categories, comparing lm and lmer methods for the retrospective group. 57
Table 18. Predicted norm based SF-36 z-scores based on the Mixed model. 58
Table 19. Number of participants by gender for the prospective group. 59
Table 20. Number of participants by age group for the prospective group. 59
Table 21. Explanatory variables for the retrospective group. 61
Table 22. Self-rating of health before start of treatment for the prospective group. 61
Table 23. Self-rating of health at first follow-up (3 months after first treatment) for the prospective group. 62
Table 24. Health transition after 3 months for the prospective group. 62
Table 25. VAS scores at two occasions for the prospective group. 66
Table 26. SF-36 scores at two occasions for the prospective group. 67
Table 27. Predicted values of PCS and MCS z-scores by the marginal model, together with observed scores and predicted scores by the fixed effect model. 79

LIST OF FIGURES

Figure 1. Approximately historical timeline of Medicine. 7
Figure 2. Illustration of how statistical analysis can be used as a tool to reduce costs and improve quality. 10
Figure 3 The different phases of drug development [22]. 14
Figure 4. Development of some HRQoL instruments. 17
Figure 5. VAS and HRQoL measurements.. 18
Figure 6. Visualization of SF-36 raw scores and SF-36 Norm based T-scores................ 27
Figure 7. Correlation between the response variables for the retrospective group........... 38
Figure 8. Self-rating of health, Retrospective group against general Norwegian population and CAM users in Norway. 40
Figure 9. Visualization of Self-rated health transition for the retrospective group. 41
Figure 10. Mean VAS scores for the retrospective group. 42
Figure 11. Visualization of SF-36 Norm based scores for the retrospective group. Higher is better, Norm = 50. 44
Figure 12. Values for significance level and power for the test illustrated for the retrospective group.. 47
Figure 13. Power curve for the hypothesis test for the retrospective group. 47
Figure 14. Correlation matrix for SF-36 and VAS scores for the retrospective group..... 49
Figure 15. Correlation matrix for SF-36 summary scores and VAS scores for the retrospective group.. 49
Figure 16. Relationship between SF-36 HRQoL z-scores and Lack of energy for the retrospective group.. 50
Figure 17. Relationship between SF-36 BP z-scores and Pain for the retrospective group. 50
Figure 18. Variance explained by the PCA components for the retrospective group....... 51
Figure 19. Biplot of the first two components for the retrospective group....................... 52
Figure 20. Predicted random effects of SF-36 Health categories and Person for the retrospective group.. 56
Figure 21. Comparing the PCR and the Mixed models for the retrospective group. 58
Figure 22. Visualizing of Health-transition for the prospective group............................ 62
Figure 23. Visualization of change in VAS scores by Person for the prospective group. z-scores: higher is better. 63
Figure 24. Visualization of change in SF-36 scores by Person for the prospective group. z-scores: higher is better. 64
Figure 25. Visualization of change in SF-36 summary scores by Person for the prospective group. z-scores: higher is better.. 65
Figure 26. Visualization of effect size for SF-36 scores, SF-36 summary scores and VAS scores for the prospective group. Effect size: higher is better. 68
Figure 27. Visualization of change T-scores for SF-36 Norm based scores for the prospective group. T-scores: higher is better... 69

Figure 28. Relationship between Change in health problems and Number of treatments for the prospective group. 71
Figure 29. Relationship between Change in HRQoL and Number of treatments for the prospective group. 73
Figure 30. Observed SF-36 summary scores for the prospective group. 76
Figure 31. Predicted PCS and MCS z-scores by Person and Time for the prospective group. 77
Figure 32. Visualization of (1) Observed values together with (2) values predicted with Random effects set to 0 and (3) values predicted with Random effects set to average. ... 80

REFERENCES

1. World Health Organization. *WHO definition of Traditional Medicine*. 2015 [cited 2015 25 March]; Available from: http://www.who.int/medicines/areas/traditional/definitions/en/.
2. White, A. and E. Ernst, *A brief history of acupuncture.* Rheumatology, 2004. 43(5): p. 662-663.
3. Taylor, K., *Chinese medicine in early communist China, 1945-63: a medicine of revolution*. 2005: Psychology Press.
4. World Health Organization, *WHO Traditional Medicine Strategy 2014-2023*. 2014.
5. Lu, G.-D., J. Needham, and V. Lo, *Celestial lancets: a history and rationale of acupuncture and moxa*. 2002: Psychology Press.
6. Reston, J. *Now, Let Me Tell You About My Appendectomy in Peking*. 1971 [cited 2015 19 July]; Available from: http://www.acupuncture.com/testimonials/restonexp.htm.
7. U.S. Food and Drug Administration. *Food and Drug Administration Modernization Act (FDAMA) of 1997*. [cited 2015 July 19]; Available from: http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/SignificantAmendmentstotheFDCAct/FDAMA/default.htm.
8. World Health Organization, *Acupuncture : review and analysis of reports on controlled clinical trials - See more at: http://apps.who.int/iris/handle/10665/42414#sthash.dySL1Pkm.dpuf*. 2002.
9. Aarbekke, J., et al., *Alternative Medicine*. NOU 1998:21.
10. Wong, W., et al., *Effectiveness of Traditional Chinese Medicine in Primary Care, Recent Advances in Theories and Practice of Chinese Medicine*, P.H. Kuang, Editor. 2012, InTech.
11. Technical University Munich *Traditionelle Chinesische Medizin (TCM) Studium*. 2015 [cited 2015 July 17]; Available from: http://portal.mytum.de/studium/studiengaenge_en/trad_chin_medizin_master?ignore_redirection=yes.

12. National Information Centre for alternative therapies (NIFAB). *Figures and facts*. 2015 [cited 2015 25 March]; Available from: http://nifab.no/hva_er_alternativ_behandling/tall_og_fakta.

13. Johansdatter Salomonsen L, Grimsgaard S, and Fønnebø V, *Use of alternative medical treatment in Norwegian hospitals (Bruk av alternativmedisinsk behandling ved norske sykehus).* Tidsskrift for den Norske Laegeforening, 2003.

14. Wikipedia. *History of Medicine*. 2015 [cited 2015 25 March]; Available from: http://en.wikipedia.org/wiki/History_of_medicine.

15. Lipton, B.H., *The biology of belief.* 2008: Hay House, Inc.

16. World Health Organization, *General Guidelines for Methodologies on Research and Evaluation of Traditional Medicine.* 2000.

17. UiT The Arctic University of Norway. *The National Research Center in Complementary and Alternative Medicine* 2015 [cited 2015 25 March]; Available from: http://en.uit.no/om/enhet/forsiden?p_dimension_id=88112.

18. World Health Organization, *WHO Traditional Medicine Strategy 2002-2005*. 2002.

19. Cleary-Guida, M.B., et al., *A regional survey of health insurance coverage for complementary and alternative medicine: current status and future ramifications.* The Journal of Alternative & Complementary Medicine, 2001. 7(3): p. 269-273.

20. Moore, W., *Past Caring: So many, so wrong.* BMJ: British Medical Journal, 2007. 334(7588): p. 318.

21. Norheim, A.J., *Acupuncture in health care - Attitudes to, and experiences with acupuncture in Norway*. 2005, Tromsø.

22. Altman, D.G., *Practical statistics for medical research*. 1990: CRC press.

23. Tang, J.-L., *Research priorities in traditional Chinese medicine.* BMJ: British Medical Journal, 2006. 333(7564): p. 391.

24. Kaptchuk, T.J., *Acupuncture: Theory, Efficacy, and Practice.* Annals of Internal Medicine, 2002. 136(5): p. 374-383.

25. World Health Organization, *Guidelines on Basic Training and Safety in Acupuncture*. 1999.

26. Chen, T., L. Li, and M.M. Kochen, *A systematic review: How to choose appropriate health-related quality of life (HRQOL) measures*

in routine general practice? Journal of Zhejiang University SCIENCE, 2005.

27. Huang, I., A. Wu, and C. Frangakis, *Do the SF-36 and WHOQOL-BREF measure the same constructs? Evidence from the Taiwan population.* Qual Life Res. , 2006. Feb.
28. Leung, K.-f., et al., *Development and validation of the Chinese Quality of Life Instrument.* Health and Quality of Life Outcomes, 2005: p. 3:26.
29. Aschero, G., et al., *The ChQoL questionnaire: an Italian translation with preliminary psychometric results for female oncological patients.* Health and Quality of Life Outcomes, 2010. 10: p. 8:106.
30. Ware, J.E. and B. Gandek, *Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project.* Journal of clinical epidemiology, 1998. 51(11): p. 903-912.
31. Fayers, P. and D. Machin, *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. 2013: John Wiley & Sons.
32. RAND Corporation. *36-Item Short Form Survey Scoring Instructions* 2014 [cited 2015 25 March]; Available from: http://www.rand.org/health/surveys_tools/mos/mos_core_36item_scoring.html.
33. Loge, J.H. and S. Kaasa, *Short form 36 (SF-36) health survey: normative data from the general Norwegian population.* Scandinavian Journal of Public Health, 1998. 26(4): p. 250-258.
34. SF-36v2. *SF-36v2_Manual_Chapter_1*. Available from: https://www.optum.com/content/dam/optum/resources/Manual%20Excerpts/SF-36v2_Manual_Chapter_1.pdf.
35. RAND Health. *Terms and Conditions for Using the 36-Item Short Form Health Survey*. 2015 [cited 2015 25 March]; Available from: http://www.rand.org/health/surveys_tools/mos/mos_core_36item_terms.html.
36. McDowell, I., *Measuring Health: A Guide to Rating Scales and Questionnaires*. 2006: Oxford University Press.
37. Hays, R.D., C.D. Sherbourne, and R. Mazel, *User's manual for the Medical Outcomes Study (MOS) core measures of health-related quality of life*. 1995: Rand Corporation.

38. Farivar, S.S., W.E. Cunningham, and R.D. Hays, *Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V.1.* Health and Quality of Life Outcomes, 2007. 5: p. 54.
39. Update, S.-H.S. *Norm-based Scoring and Interpretation*. 2007; Available from: http://www.sf-36.org/tools/sf36.shtml.
40. Ellert, U. and B. Kurth, *[Methodological views on the SF-36 summary scores based on the adult German population].* Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz, 2004. 47(11): p. 1027-1032.
41. Jenkinson, C., *Comparison of UK and US methods for weighting and scoring the SF-36 summary measures.* Journal of Public Health, 1999. 21(4): p. 372-376.
42. Kelley, K., *Methods for the Behavioral, Educational, and Social Sciences (MBESS)[Computer software and manual]. Retrievable from www. cran. r-project. org.* 2007.
43. Cohen, J., *Statistical power analysis for the behavioral sciences (rev*. 1977: Lawrence Erlbaum Associates, Inc.
44. Fitzmaurice, G.M., N.M. Laird, and J.H. Ware, *Applied longitudinal analysis*. Vol. 998. 2012: John Wiley & Sons.
45. Research Council of Norway. *People is no longer responding*. 2013 2015-06-09]; Available from: http://www.forskningsradet.no/prognett-bladetforskning/Nyheter/Folk_svarer_ikke_lenger/1253986892964.
46. Hays, R.D. and J.M. Woolley, *The concept of clinically meaningful difference in health-related quality-of-life research.* Pharmacoeconomics, 2000. 18(5): p. 419-423.
47. LINCOLN, R., *The SF-36 Health survey: A summary of responsiveness to clinical interventions*. 2000.

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no