

Out of office: A model-based cost-effectiveness analysis of return-to-work interventions in Norway

Master Thesis
European Master in Health Economics and Management

Student: Niccolò Morgante

Student number: 658141

Project supervisors: Emily Burger, Gudrun Waaler Bjørnelv, Natalia Kunst



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Erasmus School of
Health Policy
& Management




UNIVERSITETET
I OSLO

Declaration in lieu of oath

I hereby declare, under oath, that this master thesis has been my independent work and has not been aided with any prohibited means. I declare, to the best of my knowledge and belief, that all passages taken from published and unpublished sources or documents have been reproduced whether as original, slightly changed or in thought, have been mentioned as such at the corresponding places of the thesis, by citation, where the extent of the original quotes is indicated.

The paper has not been submitted for evaluation to another examination authority or has been published in this form or another.

25/05/2023	
Date	Signature

Abstract

Purpose: The sickness absence rate in Norway is at its highest point since 2009 and the decision on the reimbursement of return-to-work (RTW) programs requires a thorough analysis. This study aimed to assess the long-term cost-effectiveness of two RTW interventions for patients with musculoskeletal and psychological disorders. The interventions included in the study were I-MORE (an inpatient multimodal program) and O-ACT (an outpatient physiotherapy program).

Methods: First, we used patient-level data to estimate input parameters such as costs, HRQoL, and transition probabilities. Second, we developed a discrete-time state-transition model to perform a cost-effectiveness analysis using such inputs. Alternative scenarios and sensitivity analyses were used to assess the impact of uncertainties on the model results.

Results: Considering a healthcare perspective, over 25 years, I-MORE was not cost-effective compared to O-ACT (ICER: 1,167,887 NOK/QALY). Out of 10,000 simulations, with a threshold of NOK 500,000, I-MORE was cost-effective in 15% of the cases. Once we considered a limited societal perspective, which accounted for production loss, I-MORE not only became cost-effective but strongly dominated O-ACT. From a limited societal perspective I-MORE became cost-saving after 3 years.

Conclusion: Under current benchmark thresholds for cost-effectiveness, the inpatient nature of I-MORE drove up the costs which outweighed the small increased effects. However, the results of the evaluation were strongly influenced by the perspective of the analysis and the chosen time horizon. Our results emphasize the importance of discussing the role of the societal perspective in economic evaluations for healthcare. To our knowledge this is the first multi-state model developed to assess the cost-effectiveness of RTW in Norway.

Table of contents

1 INTRODUCTION	6
2 BACKGROUND	7
2.1 SICKNESS ABSENCE	7
2.2 RETURN-TO-WORK INTERVENTIONS	8
3 THEORETICAL FRAMEWORK	11
3.1 PRIORITY SETTING & COST-EFFECTIVENESS	11
3.2 DECISION ANALYTIC MODELLING	12
3.3 PARAMETERISATION OF STATE-TRANSITION MODELS	13
3.3.1 <i>Transition probabilities</i>	13
3.3.2 <i>Multi-state modelling</i>	14
3.3.3 <i>Costs and analysis perspective</i>	15
3.4 ADDRESSING UNCERTAINTY	16
3.5 VALUE OF INFORMATION	18
4 METHODS	20
4.1 THE HYSNES TRIAL	20
4.1.1 <i>Patient-level data</i>	22
4.2 MULTI-STATE MODEL	22
4.2.1 <i>Model structure</i>	22
4.2.2 <i>Transition probabilities</i>	23
4.2.3 <i>Alternative 3-state transition model</i>	25
4.3 COST-EFFECTIVENESS ANALYSIS	27
4.3.1 <i>Decision analytic approach</i>	27
4.3.2 <i>Healthcare costs</i>	27
4.3.3 <i>Production loss</i>	28
4.3.4 <i>Health related quality of life</i>	29
4.3.5 <i>Deterministic analysis</i>	30
4.3.6 <i>Probabilistic analysis</i>	30
4.3.7 <i>Scenario analyses</i>	32
4.3.8 <i>Value of information analysis</i>	32
4.4 MODEL VALIDATION	33
4.5 STATEMENT OF ETHICAL APPROVAL	34
5 RESULTS	35
5.1 PARAMETER ESTIMATION	35
5.1.1 <i>Time-homogeneous transition probabilities</i>	35
5.1.2 <i>Time-inhomogeneous transition probabilities</i>	38
5.1.3 <i>3-state transition probabilities</i>	40
5.1.4 <i>Cost and HRQoL estimates</i>	42
5.2 COST-EFFECTIVENESS ANALYSIS	43
5.2.1 <i>Deterministic results</i>	43
5.2.2 <i>Probabilistic analysis</i>	43
5.2.3 <i>Alternative scenarios</i>	45
5.2.4 <i>Value of information</i>	47
6 DISCUSSION AND LIMITATIONS	48
7 CONCLUSION	53
8 REFERENCE LIST	54
APPENDIX A: TABLES & FIGURES	59
APPENDIX B: VALIDATION CHECKLIST	66
APPENDIX C: STATA & R CODE	70

Figures

Figure 1 Model structure with 4 states _____	23
Figure 2 Alternative model structure with 3 states _____	25
Figure 3 States' prevalence over time for O-ACT and I-MORE in base-case model _____	36
Figure 4 Time-homogeneous transition probabilities from Sick Leave state to all states _____	37
Figure 5 States' prevalence over time for O-ACT and I-MORE in piecewise constant model _____	39
Figure 6 Time-inhomogeneous transition probabilities from Sick Leave state to all states _____	40
Figure 7 States' prevalence over time for I-MORE in piecewise constant model _____	41
Figure 8 Transition probabilities from General Benefit state to all states _____	41
Figure 9 Cost-effectiveness plane with a healthcare perspective and WPT of NOK 500,000 _____	44
Figure 10 Cost-effectiveness acceptability curves and frontiers with a healthcare perspective _____	44
Figure 11 Population EVPI and EVPPI for each proposed group of parameters _____	47

Tables

Table 1: Detailed components of the two return-to-work interventions _____	21
Table 2 Implications of a 4-state structure vs a 3-state structure _____	26
Table 3 Input parameters for the probabilistic analysis _____	31
Table 4 Time-homogeneous transition probability matrices _____	35
Table 5 State-costs from a random effects panel regression _____	42
Table 6 State-HRQoL from a random effects panel regression _____	42
Table 7 Cost-effectiveness results _____	46

1 Introduction

Since the end of the 1990s, Norway and other Nordic countries have been facing a major public health issue: high and increasing rates of sickness absence. In line with recommendations from the Organization for Economic Cooperation and Development (OECD), Norway has focused its efforts on developing policies that support and stimulate the return-to-work (RTW) of employees with temporary or permanent reduced work capacity (Kausto et al., 2008). Sickness absence can be attributed to a wide variety of disorders. Although estimates vary, research indicates that the majority of sick leave days is associated with either musculoskeletal disorders (Hagen et al., 2011; Kinge et al., 2015) or psychological disorders (Nystuen et al. 2001). With the current sick leave rate reaching the highest point (7.1%) in the past 14 years, interventions and policies focusing on RTW have become more crucial than ever.

RTW interventions are multidomain programs, developed since the 1970s, that focus on a wide variety of treatments. These include, but are not limited to, psychotherapy (individual and group-based), work-related problem solving, physical exercise, and the development of a RTW plan (Powers et al., 2009; Ammendolia et al., 2009). However, results on the efficacy of these interventions, especially in the long term, are inconsistent (Aasdahl, 2023). As a consequence, it currently remains unclear whether more resources should be allocated to the study and reimbursement of RTW programs. Policy makers face significant uncertainty when it comes to the allocation of the national budget, particularly regarding the reimbursement of different treatment options. To address the issue of resource scarcity, Norway relies on economic evaluations as a key element of priority setting (Norwegian Ministry of Health and Care Services, 2017). Although there is evidence supporting the cost-effectiveness of RTW programs (Caro et al., 2012), most economic evaluations have only been conducted alongside clinical trials. Few studies, none of which in a Norwegian setting, have developed (or are currently developing) a model that studies the effects and costs over a longer time horizon (Squires et al., 2011; Moens, 2022).

The clinical trial NCT01926574, which took place in Norway between 2013 and 2015, compared the effects and costs of two RTW interventions: I-MORE (inpatient multimodal occupational rehabilitation) and O-ACT (outpatient acceptance and commitment therapy). The purpose of our study was to use patient-level data from the trial to develop and optimize a state-transition model that could project both costs and effects over longer time horizons and under different assumptions. To the best of our knowledge, no model-based analysis has compared the cost-effectiveness of two RTW interventions for Norway. Developing such a model not only allowed us to address the specific research question but also provided the opportunity to test different assumptions, adapt the decision to alternative settings, and identify areas that may require further research. The main objective of this study is captured by the research question:

Is I-MORE, compared to O-ACT, a cost-effective intervention in the long term for individuals on sick leave due to musculoskeletal or psychological disorders in Norway?

2 Background

2.1 Sickness absence

According to the World Health Organisation (2021), Musculoskeletal disorders compromise more than 150 conditions affecting bones, muscles, joints, and more generally the locomotor system. This great variety of disorders poses a challenge to both the patient and the social-care system they are a part of. Although estimates for the general population are subject to uncertainty, research indicates that the prevalence of musculoskeletal disorders among Norwegian working adults could be as high as 80% (Hagen et al., 2011). These conditions are associated with chronic pain and other debilitating symptoms that can last for months. More specifically, Svebak et al. (2006) analysed the results of a survey that collected answers from 60,000 Norwegian adults and found that pain associated with chronic musculoskeletal disorders lasted more than three months in 45% (men) and 40% (women) of the cases. The long-term repercussions not only are likely to affect the patient's quality of life but can also have significant economic implications. Musculoskeletal disorders not only limit patients' presence at work, reducing productivity and increasing healthcare costs, but they also strain the economic system as a whole. This is mainly due to the additional costs that the system faces when providing workers with sick leave benefits. The second edition of the International Classification of Primary Care (ICPC2) groups musculoskeletal conditions under the letter "L" in codes from 0 to 99. Out of these, neck (L01) and low back (L02) complaints account for the highest share of diagnoses and the main reason for health care consumption in Norway (Kinge et al., 2015). With regards to productivity and the labour market, low back pain alone accounts for 15% of all sickness absence in the country (Werner & Côté, 2009).

Psychological disorders pose a second, but no less serious, threat to healthcare systems. These conditions are characterized by clinically relevant disturbances in an individual's cognition and behaviour. ICPC2 classifies psychological disorders under the letter "P" with numbers 70 to 99 referring to the more debilitating conditions. Mental disorders have been shown to significantly reduce patients' health-related quality of life (HRQoL) more than other common medical conditions, such as diabetes and cardiac diseases (Spitzer et al., 1995). The magnitude of the reduction in HRQoL is particularly and significant in domains such as perceived general health and social functioning. In Norway, the burden associated with mental disorders is estimated to be around 8% of the total sickness burden. The impact on the country's workforce is also significant, as more than 30% of all refunded sick days have been linked to mental disorders (Nystuen et al. 2001).

Both groups of disorders are of high relevance for Norway, where the sick leave rate has now reached the highest value (7.1%) since 2009 (Statistics Norway, 2023). Indeed, the most common diagnoses for long-term sick leave and disability benefits are those connected to mild/severe disorders that fall under the "L" and "P" categories (Øyeflaten et al., 2012).

In Norway, sick leave compensation is initially paid by the employer (first 16 days) and then by the Labour and Welfare Administration (NAV). Depending on the condition, the time needed to recover, and the way it affects the person's ability to work, Norwegian workers are entitled to different types of medical benefits. Pathways between the different medical benefits are complex, mainly due to the misalignment between the conditions and the bureaucracy behind the assessment/reimbursement process. Here we provide a general explanation of the Norwegian system and its mechanisms. Sick leave (SL) benefits are a temporary income replacement for a person that is unable to work due to an injury or an illness. SL can be graded from 20% up to 100% and usually covers up to one year (52 weeks) of leave. After the first year, workers can either go back to work (for a minimum of 6 months) or apply for work assessment allowance (WAA). This type of medical benefit is intended for workers that, after a thorough assessment, demonstrate that their ability to work has been reduced by at least 50%. Generally, this type of benefits lasts 3 years, with the possibility of extending it by a maximum of 2 years. However, in contrast to SL, which can be up to 100%, WAA only reimburses 66% of the patient's prior income (up to NOK 441,449). Lastly, disability benefits (DB) are benefits available to those that have little to no capacity to work (reduction of at least 50%). This type of benefit is intended for individuals that are unlikely to improve their condition during their lifetime. As for WAA, a thorough medical assessment from NAV will determine to what degree the person is entitled to DB (NAV, 2020).

2.2 Return-to-work interventions

Due to the broad range of conditions included in the musculoskeletal and psychological categories, establishing a standard treatment procedure becomes challenging. In the case of musculoskeletal disorders, the first point of contact is usually the patient's General Practitioner (GP), who assesses the situation and eventually refers to specialist care. Medications, physiotherapy, occupational therapy, and in worse cases surgical interventions are all feasible approaches that, depending on the severity, might be required.

Compared to musculoskeletal conditions, the diagnosis of psychological disorders presents a greater challenge. Research shows that the overlapping nature of symptoms in psychological disorders is the main cause of slower diagnoses. The identification of the condition and of effective treatment options can last months or even years (E. Baca-García et al., 2021). In addition to the long-term repercussions of musculoskeletal and psychological disorders on the individual it should be noted that comorbidity between the two groups of disorders is often high and can lead to important differences in prognosis and health care consumption. For instance, a Norwegian study on 562 sick listed patients with lower back pain, found that the prevalence of psychological disorders was of 38% (current or lifetime) with somatoform disorders accounting for 18% (Reme et al., 2011). These results emphasise the need for interventions that can be used to address both "L" and "P" disorders.

RTW interventions are rehabilitation programs designed to incentivise the patient in returning to their regular activity. These types of interventions, and their effects on sickness absence have been studied since the early 1970's. An early systematic review has found that modified work programs (often included in broader rehabilitation programs) facilitate return to work for both temporarily and permanently disabled workers (Krause et al., 1998). The authors also conclude that such rehabilitation programs have the potential to substantially reduce the costs associated with workers' compensation. Nowadays, the overall objective of these interventions is indeed to reduce sick leave and ease the burden of physical and mental conditions not only from the healthcare system but from the labour sector as well.

The unique medical circumstances that lead to sickness absence, make the design of rehabilitation programs a difficult task. A solution proposed by Ammendolia et al. (2009) involved the use of intervention mapping, a methodology which has been used since the 1990s to develop multidimensional programs. This method was successfully applied to programs aimed at AIDS prevention and smoking cessation. With such an approach the study was able to pinpoint features that aligned with the setting of the study (Ontario). These included, but were not limited to, the involvement of trained professionals to coordinate the process and empowering the patient in making RTW decisions.

Multi-domain rehabilitation programs often include individual, or group based psychological therapies. Indeed, cognitive behavioural therapy (CBT) and mindfulness-based cognitive therapy (MBCT) are recognised as cost-effective approaches that are likely to prevent relapses (National Institute for Health and Care Excellence, 2022). Acceptance and commitment therapy (ACT) stems from the "third wave" of treatments in the field of CBT. Treatments of the third wave are characterized by a focus on contextual change, the development of flexible repertoires, and the emphasis of function over form (Hayes, 2004). The main goal of ACT is to support individuals that are struggling with their internal experiences, such as thoughts, and provide them with the tools needed to implement behavioural changes. ACT proved to be effective in many domains related to mental health, showing effects comparable to the more traditional CBT (Shand et al., 2013). Moreover, ACT repeatedly outperformed the effects of treatment as usual for a variety of mental and physical disorders (Powers et al., 2009).

In addition to psychological therapies, and pharmacological treatments, physical exercise is often linked to improvements in HRQoL. In a systematic review, Cooney et al, (2013) reported that physical exercise was moderately more effective in reducing symptoms of mental disorders (specifically depression) when compared to no treatment or control groups. However, there is a limited number of studies that directly compared the effects of physical exercise with psychological and pharmacological treatments. In these, results showed no significant effects for the exercise groups.

A variety of methods have been developed to stimulate RTW and address the symptoms of musculoskeletal and psychological conditions. For instance, multi-domain interventions proved to be effective rehabilitation programs. The interventions that had better rates of success included components such as cognitive behavioural therapy (CBT/ACT), physical exercise, work-related problem solving, and a service coordination component, in which the patient is assisted in the development of a return-to-work programme (Cullen et al., 2017).

3 Theoretical Framework

3.1 Priority setting & cost-effectiveness

Resources are scarce and investment decisions are possibly unlimited. In 1935, British economist Lionel Robbins defined the field of economics as the study of the “relationship between ends and scarce means” (Robbins, 1935, p.15). Healthcare decisions are no different, and the issue of resource allocation is here to stay (van Delden, 2004). Economic evaluations play a crucial role in decision-making processes related to healthcare policy, practice, and resource allocation. These evaluations provide valuable insight into the costs and benefits associated with various health interventions, aiding decision-makers in prioritizing their reimbursement.

In Norway, healthcare resource allocation has been addressed by the Government in a guiding report (white paper) published in June 2016 and approved by the Parliament in November of that same year (Norwegian Ministry of Health and Care Services, 2017). The report highlights three main criteria to be used in decision-making processes: the benefit criterion, the resource criterion, and the severity criterion. All three criteria resolve into health technology assessment (HTA), a multi-dimensional systematic approach to new health technologies and services.

Economic evaluations contribute to HTA and often rely on a cost-effectiveness analysis (CEA) (Briggs et al., 2006), in which the clinical effects of two (or more) technologies are compared to their resource use. Effects are usually expressed in life years, disability-adjusted life years (DALYs), or quality-adjusted life years (QALYs). QALYs are a combination of quantity and quality of life (Drummond, 2015) and are calculated by multiplying life years (LY) with a HRQoL value usually bound between 0 and 1. Generic questionnaires such as the 15D and the EQ5D, can be used to elicit those values. Specifically, the 15D is a self-reported questionnaire that assesses fifteen domains: mobility, vision, hearing, breathing, sleeping, eating/drinking, speech, elimination, vitality, mental function, discomfort and symptoms, depression, distress, usual activity, and sexual activity. Each domain presents 5 response options ranging from no problems to severe problems. Using a set of preference weights, as expressed by the general Finnish population, an aggregation formula generates the final 15D score. The scores are expressed on a scale from 0 to 1, where 0 corresponds to being dead and 1 to no problems in any dimension (Sintonen, 2001).

The main outcome of a cost-effectiveness analysis, also referred to as cost-utility analysis when QALYs are used, is the incremental cost-effectiveness ratio (ICER). The ICER is the difference in costs between two technologies divided by the difference in effects. Equation 1 shows the general formula for the ICER.

$$ICER = \frac{\Delta C}{\Delta E} = \frac{(Cost_B - Cost_A)}{(QALY_{S_B} - QALY_{S_A})}$$

By evaluating different options based on their relative effectiveness for achieving desired outcomes, we can identify where resources are best spent. However, the decision does not merely rely on the ICER. As mentioned, scarcity is an underlying issue. To be considered cost-effective, the ICER is compared with a pre-specified willingness-to-pay (WTP) threshold value (λ). Conceptually the threshold represents the maximum amount that decision-makers are willing to pay for an additional unit of health benefit. The threshold is typically based on societal values and reflects the opportunity cost of allocating resources to one intervention over another (Simoens, 2012). Due to their nature, thresholds vary between countries, societies, and even by disease severity (Raftery, 2008). In Norway, WTP thresholds range from NOK 275,000 to NOK 825,000 depending on the severity of the condition (Norwegian Ministry of Health and Care Services, 2015), with an often-reported average value of NOK 500,000 (Barra & Rand-Hendriksen, 2016). The decision rule based on the ICER and the threshold can also be expressed in monetary terms. At a chosen WTP threshold the net monetary benefit (NMB) expresses the incremental effects of an intervention on the same scale of incremental costs. If their difference is positive the intervention is cost-effective (Claxton, 1999). The intervention with the highest NMB is the optimal strategy.

$$NMB = (\lambda * \Delta E) - \Delta C$$

[2]

CEAs have been widely used to evaluate RTW interventions and rehabilitation programs (Dewa et al, 2020; Carroll et al, 2009). Yet, methodologies differ, and comparability is low due to different preferences in terms of outcomes and health-related quality of life (HRQoL) instruments. Most of the research conducted so far focuses on CEAs alongside clinical trials, without fully developing a simulation model to study and evaluate the long-term consequences of the interventions (Hoefsmit et al., 2012). Overall, economic evaluations provide decision-makers in healthcare with a systematic approach to evaluate the costs and benefits of different interventions, address the problem of resource scarcity, and allocate resources in a more effective and equitable manner.

3.2 Decision analytic modelling

To be valuable sources of information for policy makers, evaluations must rely on appropriate evidence. In addition, evaluations should be fully transparent on the limitations and uncertainty intrinsic to models that seek to simplify real world phenomena (Briggs et al., 2011). Many approaches have been proposed through the years: state-transition models (Markov models), discrete event simulations, dynamic transmission models, microsimulations, and decision trees are just some of the approaches to combine quantitative information from various sources and make informed decision in healthcare (Caro et al., 2012).

Decision trees are the simplest form of decision analytic models and have been widely used to represent screening programmes and diagnostic tests with relatively short time horizons. In

decision trees the different clinical pathways are visualised with branches (hence the name) and nodes where a decision/event occurs. Although initially simple and transparent, these models can quickly become overly complex as the time horizon increases and the states multiply (state explosion) (Briggs et al, 2011; Petrou & Gray, 2011). State–transition models allow for a flexible sequencing of states with the possibility of including recurring outcomes (Petrou & Gray, 2011).

In discrete time state–transition models (DTSTMs) patients move between mutually exclusive health states over discrete time intervals known as cycles (Drummond, 2015). Cycle length and states will depend on the specific decision problem being addressed. One major limitation to the basic approach to DTSTMs is that the transition between states is governed only by the current health state, this is known as the Markovian assumption.

Once that the right model specification has been identified, different types of data can be used to inform it. Data relying on published literature are usually collected with systematic reviews of studies closely linked to the intervention/disease being modelled. If the topic has been researched for a long time, published literature can provide many insights into the clinical effectiveness, safety, and costs associated with a technology. However, this kind of information might not always be up-to-date or relevant for the patient population being considered in the evaluation. Primary data are generated when specific studies designs (e.g., randomized controlled trials, surveys, etc.) are carried out. Such an approach offers more control on the type of information that will be collected and ensures that the patient population is in line with the population being considered in the evaluation. Nonetheless, studies aimed at collecting primary data are both time and resource consuming. Depending on the context, it might not be feasible to rely on the generation of primary data.

Overall, many approaches to decision analytic modelling and data collection have been proposed and developed through the years. Each one with its number of assumptions and limitations. However, it is good to remember that a model should be as simple as possible but no simpler.

3.3 Parameterisation of state–transition models

3.3.1 Transition probabilities

On top of costs and effects, CEAs often rely on parameters that describe the patterns (clinical and non-clinical) that patients are expected to follow. To do so, studies often report transition rates and probabilities. In economic modelling, rates represent an instantaneous measure that can take values between 0 and infinity (Fleurence & Hollenbeak, 2007). The rate indicates the instantaneous potential occurrence of an event. Specifically, transition rates (or intensities) represent the instantaneous risk of moving from state 1 to state 2. These values can be easily added and subtracted. Differently, probabilities range between 0 and 1 and represent the likelihood of an event happening over a set period of time. When rates are assumed to be

constant over time without competing risks, rates can be converted into probabilities and vice versa (Briggs et al., 2006)

$$p = 1 - \exp(-rt)$$

$$r = -[\ln(1 - p)]/t$$

[3]

Where:

p is the probability

r is the rate

A transition rate matrix also referred to as an intensity matrix, is a set of instantaneous rates that describe possible transitions between pre-specified states (Jackson, 2011).

3.3.2 Multi-state modelling

Multi-state models, also referred to as DTSTMs when the transitions happen at discrete time intervals, are a way to extend survival analysis to incorporate more states and effectively model competing events (Gran et al., 2015). Even when time-to-event data are not available, these models offer flexible structures to fit panel data (Gran et al., 2015; Incerti & Jansen, 2021). As we introduced, DTSTMs are a simplification of often complex clinical patterns in which transition probabilities are associated with each state. However, in their basic forms, these models are subject to one important limitation. That is, future transitions only depend on the current state (Markov assumption) and that sojourn times are exponentially distributed (Jackson, 2011; Briggs et al., 2006). Multi-state models can have one (or more) final state from which patients cannot come back. These states, where the probability of remaining in the state equals 1, are called absorbing states.

msm is an R package that can be used to fit multi-state models when working with panel data. Panel data can be seen as observations/snapshots of a continuous-time process at arbitrary times (Jackson, 2011). To explain how *msm* works, we first assume a time-homogeneous model, in which the transition intensities are not dependent on time. When exact transition times are unknown (panel data), the transition probability matrix is calculated using a transition intensity matrix Q (Cox & Miller, 1977). If the Q matrix remains constant over time, the case for time-homogeneous models, the equations associated with it are solved by the matrix exponential of Q (scaled for time). In such a scenario, it is the exponential of Q that contributes to the likelihood function. When transition times are exactly observed, the intensity matrix Q directly contributes to the likelihood function (Kunst et al., 2020). In *msm*, the constant intensity assumption (exponential assumption) can be relaxed by allowing piecewise constant transition rates. Having a piecewise constant model is one way of partly introducing time dependency in models fitted with *msm*. With this option, *msm* specifies cut-offs at which the Q matrix is expected to change. Nevertheless, the models estimated between the cut-offs remain based on exponential distributions. The fact that *msm* can only fit exponential distributions (or piecewise exponential)

is a highly constricting feature that limits modelling options. This is why, in the case of exactly observed transitions (time-to-event data), more flexible and fully time-inhomogeneous models should be considered. For instance, packages like *mstate* allow to fit continuously changing intensities (e.g, Weibull) (de Wreede et al., 2011).

The likelihood function quantifies how well a particular set of parameters explains the observed data. The likelihood function represents the joint probability density of all observations given the chosen model parameters. In statistics, the parameters of a model are often estimated with the maximum likelihood estimation (MLE) method. MLE is the basis of many well-known inference methods, such as the chi-square test, models of random effects, and selection criteria such as the Akaike information criterion (Myung, 2003). The MLE method determines parameters' values such that they maximize the likelihood that the process described by the model matches the observed data. In practice, this process typically involves solving a non-linear optimization problem subject to certain constraints on the parameters. The MLE algorithm generates the parameters by improving initial values, either chosen at random or identified by an educated guess. However, there is no guarantee that the parameters' set that maximizes the likelihood will be found. This is known as the local maxima problem. One approach is to choose different starting values over multiple attempts. If the solution stays the same, it can be assumed with some confidence that the algorithm identified a global maximum (Myung, 2003).

The value derived from maximizing the likelihood function provides us with information about which set(s) of specific parameter estimates are more closely related to explaining our data than others. This allows us to perform AIC and likelihood tests, and decide which model, and corresponding set of parameters better fits the data (Claeskens & Hjort, 2008). Transition parameters can be elicited from models that include two or more states. However, there exist a trade-off between number of observations, number of states, and the chances of maximising the likelihood function.

3.3.3 Costs and analysis perspective

The perspective assumed in a cost-effectiveness analysis affects the parameters selected to populate the model. Choosing the relevant perspective is crucial, as a programme that looks unattractive from one perspective could indeed be significantly better when a different perspective is considered (Drummond, 2015). The definitions of these perspectives vary between countries but can generally be grouped into two main categories: the healthcare perspective and the societal perspective.

On the one hand, the healthcare perspective includes all the costs associated with healthcare interventions, treatments, and services. Usually, expenses incurred by both the patient and the healthcare payer are considered. This perspective is usually regarded as the standard approach

by organisations like the National Institute for Health and Care Excellence in the United Kingdom (with exceptions). On the other hand, the societal perspective, adopted in the Netherlands since 2016, includes costs and benefits that do not strictly belong to the healthcare sector (Versteegh et al., 2016). This broad perspective would normally include travel costs, costs related to informal care, time lost from paid work, unpaid work, and leisure time.

In Norway, as of today, health technology appraisals are recommended to follow an extended healthcare perspective (Norwegian Institute of Public Health, 2021). This perspective, in addition to costs directly associated with the treatments, accounts for transport costs, and the patients'/relatives' use of time. A limited societal perspective would consider cost components beyond those captured by the (extended) healthcare perspective (i.e., loss of productivity), but it may not encompass the full range of costs and benefits included in the broader concept of societal perspective (Garrison et al., 2010; Kim et al., 2020).

3.4 Addressing uncertainty

So far, we presented decision analytic models and estimation procedures that try to simplify reality. For that reason, they end up incorporating a great amount of uncertainty and their use to inform policymakers still raises concerns. The inappropriate use of clinical data, the difficulties of extrapolation procedures, and the transparency/validity of the model are all points of reflection that have been raised in the past twenty-five years (Buxton et al., 1997). Nonetheless, the use of models to evaluate health technologies steadily increased (Petrou & Gray, 2011) and good-practice guidelines have been developed to address uncertainty (Briggs et al., 2012). Uncertainty in economic evaluations can be conceptually divided into four categories. Stochastic uncertainty also referred to as first-order uncertainty, captures random variability in outcomes between patients. Heterogeneity, expresses patients' variability caused by the characteristics of those patients. Parameter uncertainty (second-order uncertainty) is associated with the estimation of the parameters. Structural uncertainty (model uncertainty) is generated by all the underlying assumptions inherent in the model (Briggs et al., 2012).

Probabilistic modelling can help in understanding the effects of the internal parameter uncertainty on the outcomes of interest. Probabilistic analyses propagate joint parameter uncertainty through the model; however, the underlying structure was still based on assumptions imposed by the modelling framework (Briggs et al., 2006). Sensitivity analyses can be performed to understand the impact of structural assumptions. This is also relevant in the case of a long/lifetime horizon, where the impact of alternative extrapolation methods beyond the trial/observed data should be explored (Caro et al., 2012). In practice, during a probabilistic analysis, all input parameters are resampled at the same time from chosen probability distributions (Fenwick et al., 2020).

The Beta distribution is bound between 0 and 1 and can be used to sample health-state and probability values. This distribution is characterised by two distinct parameters commonly referred to as alpha and beta. Alpha represents the number of events of interest while beta the number of non-events (total number of events minus alpha). When alpha and beta are not directly available they can be approximated using the method of moments. Using mean (μ) and standard error (SE) the method of moments generates alpha and beta following the calculations in equation 4 (Briggs et al., 2006).

$$\alpha = \mu * \frac{\mu * (1 - \mu)}{SE^2 - 1}$$

$$\beta = (1 - \mu) * \frac{\mu * (1 - \mu)}{SE^2 - 1}$$
[4]

The gamma distribution is commonly regarded as an appropriate fit for costs because it is constrained to the interval 0 to positive infinity. This property aligns well with costs, which are often right skewed and by definition never negative. The distribution is characterized by two parameters, shape and scale. Knowing the mean and standard error, shape and scale can be calculated as:

$$shape = \left(\frac{\mu}{SE}\right)^2$$

$$scale = \frac{SE^2}{\mu}$$
[5]

Beta and gamma distributions are not the only type of distributions that can be used to generate input for the probabilistic analysis. For instance, the Lognormal distribution is a valid and often chosen distribution to inform the model on input parameters such as relative risks, costs, and probabilities. In addition, a more generic approach could also rely on the traditional normal distribution (Briggs, 2011).

Each of the simulations generates an ICER that can be displayed in a cost-effectiveness plane where the y-axis represents increments in costs between interventions (ΔC), and the x-axis increments in effects (ΔE). The results of the simulations can also be displayed through a cost-effectiveness acceptability curve (CEAC). The curve represents the probability of each intervention being cost-effective at a given willingness to pay threshold (λ) per QALY gained (Al, 2012). It should be noted that when comparing technologies, the intervention with the highest probability of being cost-effective might not be the optimal option from a net monetary benefit perspective (Fenwick et al., 2001). To account for this, CEACs are often complemented with a cost-effectiveness acceptability frontier (CEAF). The CEAF plots the points of the CEAC in which the incremental net monetary benefit is higher.

3.5 Value of Information

Economic models based on single trials often incorporate incomplete and imperfect evidence. This means that there is always some underlying risk that decisions made on available information will be suboptimal (Fenwick et al., 2020). One way of addressing such an issue is by acquiring more information and reducing uncertainty. Information, however, comes at a cost and further research should be considered only when deemed valuable. The value of information (VOI) analysis serves as a framework to quantify in monetary terms the value of collecting additional evidence. In the VOI framework, EVPI is a measure that quantifies the value of eliminating uncertainty from all the parameters in the model (Rothery et al., 2020). EVPI can be computed as the difference between the expected value of a decision that assumes perfect information and a decision that is based on current knowledge.

$$EVPI = E_{\theta} \left(\max_i NMB_i(\theta) \right) - \max_i E_{\theta} (NMB_i(\theta))$$

[6]

Where:

θ is a vector of all model parameters

i represents the available decision options

In practice, EVPI's calculations follow a single-loop Monte Carlo scheme and are based on the output of the probabilistic analysis (Rothery, 2020). EVPI reaches its maximum value when the ICER is equal to the WTP threshold. This is because at that value we are most uncertain on whether to consider the technology cost-effective (Briggs et al., 2011). When EVPI is adjusted to account for the actual population that would be affected by the technology, it generates a hypothetical upper-bound value to be gained by investing in additional research. Studies that are expected to cost more than population-adjusted EVPI (pEVPI) should not be carried out (Briggs, 2006).

Additional research will not eliminate all uncertainty from the model. With that in mind the theoretical upper bound value of EVPI is not fully informative as it does not represent what can be reasonably achieved. The expected value of partial perfect information (EVPPI) is the difference between the expected value of a decision that assumes perfect information regarding a parameter or a group of parameters and a decision that was based on the currently available knowledge.

$$EVPPI_{\Phi} = E_{\Phi} \left(\max_i E_{\Psi|\Phi} (NMB_i(\theta)) \right) - \max_i E_{\theta} (NMB_i(\theta))$$

[7]

Where:

θ is a vector of model parameters

Φ is the parameter(s) of interest

Ψ represents the remaining model parameters

EVPPI calculations can be carried out with a nested double-loop Monte Carlo approach. The simulation runs two loops. In the outer loop, the parameter(s) of interest is sampled from its

distribution, in the inner loop the remaining complementary parameters are selected from a distribution conditional on the value sampled in the outer loop (Rothery et al., 2020). Alternatively, EVPPI can also be computed by fitting a non-parametric regression model (generalized additive model) between the NMB and the parameter(s) of interest (Φ).

Groups of parameters for the estimation of EVPPI should be defined according to their nature (costs, HRQoL, transitions, etc.) and in a way that matches the type of research design that would be needed to collect more information (Briggs et al., 2011). Once that different scenarios have been hypothesized, the groups' EVPPI can be used to determine the drivers of current decision uncertainty. As for EVPI, these estimates should also account for the population affected by the decision (pEVPPI) and the useful time of the technology.

4 Methods

This section is structured in three main components. First (Section 4.1), we introduce the trial and the results of the economic evaluation carried out for the 24 months of follow up. Data from these studies will be used to estimate not only transition probabilities but also states' costs and HRQoL.

Second (Section 4.2), we describe the steps towards the conceptualization and optimization of a model structure that extends the results of the trial in terms of state progression. This allows us to project health and economic outcomes beyond the follow-up period. Parameters closely linked to the states such as transition probabilities are estimated using multi-state modelling under different assumptions. The outcomes of these analyses will be presented in the first part of the Results section.

Third (Section 4.3), we present the methodology for carrying out the base-case cost-effectiveness analysis, and for addressing uncertainty. The cost-effectiveness analysis relies on the input parameters generated in sections 4.1 and 4.2. Uncertainty is explored through probabilistic analyses, alternative scenarios, and a preliminary VOI analysis. Outcomes of the economic evaluation will be presented in the second part of the Results section. Parameters' estimations and the cost-effectiveness analyses were carried out in STATA (Stata/SE 17.0) and R (2022.12.0+353), the full code is reported in Appendix C.

4.1 The Hysnes trial

The clinical trial of interest (ClinicalTrials.gov, NCT01926574) was conducted between January 2013 and June 2015. Its aim was to investigate whether group-based rehabilitation programs could promote a sustainable return to work. Previous research on these interventions focused on specific groups of disease, mostly musculoskeletal or psychological disorders. However, the high rates of comorbidity raised the need for interventions able to address the two groups of disorders at the same time. Studies connected to the trial analysed the efficacy (and costs) of two multi-domain interventions, O-ACT and I-MORE (Aasdahl, 2021; 2023).

The O-ACT intervention carried out in the trial, consisted in a 6 weeks-long rehabilitation program with one weekly session of 2.5 hours (Gismervik et al., 2020). In addition to the weekly sessions, patients were given (unsupervised) home assignments, access to group discussions with a physiotherapist, and an individual session at the end of the program with a social worker and a group leader. 80 patients were randomly allocated to O-ACT and 61 of them completed the program.

The second intervention, I-MORE took place at the Hysnes rehabilitation centre, one hour away from the city of Trondheim. The centre was open for 6 years between 2010 and 2016 during

which I-MORE and other rehabilitation programs were carried out (St. Olavs Hospital, 2016). I-MORE lasted 3.5 weeks with sessions of 6 to 7 hours each day. It consisted of several components, such as group-based ACT, education on various topics, mindfulness, group-based physical training, and individual sessions of work-related problem-solving. 86 patients were randomly assigned to I-MORE and 69 completed the program. Table 1 provides a detailed list of the different components of O-ACT and I-MORE.

Table 1: Detailed components of the two return-to-work interventions

Outpatient acceptance and commitment therapy (O-ACT)	Inpatient multimodal occupational rehabilitation (I-MORE)
Location: Outpatient Hospital clinic	Location: Hysnes rehabilitation centre
Duration: 6–7 weeks	Duration: 3.5 weeks
Acceptance and commitment therapy (group sessions: 15h)	Acceptance and commitment therapy (group sessions: 16h)
Discussion and advice on physical activity (group session: 1h)	Sessions with social workers (group sessions and individual guidance: 12h)
Sessions with social workers (individual: 2h)	Work-related problem solving (individual: 5h)
Sessions with social workers and ACT moderator (individual: 0.5h)	Meeting with physician (individual: 0.5h)
Mindfulness session (group: 1.5h)	Mindfulness sessions (group: 3.5h)
Home practice	Outdoor activities day (5h)
Short resume to the GP	Individual return-to-work plan & resume to GP
	“Network day” activity
	“Walking to work” activity
	Lectures (stress, nutrition, pain) (6.5h)

Notes: Reported hours are total hours per type of activity

Main outcomes considered in the study were number of days on medical benefit and time until sustainable RTW (30 days without relapse). During two years of follow-up, I-MORE had a lower value for median number of days on medical benefit (159) compared to O-ACT (249). In addition, a smaller share of patients in I-MORE (54%) transitioned to the more permanent medical benefit (WAA) compared to O-ACT (69%). The hazard rate for sustainable RTW was 1.77 (p -value = 0.01) and in favour of I-MORE (Aasdahl, 2021).

In a following study, the trial was evaluated not only in terms of efficacy but in terms of healthcare consumption and production loss, we reported monetary values in NOK 2023. I-MORE, being an inpatient intervention, was considerably more expensive (NOK 174,000) than O-ACT (NOK 13,570). Primary and secondary healthcare consumption was NOK 77,842 for O-ACT and 55,150 for I-MORE. The difference was mainly driven by a higher use of secondary care

(somatic and psychiatric hospital costs) in O–ACT (difference of NOK 19,635 in favour of I–MORE). Total healthcare cost (including treatment cost) was NOK 91,412 (O–ACT) and NOK 229,150 (I–MORE). The analysis took a step further and calculated production losses over the 2 years. Using daily wage as a proxy for production loss, I–MORE had smaller costs associated with production loss (NOK 810,231) than O–ACT (NOK 969,819). Considering all costs (treatment, healthcare, and production) the difference was NOK 305,350 in favour of I–MORE (Aasdahl, 2023).

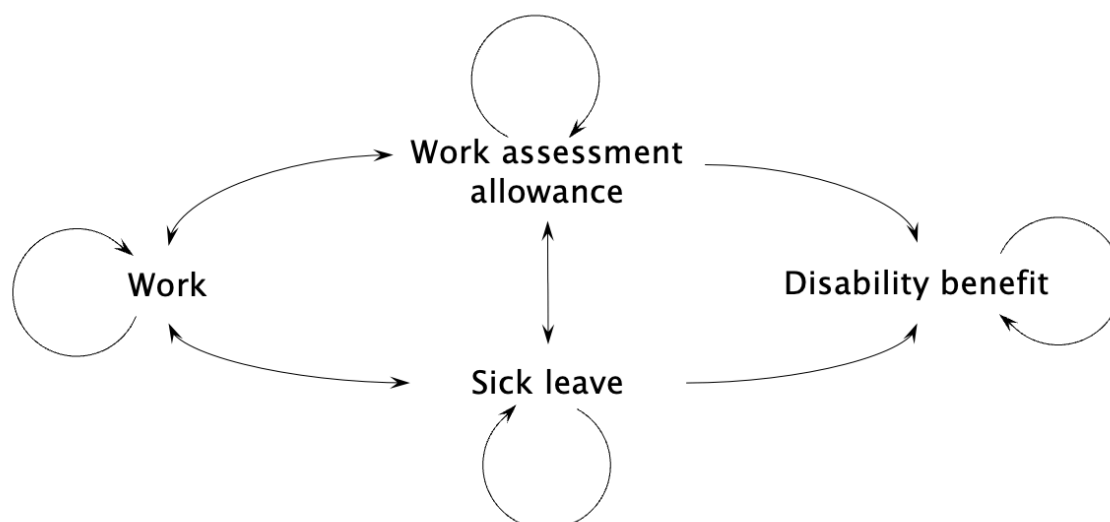
4.1.1 Patient–level data

Patient–level data from the trial were provided in the form of panel data. In our case, monthly observations per patient were available for a period of 24 months (2 years of follow–up). Eligible patients were adults (18–60 years) that at the time of inclusion had been sick listed for 2 to 12 months. Specifically, individuals with an ICPC2 diagnosis within the L (musculoskeletal), P (psychiatric), or A (unspecific disorders) categories were eligible (Aasdahl et al., 2021). Exclusion criteria included substance abuse, serious somatic/psychological disorders, pregnancy, insufficient Norwegian language skills, scheduled surgeries, and serious behavioural problems in group settings. Each observation in the dataset included information on days on sick leave, days on work assessment allowance, percentage of assessed disability, and percentage of employment. The variable capturing the percentage of employment comes from a period during which the registry data were deemed as unreliable. Hence, days on benefits (either sick leave or work assessment allowance) were adjusted for employment percentage only when a second variable on job status matched the information from NAV (Aasdahl, personal communication, 2023). The dataset also included information on patients' characteristics such as age, gender, education, and type of diagnosis at inclusion.

4.2 Multi–State Model

4.2.1 Model structure

Depending on the type of benefit that the patient received during the month, we defined four mutually exclusive and collectively exhaustive model states (i.e., a patient cannot be in more than one state during the month). In the Work (WK) state the patient did not receive any benefit and did not present any disability. In the Sick Leave (SL) state the patient received sick leave benefits for any number of days during the month and did not have any disability. For the Work Assessment Allowance (WAA) state the patient received a specific type of benefit that differed from standard sick leave, they also did not have any disability. In the Disability Benefit (DB) state, the patient presented any percentage of assessed disability, in the model this was an absorbing state. These states were partly based on the Norwegian regulatory context. In absence of any type of benefit, patients were assumed to be fully back at work. Based on these states, the multi–state model was conceptualized as in Figure 1

Figure 1 Model structure with 4 states

These model states and the corresponding model structure, aligned well with other studies in the field of return-to-work interventions and sick leave analysis (Gran et al., 2015; Øyeflaten et al., 2012; Squires et al., 2011). To account for a relatively small sample size, we also considered pooling together people on sick leave and working assessment allowance, this generated an alternative model structure which is further explored at the end of this chapter (4.2.3)

4.2.2 Transition probabilities

We used patient-level data to fit a multi-state model and extrapolate transition probabilities. The multi-state model was fitted using the *msm* package (R). Although the *msm* package was initially developed to model the progression of chronic diseases, the possibility of moving back and forth between states aligned with the conceptualization of our model structure and made *msm* a useful tool to extrapolate sojourn times and transition probabilities. As mentioned, the different states that a patient can transition to are also governed by underlying policies and regulations. For instance, in Norway, workers can be on sick leave for a maximum of one consecutive year. This means that sojourn times in one state affect the intensity of leaving that state, this simple but essential consideration, violates the Markov assumption. The extent to which this violation leads to errors in the estimates depends on how often it is the case that patients stay on sick leave long enough for the regulation to be enforced. However, while such a violation is possible, research shows that it is uncommon (Gran et al., 2015). In our study, the limitation on 12 months of sick leave, should not have too big of an impact as the mean sojourn time in sick leave was of 4.3 months for the O-ACT intervention and 3.87 months for I-MORE. Nevertheless, later in this section, we present one possible approach to the issue of time dependency.

Time-homogeneous models

Our first approach was to assume a scenario in which transition probabilities did not change over time. To fit such base-case models for the O-ACT and I-MORE groups we first had to define the initial intensity matrix iQ . Values in the matrix were assigned 1s when the transition was allowed and 0s when the transition was not allowed. To increase the chances of maximizing the likelihood, initial values for non-zero elements of the iQ should be chosen based on reasonable assumptions (Jackson, 2011; Myung, 2003). However, when only 1s and 0s are provided, and enough data are available, *msm* can estimate initial values. This is the approach chosen for the base-case models for O-ACT and I-MORE. More complex models presented in this paper are built using an initial intensity matrix iQ based on estimates of the base-case models.

$$iQ = \begin{pmatrix} x & 1 & 1 & 0 \\ 1 & x & 1 & 1 \\ 1 & 1 & x & 1 \\ 1 & 0 & 0 & x \end{pmatrix}$$

[8]

The initial matrix is then used for the estimation of the intensity matrix Q and the corresponding probability matrix, which changes according to the length of the time interval chosen for the estimation of the probabilities. To visually assess the goodness of fit of the base case models we plotted the estimated and observed prevalence for each state over time. The probability matrix, calculated using the intensity matrix Q and a cycle length of 1 month, was used to inform transition parameters in the base-case model.

Time-inhomogeneous models

To introduce time dependency, we adapted the base-case models to be piecewise constant. Before generating piecewise constant estimations, we needed to specify time cut-offs at which intensities could vary. The decision on when to set such cut-offs is arbitrary and inevitably leads to different estimates. However, research shows that multi-state models with cut-offs at times that reflect clinical patterns have the best fit when assessed through a Pearson-type goodness-of-fit test (Kunst et al., 2020).

Using Norwegian guidelines on medical benefits we could make an apriori choice on when transition intensities were expected to be significantly different. In Norway, for instance, after 12 months of sick leave, patients can apply for work assessment allowance benefits. As we previously mentioned, at the beginning of the trial patients had been sick listed for 2 to 12 months. More specifically, they had been on sick leave for an average of 7.2 months in the O-ACT group, and 6.9 months in the I-MORE group. This suggested that 5 months after randomization (12 - 7) could represent a time point for a change in the transition intensity. With a greater number of patients switching to work assessment allowance. After the first 12 months of sick leave, patients that want to apply for further sick leave benefits have to go back to work for at least 6 months. This is often done by patients that want to avoid the more permanent

situation and consequences that come with being on work assessment allowance. Based on this information, 11 months (5+6) was selected as a second time point for the piecewise constant model.

We specified piecewise constant models for both groups with cut-offs at 5 and 11 months, in these models, a different intensity matrix (Q) is associated with each time interval. With two cut-offs, we generated three intervals: 0 to 5 months, 5 to 11 months, and 11 months onwards. For each interval, a different probability matrix, with a cycle length of 1 month was estimated. However, splitting the data into three pieces reduced the already limited number of observations (see Discussion).

The goodness of fit between time homogeneous models and piecewise constant models was assessed with likelihood tests, and by comparing the models' Akaike information criteria (AIC) (Claeskens & Hjort, 2008). However, these tests only inform on how well the models fit the data. The visual inspection of the estimated prevalence complemented our decision on which model to use for the extrapolation of transition probabilities.

4.2.3 Alternative 3-state transition model

The Hysnes trial highlighted how I-MORE was indeed more effective in preventing people from transitioning to the work assessment allowance state. This, in combination with the regulatory framework in Norway, led us to define the four states as our base case model. However, due to the limited sample size, uncertainty around model parameters might affect the results of the CEA. To increase observations, and generate more precise transitions, we considered pooling together two states, specifically, SL and WAA. Work (WK) and Disability Benefit (DB) states were not changed. The new General Benefit (GB) state included patients that received sick leave benefits or work assessment allowance for any number of days during the month. In the GB state patients did not present any percentage of disability. As for the previous specification, the states are mutually exclusive. The cost-effectiveness of this model was addressed in a separate scenario analysis.

Figure 2 Alternative model structure with 3 states

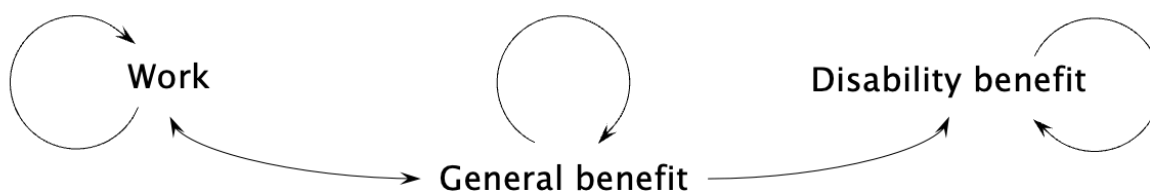


Table 2 Implications of a 4–state structure vs a 3–state structure

4–state model specification	
Pros	Cons
Alignment with regulatory framework	Fewer observations per state
Captures difference between SL and WAA	Increased uncertainty around transition estimates
Comparable with similar studies	
3–state model specification	
Pros	Cons
More observations per state	Cost difference between SL and WAA is not captured
Increased statistical power	

4.3 Cost-effectiveness analysis

4.3.1 Decision analytic approach

Using the same states and transitions identified using multi-state modelling, we developed a decision-analytic model and conducted a cost-effectiveness analysis that compared the health and economic outcomes of O-ACT and I-MORE. A model-based approach allowed us to extrapolate the results of the clinical trial, and its two years of follow-up, to a longer time horizon. The analysis focused on the interventions considered in the clinical trial and did not include any additional comparators.

To achieve such an extension, we opted for a state-transition model. Specifically, given the nature of the data, we modelled a discrete time state-transition model. The base case cost-effectiveness analysis was performed considering an extended healthcare perspective, consistent with Norwegian guidelines (Norwegian Institute of Public Health, 2021). However, to account for broader societal repercussions associated with RTW interventions, we also assessed the cost-effectiveness of the interventions accounting for production loss (limited societal perspective). Following the information from patient-level data, the average patient in the model started the intervention at age 40 years in the sick leave state. Estimated transition probabilities from multi-state modelling were used to generate a Markov trace to which we attached state-related costs and HRQoL. The time horizon was set to 25 years, reflecting the fact that the average retirement age in Norway is 65 years (Storeng et al., 2020). We deemed half-cycle correction unnecessary given that the cycle length was set at 1 month. The main outcomes were total QALYs gained and total costs per intervention group. Due to the long-time horizon, both costs and effects were discounted at a 4% yearly rate (Norwegian Institute of Public Health, 2021). Discounted and undiscounted differences in QALYs and costs were used to compute the ICER. The ICER between I-MORE and O-ACT was compared to a threshold value of NOK 500,000. We also considered a lower threshold of NOK 275,000 and a higher threshold of NOK 825,000. Uncertainty around model parameters was explored through a probabilistic analysis. Separate scenario analyses allowed us to explore assumptions on the underlying structure of the transitions between states. Finally, a preliminary VOI analysis was used to address decision uncertainty and redirect research to specific groups of parameters. All monetary values presented are expressed in NOK 2023 and were computed using a price inflator value of 23.2% between 2016 and 2023 (Statistics Norway, 2023).

4.3.2 Healthcare costs

The trial dataset included individual information on monthly total expenditure by type of healthcare service. Data on primary healthcare consumption was obtained from the Norwegian Health Economics Administration. Primary healthcare consumption included the use of general practitioners, other physicians, psychologists, physiotherapy, psychomotor physiotherapy, manual therapy, chiropractors, and medical imaging. Information on inpatient and outpatient

secondary care was collected from the Norwegian Patient Registry and included the use of both somatic and psychiatric care, rehabilitation, and visits to private specialists.

The costs of the two interventions were calculated with the method of time-driven activity-based costing (Kaplan et al., 2009). A detailed analysis can be found in the supplementary material of the trial's economic evaluation (Aasdahl, 2023). The average cost per patient in I-MORE was NOK 174,000, while O-ACT presented a much lower cost of NOK 13,570. In the model, intervention costs were considered as an initial lumpsum cost and were not affected by discounting.

We grouped costs associated with each health service, primary and secondary, in a variable that captured monthly total healthcare consumption per patient. We first performed a t-test (Appendix A, Table 1 and 2) to determine whether healthcare consumption was different between the first and second year after the intervention. The cost difference was not significant at any usual level (p -value=0.25). However, given that costs are often skewed, we transformed them on the logarithmic scale. Even after transforming them, the difference in costs between the 2 years was not statistically significant (p -value=0.39). We then proceeded to the estimation of time-invariant state costs. We first ran a panel regression of intervention group on total costs (we repeated the regression for the transformed costs). In both cases the intervention group was not significant in reducing/increasing healthcare consumption. With this in mind, we proceed to estimate state-related costs that did not depend on the intervention group. To do so, we performed another panel regression of benefit type (model states) on healthcare consumption. Regression coefficients were estimated using both fixed effects and random effects methods. After we generated the two models, we performed Hausman test to determine whether a fixed effects model would be preferred. Given the test's non-significance (p -value= 0.1255) we opted for the random effects model. Estimates generated with the fixed effects model can be found in Table 3 of Appendix A. The standard errors of the regression coefficients were generated using a bootstrap algorithm with 1000 iterations. Using the delta method, we combined the standard errors of the coefficients into standard errors associated with each state. These values were used to inform cost distributions in the probabilistic analysis.

To ensure the reliability of the random effects estimates and test their robustness, we also approached the regression of costs and HRQoL with a generalized linear model (GLM). Specifically, we regressed model states on costs and HRQoL using a GLM with family gamma and link log (Appendix A, Table 3).

4.3.3 Production loss

For each state in the model, costs associated with production loss were calculated by multiplying the average number of absence days with a proxy for daily production loss. Absence was based on days on sick leave, work assessment allowance, and disability benefits. To better capture the

difference in absence between I-MORE and O-ACT we also stratified by intervention group. Per definition, working individuals (WK state) had no absence days. In the O-ACT group, absent days were on average 20, 21, and 13 for the SL, WAA, and DB states respectively. In I-MORE absence was slightly reduced, with absent days going down to 18, 21, and 10 for the SL, WAA, and DB states, respectively. Daily production loss was based on reported national wage multiplied by social expenses (40%). Since our data on days absent from work could range from 0 to 31, we computed daily wage (NOK 1,767) by dividing the average monthly wage (NOK 53,765) by 31 (30.42). Total production loss for one day of absence was NOK 2,473.

4.3.4 Health related quality of life

Information regarding patients' HRQoL was collected using the 15D instrument within the Hysnes trial. The original 15D algorithm relied on Finnish preferences to elicit HRQoL values. However, more recently, the instrument has also been validated for the general Norwegian population (Michel et al., 2019). The Norwegian value algorithm was used in the Hysnes trial, during which no negative scores were registered, and the lowest obtained was 0.153.

Data from the 15D were available at the beginning of treatment (\approx 1 month), 5, 8, and 14 months. However, a significant portion was missing due to loss of follow-up. To partially account for missing data, missing values were predicted with a single imputation approach. A variety of methods have been proposed throughout the years, from mean imputation to regression imputation (Faria et al., 2014; Baraldi and Enders, 2010; Molenberghs et al., 2004). Given the high association of HRQoL with baseline characteristics such as age, gender, (etc...) regression is often the preferred approach for single imputation (Faria et al., 2014). In our case predictions were based on a regression model that included previous HRQoL, age, gender, education, marital status, disability status, pain level, anxiety and depression levels, intervention group, ICPC2 diagnosis, and absence.

As for healthcare consumption we first conducted a panel regression of intervention group on 15D scores. The estimated coefficient was not statistically significant at any usual level. Besides negligible initial differences, possibly due to the inpatient nature of I-MORE, HRQoL was assumed to be driven exclusively by the type of working state. We pooled data from both groups to obtain point estimates and ran a panel regression (both fixed effects and random effects) of states on 15D scores. The non-significance of the Hausman test (p -value=0.051) led to the decision of informing the model using the output of the random effects model. Given that during the second year of follow-up no HRQoL data were available, all regressions only considered the first 14 months. Following the same approach adopted for cost parameters, bootstrapped standard errors were based on 1000 iterations. Standard errors for the states' HRQoL were computed with the delta method.

4.3.5 Deterministic analysis

The deterministic analysis assumed that transition intensities did not change over time. Hence, state-related transition probabilities were fixed over the entire time horizon. Although Norwegian guidelines recommend an extended healthcare perspective, transportation costs and time in connection with treatment were not included in our analysis due to the unavailability of these data. Therefore, state costs were based on primary and secondary care only. First, we first simulated the model for 2 years and compared the results with the outcomes of the trial's economic evaluation. Direct comparison was only possible for costs, as the previous evaluation expressed effects as number of working days but not in QALYs (Aasdahl, 2023). Second, we extended the time horizon to 25 years. To assess the impact of discounting we reported both discounted and undiscounted deterministic ICERs.

We also considered a limited societal perspective, which in addition to the costs of the healthcare perspective, accounted for production losses. Again, we first computed the ICER for a time horizon of 2 years and then for a time horizon of 25 years. For the ICER based on the longer time horizon, we reported both discounted and undiscounted values.

4.3.6 Probabilistic analysis

To address joint parameter uncertainty and evaluate the robustness of the results we conducted a probabilistic analysis. The base-case probabilistic analysis was performed by running 10,000 iterations of the model. Each iteration was based on random draws from the parametric distributions assigned to the input parameters (Table 3). We assigned beta distributions to all the transition probabilities to ensure that random values could not be generated outside the interval 0–1 (Briggs et al., 2006). We used the method of moments to inform the parameters of each beta distribution. The random sampling of costs was based on gamma distributions. Mean and standard errors were used to compute shape and scale. HRQoL values elicited from the 15D questionnaire are bounded between 0 and 1. Similarly to probabilities, we opted for beta distributions. The alpha and beta parameters for HRQoL distributions were calculated using the output from the previously addressed panel regressions and the method of moments.

Table 3 Input parameters for the probabilistic analysis

Intervention	Input parameter	Mean	SE ^a	Distribution ^b	Type of parameter
O-ACT	WK-SL	0.05985	0.0108	Beta	Probability
	WK-WAA	0.00815	0.0035	Beta	Probability
	SL-WK	0.1070	0.0137	Beta	Probability
	SL-WAA	0.09815	0.0137	Beta	Probability
	SL-DB	0.00034	0.0001	Beta	Probability
	WAA-WK	0.01405	0.0041	Beta	Probability
	WAA-SL	0.00046	0.0002	Beta	Probability
	WAA-DB	0.00588	0.0027	Beta	Probability
	Treatment cost	11,033		Fixed	Cost
	Production loss SL	49,460		Fixed	Production loss
	Production loss WAA	51,933		Fixed	Production loss
Production loss DB	32,149		Fixed	Production loss	
I-MORE	WK-SL	0.09563	0.0123	Beta	Probability
	WK-WAA	0.00502	0.0021	Beta	Probability
	SL-WK	0.15511	0.0155	Beta	Probability
	SL-WAA	0.06195	0.0171	Beta	Probability
	SL-DB	0.00171	0.0001	Beta	Probability
	WAA-WK	0.01496	0.0048	Beta	Probability
	WAA-SL	0.00081	0.0003	Beta	Probability
	WAA-DB	0.00174	0.0022	Beta	Probability
	Treatment cost	141,455		Fixed	Cost
	Production loss SL	44,514		Fixed	Production loss
	Production loss WAA	51,933		Fixed	Production loss
Production loss DB	24,730		Fixed	Production loss	
O-ACT & I-MORE	c_WK	785	198	Gamma	State cost
	c_SL	3,229	665	Gamma	State cost
	c_WAA	2,922	496	Gamma	State cost
	c_DB	1,852	480	Gamma	State cost
	u_WK	0.699	0.011	Beta	HRQoL
	u_SL	0.608	0.006	Beta	HRQoL
	u_WAA	0.628	0.009	Beta	HRQoL
	u_DB	0.582	0.023	Beta	HRQoL

Notes:

All costs (and respective SE) are reported in NOK 2016 (In the analysis we accounted for a price inflator value of 23.2%)

Production loss is reported in NOK 2023

a) Standard errors for transition probabilities were computed from 95% CI following Briggs (2011)

a) Standard errors for HRQoL were computed using the Delta method (STATA)

b) "Fixed" parameters were not varied in the probabilistic analysis

4.3.7 Scenario analyses

Throughout our analysis, we had to make several assumptions both in terms of model structure and time dependency. Therefore, we defined two alternative scenarios to address structural uncertainty and explore how it affected our base-case probabilistic results.

Scenario A: piecewise constant model

The base-case analysis assumed that transition intensities did not change over time. Hence, transition probabilities associated with each cycle were fixed. However, in the first part of the Methods section, we introduced a partial solution to time independency, the piecewise constant exponential model. With it, we generated an array of transition probabilities matrices that changed at specific cut-offs (5 & 11 months). Similarly to the base-case scenario, we used the array to populate the model and generate a Markov trace to which we attached costs and HRQoL. In this scenario, only the transition parameters were affected. Costs and HRQoL associated with model states remained the same.

Scenario B: 3-state model

As the small sample size was the main limitation of our data, we decided to decrease the number of transition parameters being estimated. We pooled together people on sick leave and work assessment benefits and defined the general benefit (GB) state. This model structure (Figure 2) included 3 states: Work, General Benefits, and Disability Benefits. As in the four-state structure, people at work cannot directly transition to disability benefits. Cost parameters and HRQoL parameters for this specification were computed with the same approach used for parameters in the base-case analysis, this time using 3 states. An in-depth list of the input parameters for scenario A and scenario B can be found in the appendix (Appendix A, Tables 4 and 5).

4.3.8 Value of information analysis

Following the results from the probabilistic analysis, we computed the NMB between I-MORE and O-ACT. Calculations of the NMB were based on equation [1] in which we first multiplied the average QALYs of the interventions with the chosen threshold value (i.e., NOK 500,000) and then subtracted the average cost. To quantify the value of perfect knowledge we used the output of the probabilistic analysis to calculate the NMB for each of the 10,000 iterations. Then, following equation [2] we computed EVPI.

Using the Sheffield Accelerated Value of Information (SAVI) tool (Strong et al., 2013), we initially calculated EVPPI for individual parameters, parameters linked to the transitions between WAA and WK (for both O-ACT and I-MORE) and those associated with the DB were the one with higher EVPPI (Appendix A, Figure 2). However, when we consider parameters individually their impact on the NMB might not be sufficient to be addressed. Moreover, it is unlikely that further research will focus on a single parameter of the model. The next step was to calculate EVPPI for groups

of parameters that could be reasonably explored together. Groups' EVPPI was calculated using a GAM regression in R.

The first group of parameters (Group 1) included the costs associated with healthcare consumption in the WK, SL, WAA, and DB states. The way we defined these states allowed for comparability with other studies. Indeed, to gather this type of information we could use existing literature, cutting down on the costs of additional research. Similarly, we grouped state-HRQoL (Group 2). In our analysis, due to a significant loss of follow up, these values were based on imputed data. Further research (e.g., surveys) could provide more insight into the perceived HRQoL of patients on medical benefits. The final group (Group 3) included all the transition parameters associated with I-MORE. To explore this scenario, we would have to repeat the trial (or variations of it). However, it is possible that such a study would also reduce uncertainty around other parameters. All specific parameters and the way they were grouped can be found in the appendix (Appendix A, Table 6)

To account for the population that could benefit from the intervention. We had to rely on several assumption. In 2023, more than 4 million people were working in Norway (Statistics Norway). Estimates of the prevalence of psychological disorders in Norway have great variation. However, following the results from large population studies such as the HUNT (NTNU, 2019) we assumed a prevalence of 22.9% among the adult population. Yearly incidence was set to 2.47% (Nystuen et al., 2001). Values for musculoskeletal disorders present even higher uncertainty. Nevertheless, following the results of the HUNT study, we assumed the prevalence to be 47.9%. Incidence for this group of disorders was set to 7.9% (Hagen et al., 2006).

Using an arbitrary number of useful years, in our case 10, and a discount rate of 4%. We estimated that the population affected by the decision could be as high as 6,008,657 individuals. We used this value to calculate pEVPI and pEVPPI. Due to the discrepancies around the information on sick leave prevalence (and incidence) in combination with musculoskeletal and psychological disorders, this population value should only be seen as an upper bound value (see Discussion).

4.4 Model validation

Throughout the various steps of the analyses, we carried out checks to assess the internal and external validation of our model. For instance, in line with the AdViSHE guidelines for health-economic models (Vemer et al., 2015), we included in the code a function that checked whether the array of transition matrices (homogeneous and inhomogeneous) added up to 1 for each cycle in the model. Although not many models have been developed for RTW treatments, cross validity has been addressed by comparing our model states with models developed to assess either the cost-effectiveness of the interventions (Squires et al., 2011) or with models that aimed at eliciting transition probabilities between the different types of benefit (Gran et al., 2015).

However, comparability with other studies is limited, as I-MORE was an intervention tailored to the Hysnes facility and the trial. We performed scenario analyses to explore structural uncertainty and also compare the outcomes of the different models against empirical data. A detailed validation section is reported in Appendix B using the TECH-VER checklist (Büyükkaramikli et al., 2019)

4.5 Statement of ethical approval

For the various analyses we used patient level data stored at the Norwegian University of Science and Technology (NTNU). To access the dataset on a secured server, we requested the approval of the regional committees for medical and health research ethics (REK). Approval was granted in December 2022.

5 Results

5.1 Parameter estimation

5.1.1 Time-homogeneous transition probabilities

For each intervention, we obtained a matrix that captured time-independent transition probabilities between states assuming a 1-month cycle length (Table 4). The fitted multi-state model estimated that the O-ACT WK prevalence peaked at around 10 months (37%) and then decreased over time. A similar trend was highlighted for I-MORE, although the 10-month peak registered a higher prevalence (50%) (Figure 3).

Table 4 Time-homogeneous transition probability matrices

O-ACT					I-MORE				
	WK	SL	WAA	DB		WK	SL	WAA	DB
WK	0.9320	0.0598	0.0081	0.0000	WK	0.8992	0.0957	0.0051	0.0000
SL	0.1070	0.7945	0.0981	0.0003	SL	0.1552	0.7812	0.0635	0.0001
WAA	0.0140	0.0005	0.9796	0.0059	WAA	0.0149	0.0008	0.9809	0.0034
DB	0.0000	0.0000	0.0000	1.0000	DB	0.0000	0.0000	0.0000	1.0000

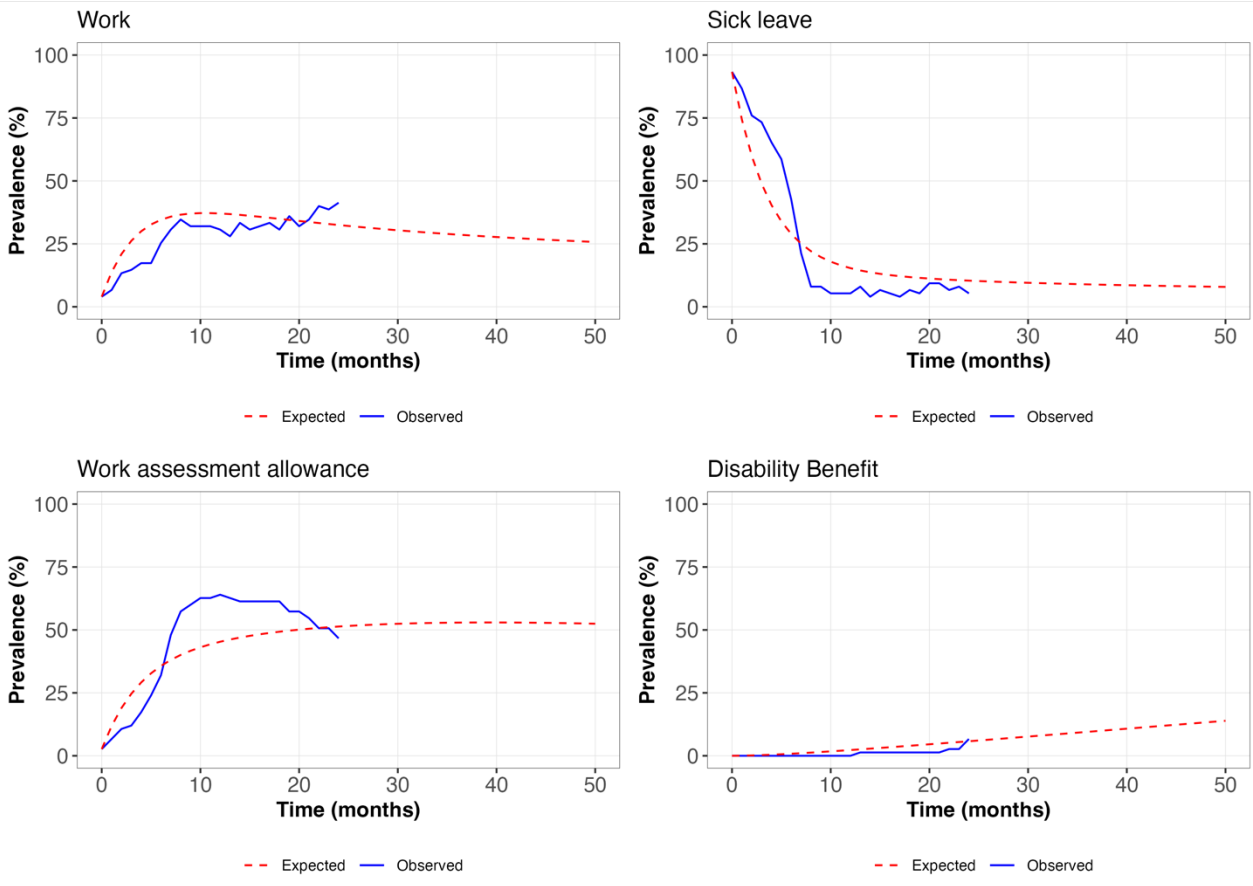
Notes:
O-ACT (outpatient acceptance and commitment therapy)
I-MORE (inpatient multimodal rehabilitation program)
WK: work
SL: sick leave
WAA: work assessment allowance
DB: disability benefit

When we compared the observed and expected state-prevalences over time, we found that the observed and expected trends had good overall correspondence except for several time points (Figure 3). For instance, at 12 months, the observed SL prevalence for O-ACT was 5.3% whereas the expected value generated by the matrix was 15%. Likewise, at 12 months, the values for I-MORE were also misaligned. The observed prevalence for the SL state was 15%, while the expected value was 22%. At 24 months, the last available follow-up time, the prevalence estimates improved for both interventions. For instance, in O-ACT, the observed and expected values for SL were 5.3% and 10.4%, respectively. In I-MORE SL prevalence was 12% observed and 17% expected.

The estimated transition probabilities can also be visually presented in terms of stacked probability plots. In Figure 4 we showed the probability of transitioning to all the model states starting from the SL state. Importantly, I-MORE reduced the probability of transitioning both to WAA and DB as shown by the larger WK region (Figure 3, right panel). Stacked probabilities starting from other states can be found in the appendix (Appendix A Figure 1).

Figure 3 States' prevalence over time for O-ACT and I-MORE in base-case model

O-ACT



I-MORE

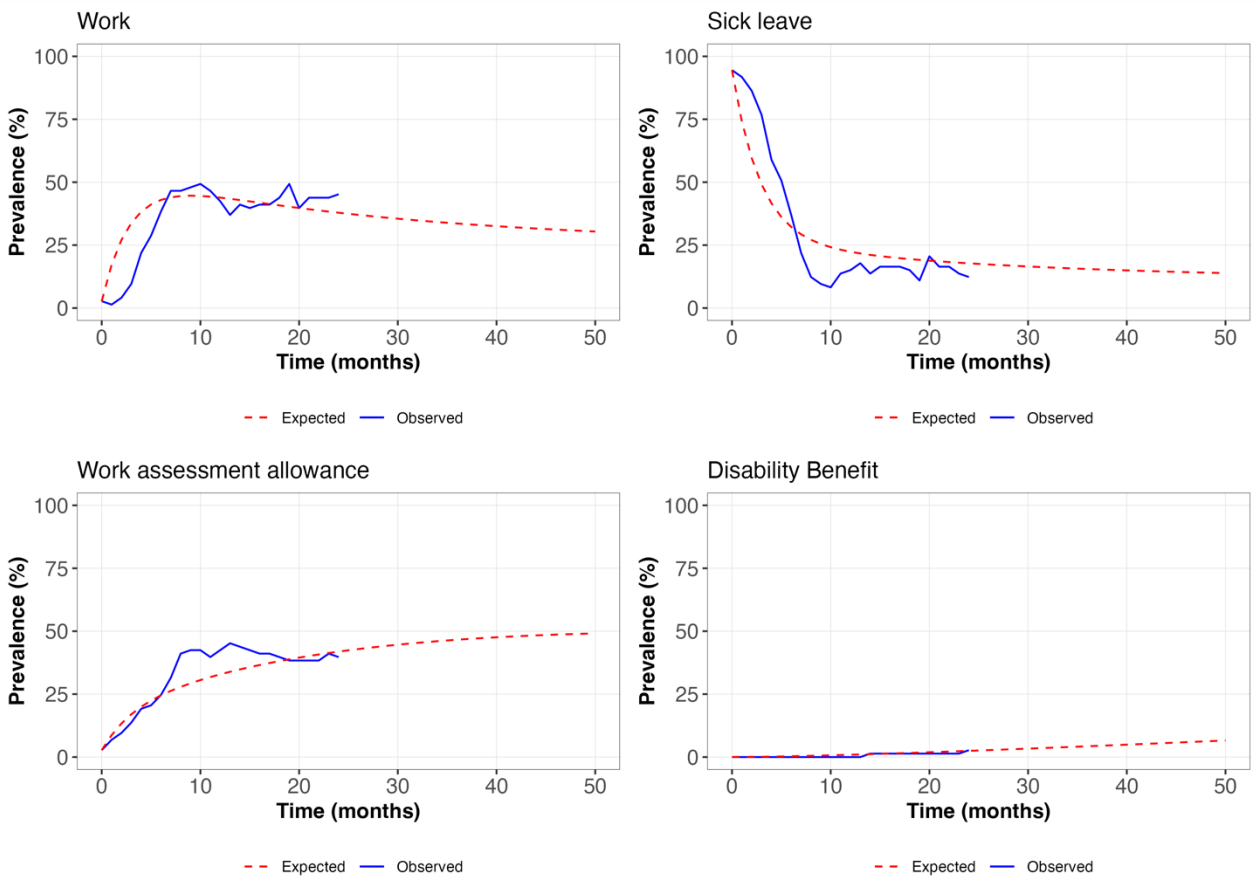
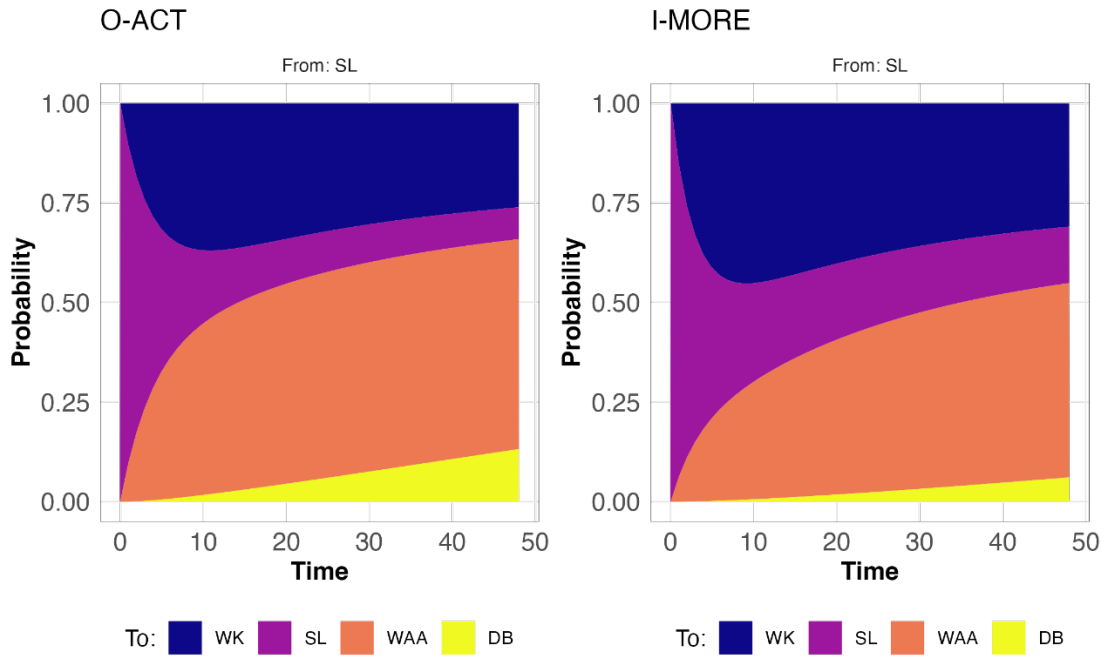


Figure 4 Time-homogeneous transition probabilities from Sick Leave state to all states



5.1.2 Time-inhomogeneous transition probabilities

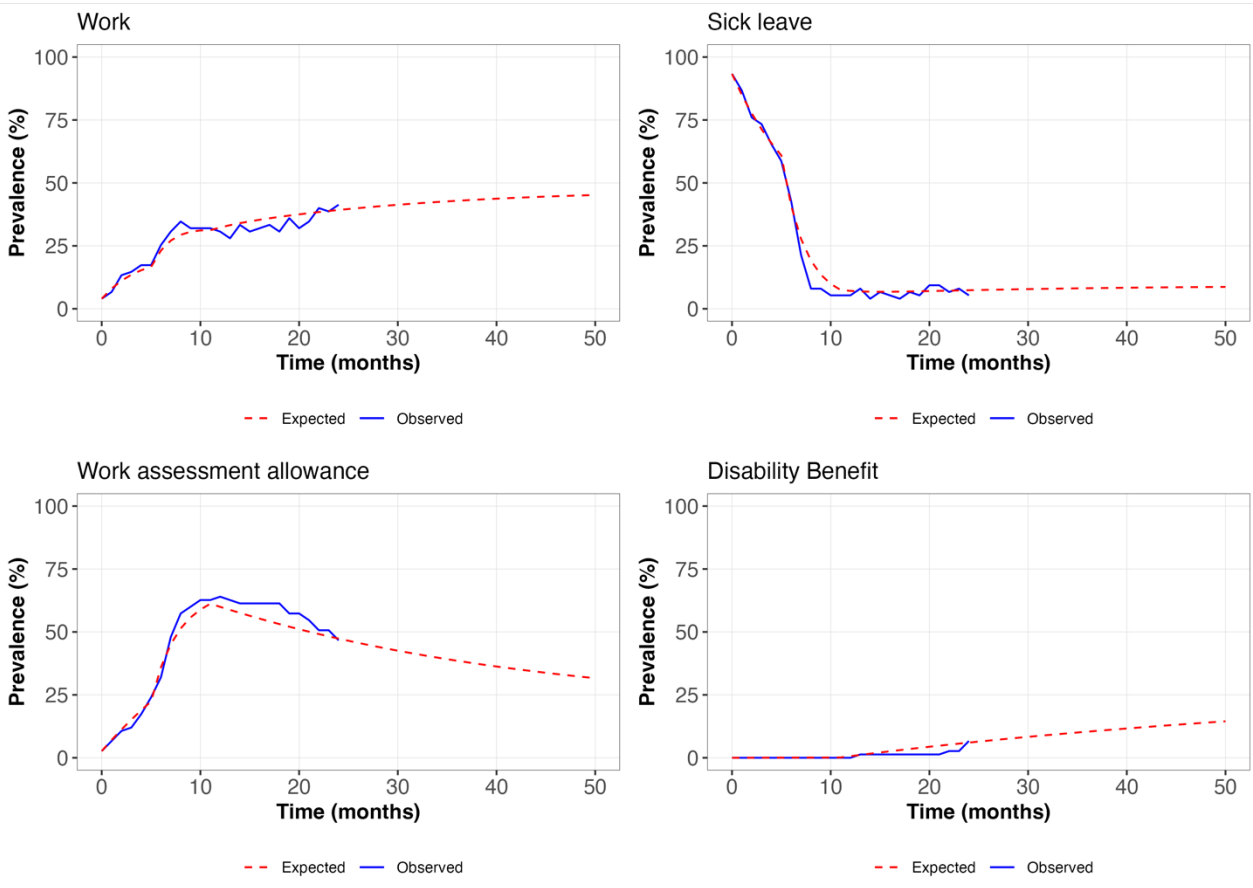
The piecewise exponential model, with cut-offs at 5 and 11 months, relaxed the assumption of constant baseline transitions and generated three transition probability matrices for O-ACT and three matrices for I-MORE (Appendix A, Table 4). When compared to the time-homogeneous model, the fitted piecewise constant model changed the expected prevalence trends, especially for O-ACT. For instance, instead of peaking at 10 months and then slowly decline, the WK prevalence for O-ACT reached its peak at month 70 (46%). On the other hand, I-MORE maintained its peak at 10 months (45%), with the overall trend for this state resembling the one in the time-homogeneous case (Figure 5).

The comparison between observed and expected prevalence (Figure 5) demonstrated that having time-inhomogeneous probabilities significantly reduced the disparity between the two values. For instance, at 12 months, the observed SL prevalence for O-ACT was 5.3% while the expected value was 5.5%. For I-MORE, the observed and expected values were 15% and 13% respectively.

The overall visual assessment of the time-homogeneous and time-inhomogeneous models (Figure 3 and Figure 5) showed that the piecewise constant model better fitted the events of the trial. We also performed a likelihood ratio test between the time-homogeneous and time inhomogeneous models which confirmed that, at any usual level of significance (O-ACT p -value= $5e-13$; I-MORE p -value= $2e-07$), the piecewise constant estimates were preferred. For O-ACT the corresponding Akaike information criterion was 1079 (homogeneous) and 1018 (inhomogeneous). While for I-MORE, 1286 (homogeneous) and 1255 (inhomogeneous).

Figure 5 States' prevalence over time for O-ACT and I-MORE in piecewise constant model

O-ACT



I-MORE

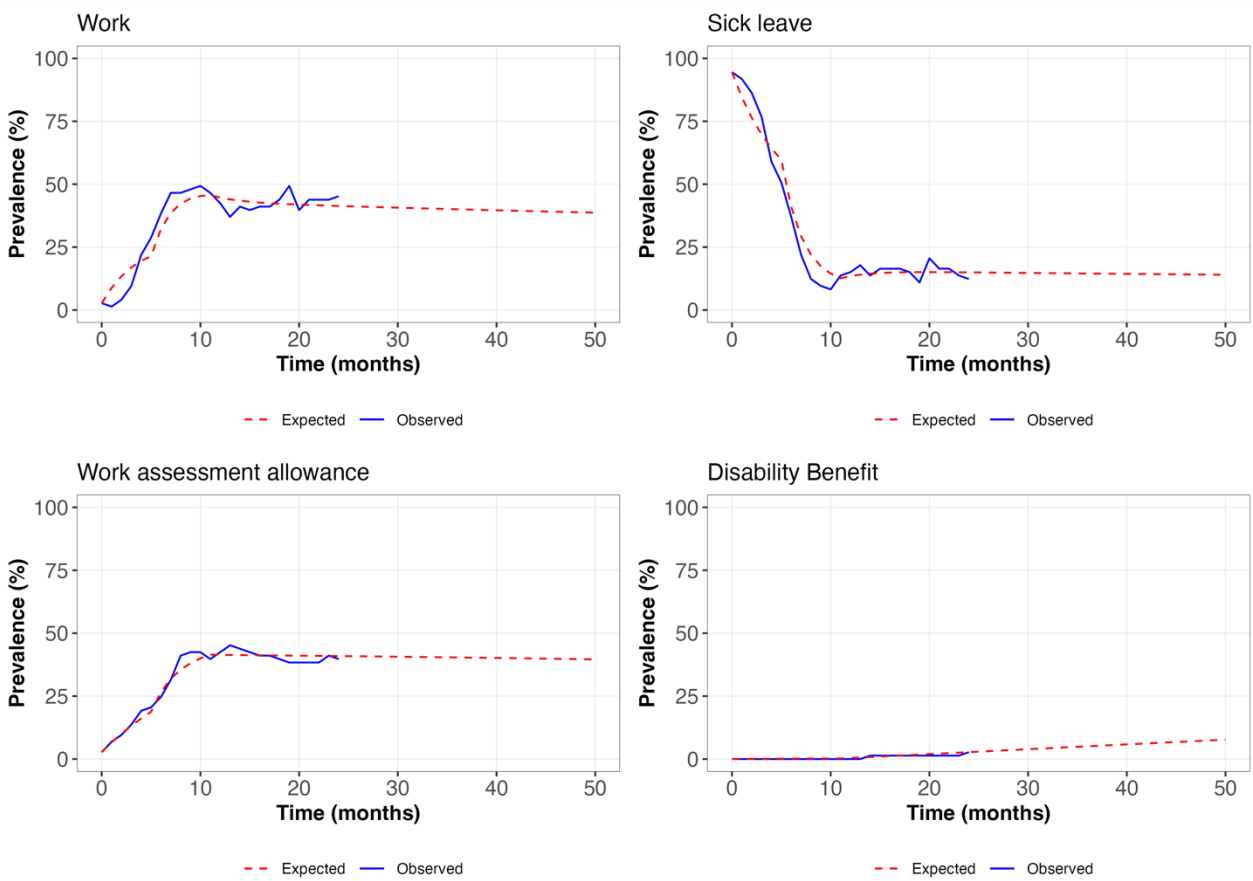
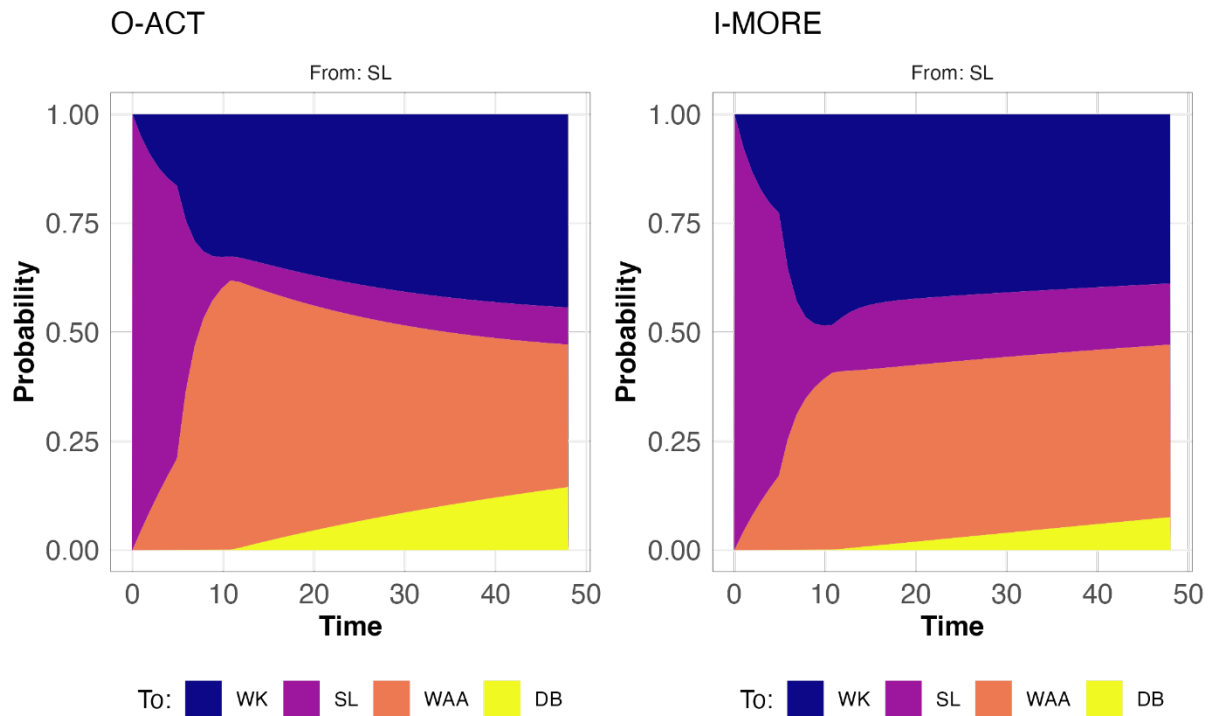


Figure 6 Time-inhomogeneous transition probabilities from Sick Leave state to all states

5.1.3 3-state transition probabilities

When we pooled together patients in the SL and WAA states in a 3-state model (Figure 2), we found that WK prevalence for both O-ACT and I-MORE peaked at month 20 with an expected prevalence of 37% and 44% respectively. While they had a different timing, these values resembled the peak values registered in the base-case scenario (37% and 50%).

Although I-MORE reached a higher WK prevalence (and faster than O-ACT), the overall trends for the WK and GB states were reasonably similar between the interventions. For instance, at the end of the follow-up the observed prevalence in GB was 54% (vs 56% expected) for O-ACT and 52% (vs 53% expected) for I-MORE. The increased number of observations produced estimates that, even without time-dependency (piecewise constant), accurately matched the observed data (Figure 7).

The expected prevalence of the DB state was almost identical to the 4-state model. The effects of I-MORE on the transition probabilities can also be visualised in terms of stacked probability plots. Starting from the GB state, I-MORE reduced the probability of transitioning to the absorbing state (smaller DB region) and led to an increased initial probability of entering the WK state (Figure 8).

Figure 7 States' prevalence over time for I-MORE in piecewise constant model

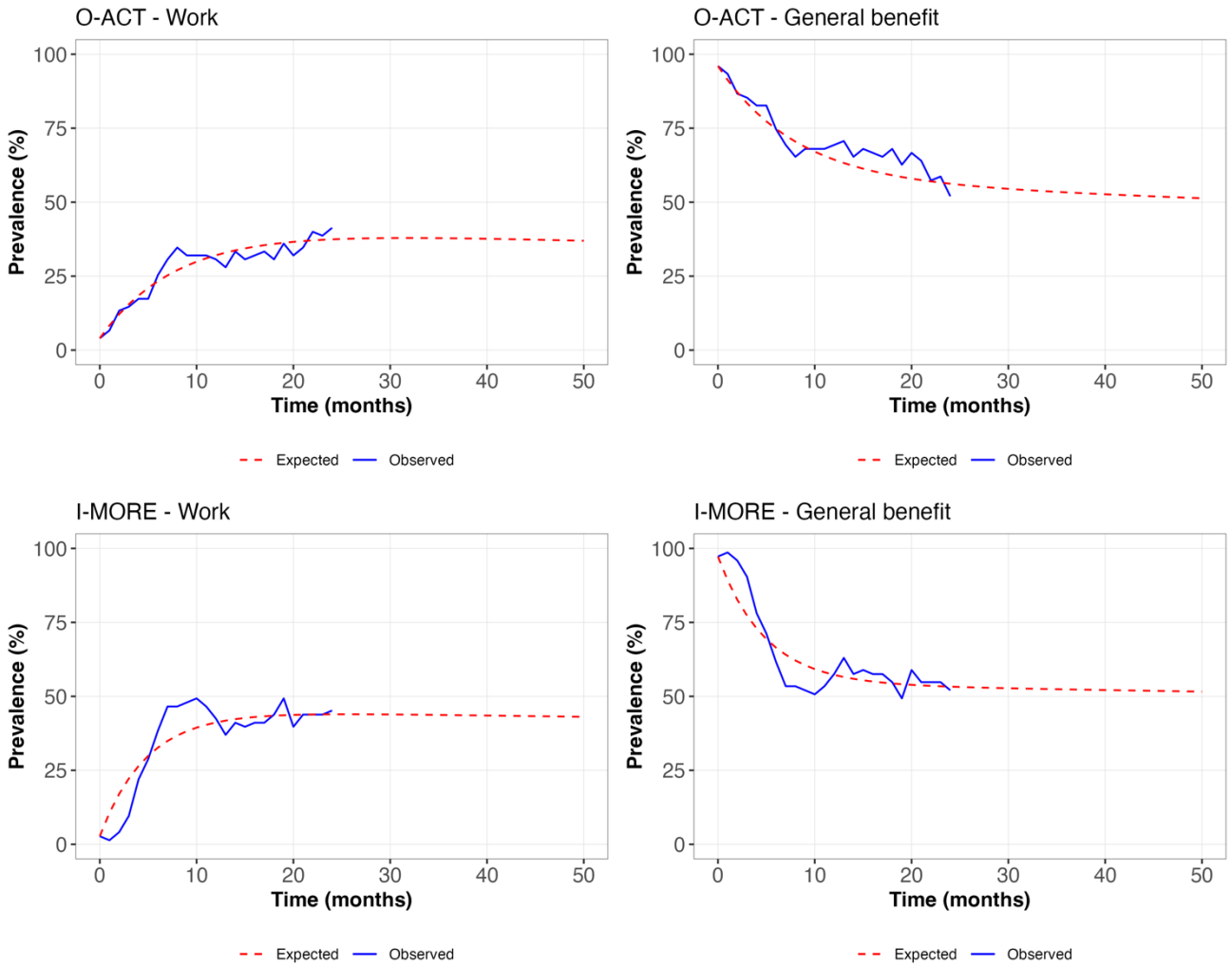
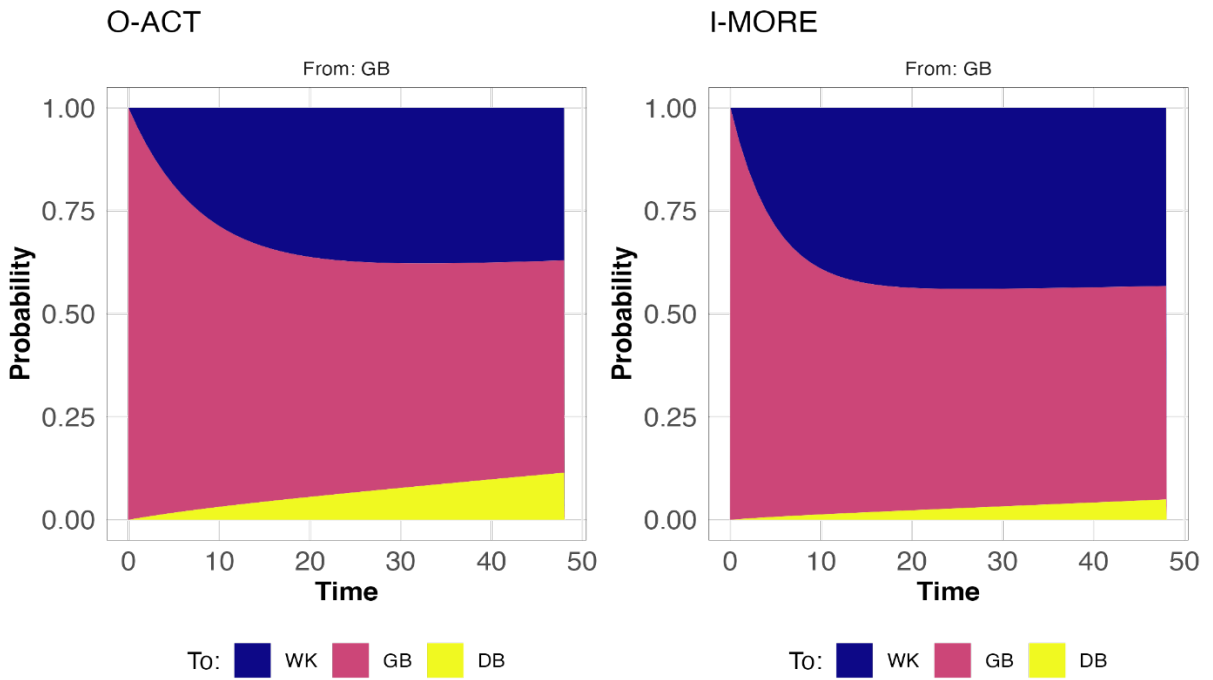


Figure 8 Transition probabilities form General Benefit state to all states



5.1.4 Cost and HRQoL estimates

The bootstrapped estimates (Table 5) highlighted that the WK state was associated with the lowest healthcare consumption and resulted in a monthly state-cost of NOK 785. In contrast, the SL and WAA states were the most expensive ones, with monthly healthcare costs of NOK 3,229 and NOK 2,922 respectively. The DB state, with a cost of NOK 1,852 was lower than SL and WAA but higher than WK.

Table 5 State-costs from a random effects panel regression

State	Cost (NOK)	SE	95% CI	
WK	785	198	380	1,190
SL	3,229	665	1,905	4,554
WAA	2,922	496	1,937	3,906
DB	1,852	480	888	2,815

Notes:

WK: work

SL: sick leave

WAA: work assessment allowance

DB: disability benefit

When we tested whether we should assume random or fix effects for the estimation of state-HRQoL, we found a borderline significant Hausman test (p-value=0.051), however, given that the fixed effect model produced inconsistent estimates (highest HRQoL associated with the disability benefit state) we proceeded with the random effects approach. Starting from the highest value in the WK state (0.7), scores decreased in SL and WAA (≈ 0.61) and reached the lowest value in the DB state (0.58).

Table 6 State-HRQoL from a random effects panel regression

State	HRQoL	SE	95% CI	
WK	0.699	0.0110	0.678	0.721
SL	0.608	0.0056	0.597	0.619
WAA	0.628	0.0091	0.608	0.648
DB	0.582	0.0227	0.538	0.627

Notes:

WK: work

SL: sick leave

WAA: work assessment allowance

DB: disability benefit

For both costs and HRQoL, the estimates generated with the GLM models were systematically consistent with the results of the random effects model (see Appendix A Table 3). Indeed, this confirmed that the estimates of the fixed effects regressions would have not been robust.

5.2 Cost-effectiveness analysis

5.2.1 Deterministic results

When we initially performed the deterministic analysis over a 2-year time horizon. Considering a healthcare perspective, the total cost of I-MORE was NOK 241,589 and the total cost of O-ACT was NOK 84,714. These values were consistent with the costs estimated in the trial, i.e., NOK 229,498 (I-MORE) and NOK 91,449 (O-ACT). The difference in costs between the two interventions was mainly driven by the higher price of the I-MORE intervention, NOK 174,273 vs NOK 13,593 for O-ACT. We estimated positive incremental QALYs over the 2-year for I-MORE compared with O-ACT of 0.0094, yielding the deterministic ICER of 16,691,624 NOK per QALY gained. When we increased the time horizon to 25 years, the deterministic discounted ICER went down to NOK 870,996 per QALY gained (0.201 QALYs gained) but still remained higher than benchmark WTP thresholds in Norway.

Once we accounted for additional costs savings, through averted production loss over a 25-year period, I-MORE dominated O-ACT. Total costs associated with I-MORE (healthcare consumption and production loss) amounted to NOK 6,995,929, while O-ACT amounted to NOK 7,185,559. The incremental effect remained the same as in the healthcare perspective (0.201 QALYs).

5.2.2 Probabilistic analysis

Over a 25-year period, the base-case ICER resulting from the probabilistic analysis with a healthcare cost perspective was NOK 1,167,887 per QALY gained. I-MORE would not be considered cost-effective according to the benchmark thresholds. Indeed, at a WTP threshold of NOK 275,000 I-MORE was cost effective in only 3% of the cases. At NOK 500,000 the proportion of cost-effective simulations went up to 15%. With the higher WTP threshold of NOK 825,000 it reached 37% (Figure 10).

When we included the costs associated with production loss, we found that similarly to the deterministic results, I-MORE was considered dominant (negative incremental costs of NOK -284,969 and positive incremental effects of 0.145 QALYs). With this perspective at a WTP threshold of NOK 500,000, I-MORE was cost-effective in 72% of the simulations (Appendix A, Figure 2).

Figure 9 Cost-effectiveness plane with a healthcare perspective and WPT of NOK 500,000

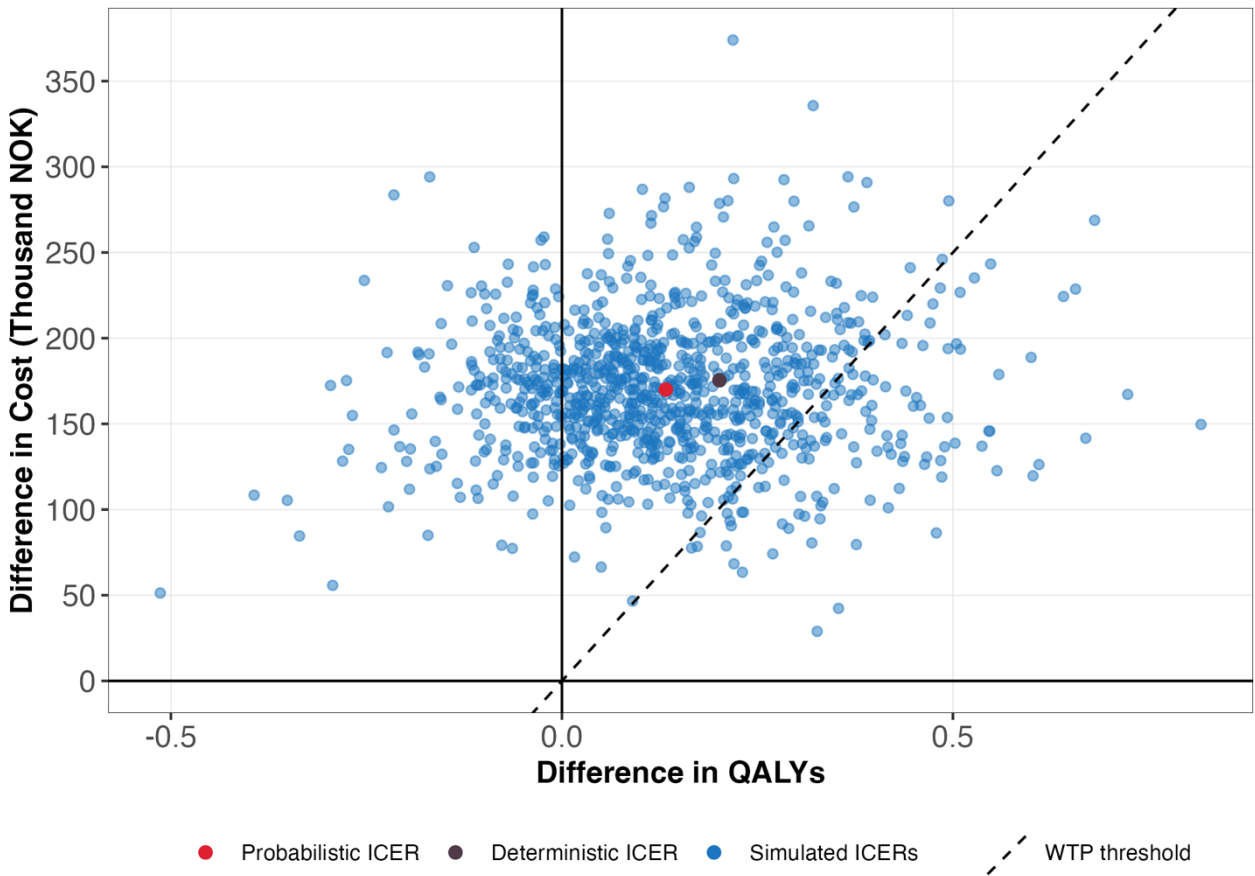
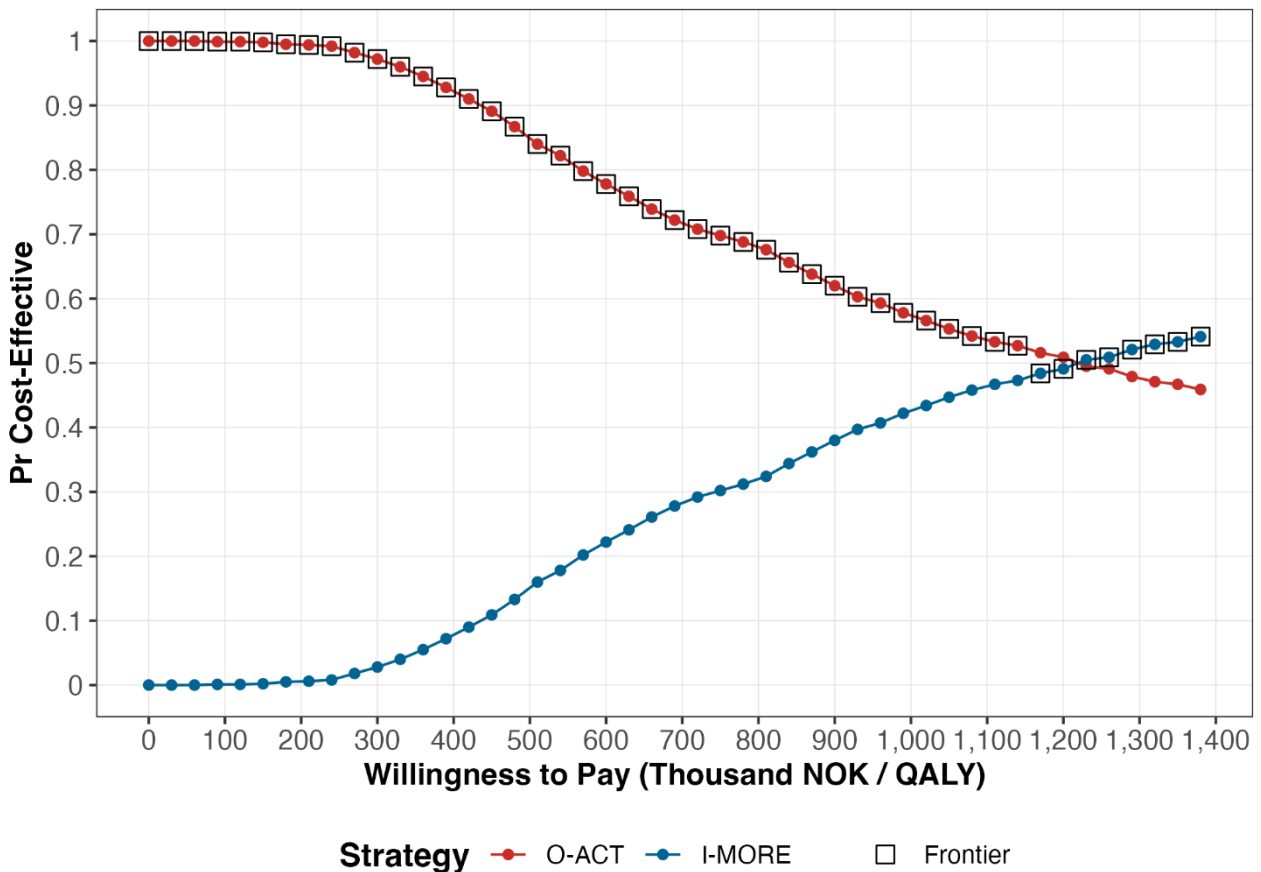


Figure 10 Cost-effectiveness acceptability curves and frontiers with a healthcare perspective



5.2.3 Alternative scenarios

To explore the effects of structural uncertainty in our model, we defined scenario A (piecewise constant model) and scenario B (3-state model). The numbers presented refer to the outcomes of probabilistic analyses (10,000 simulations).

Scenario A (piecewise constant model)

When we ran the piecewise constant model for 2 years, considering a healthcare cost perspective, we found that total cost was NOK 86,564 for O-ACT and NOK 243,017 for I-MORE (incremental cost of NOK 156,452). These numbers matched the outcomes of the trial's evaluation which estimated incremental costs of NOK 138,049. Once we extended the time horizon to 25 years, we found that although the incremental cost remained positive (NOK 215,289), I-MORE generated fewer QALYs than O-ACT. In this scenario, once we accounted for production loss I-MORE became strongly dominated by O-ACT (more costly and less beneficial). More specific model outcomes and the consequences of including production loss are reported in Table 7.

Scenario B (3-state model)

The 3-state model was used to determine the same outcomes of the previous analyses. Over 25 years, and with a healthcare perspective, total cost was NOK 502,108 for O-ACT and NOK 650,661 for I-MORE. The difference in healthcare costs between I-MORE and O-ACT was NOK 148,554 and moderately similar to the incremental cost of the base case scenario (NOK 169,523). However, the gain in QALYs (0.18) was higher than the gain in the base case scenario (0.14), which led to a lower ICER of 822,939 NOK/QALY. Both base case and scenario B accurately reproduce the results of the trial in terms of healthcare consumption costs (Table 7). Similarly to the base-case scenario, accounting for production losses over 25 years, resulted in O-ACT being strongly dominated by I-MORE.

Table 7 Cost-effectiveness results

Scenario	Years	Strategy	Cost ^a	Effect ^b	Incremental		ICER ^c	ICER
					Cost	Effect	(no discount)	(discount)
Trial results (healthcare)	2	O-ACT	91,412					
		I-MORE	229,150		137,738		/	/
Trial results (societal ^d)	2	O-ACT	1,061,251					
		I-MORE	1,030,716		-30,535		/	/
Base-case (healthcare)	2	O-ACT	85,125	1.34				
		I-MORE	241,765	1.35	156,639	0.010	16,413,333	/
	25	O-ACT	525,215 (406,157;655,373)	9.94 (9.59;10.28)				
		I-MORE	694,738 (572,449;852,672)	10.09 (9.77;10.44)	169,523	0.145	702,774	1,167,887
Base-case (societal)	2	O-ACT	947,580	1.34				
		I-MORE	974,557	1.35	26,977	0.010	2,787,119	/
	25	O-ACT	7,190,277 (6,360,162;7,987,552)	9.95 (9.57;10.26)				
		I-MORE	6,899,702 (5,945,800;7,924,234)	10.1 (9.77;10.36)	-290,575	0.145	Dominant	Dominant
Scenario A ^e (healthcare)	2	O-ACT	86,531	1.34				
		I-MORE	242,980	1.35	156,449	0.010	15,085,220	/
	25	O-ACT	444,683 (336,532;562,727)	10.23 (9.79;10.56)				
		I-MORE	656,973 (544,581;788,168)	10.2 (9.84;10.47)	215,289	-0.033	Strongly dominated	Strongly dominated
Scenario A (societal)	2	O-ACT	983,771	1.34				
		I-MORE	1,003,914	1.35	20,143	0.010	1,942,267	/
	25	O-ACT	5,345,655 (4,109,247;6,606,370)	10.23 (9.79;10.56)				
		I-MORE	609,630 (4,859,532;7,410,113)	10.2 (9.84;10.47)	750,648	-0.033	Strongly dominated	Strongly dominated
Scenario B ^f (healthcare)	2	O-ACT	87,155	1.329				
		I-MORE	242,859	1.344	155,705	0.016	10,076,730	/
	25	O-ACT	502,108 (390,780;639,077)	10.07 (9.76;10.32)				
		I-MORE	650,661 (538,553;777,307)	10.25 (10.01;10.47)	148,554	0.181	467,216	822,939
Scenario B (societal)	2	O-ACT	1,018,797	1.329				
		I-MORE	1,035,523	1.344	16,726	0.016	1,082,442	/
	25	O-ACT	6,487,137 (5,690,660;7,208,566)	10.06 (9.74;10.33)				
		I-MORE	5,800,528 (5,141,959;6,466,361)	10.25 (10.0;10.5)	-686,609	0.190	Dominant	Dominant

Notes:

All values with a 25-year time horizon are reported discounted

All values with a 2-year time horizon are reported undiscounted

a) Costs are reported in NOK 2023; for the 25-year time horizon 95% credible intervals are reported

b) Effects are reported in QALYs; for the 25-year time horizon 95% credible intervals are reported

c) ICERs are reported in NOK 2023 per QALY gained

d) Limited societal perspective (does not include time and transport costs)

e) Piecewise constant transition probabilities

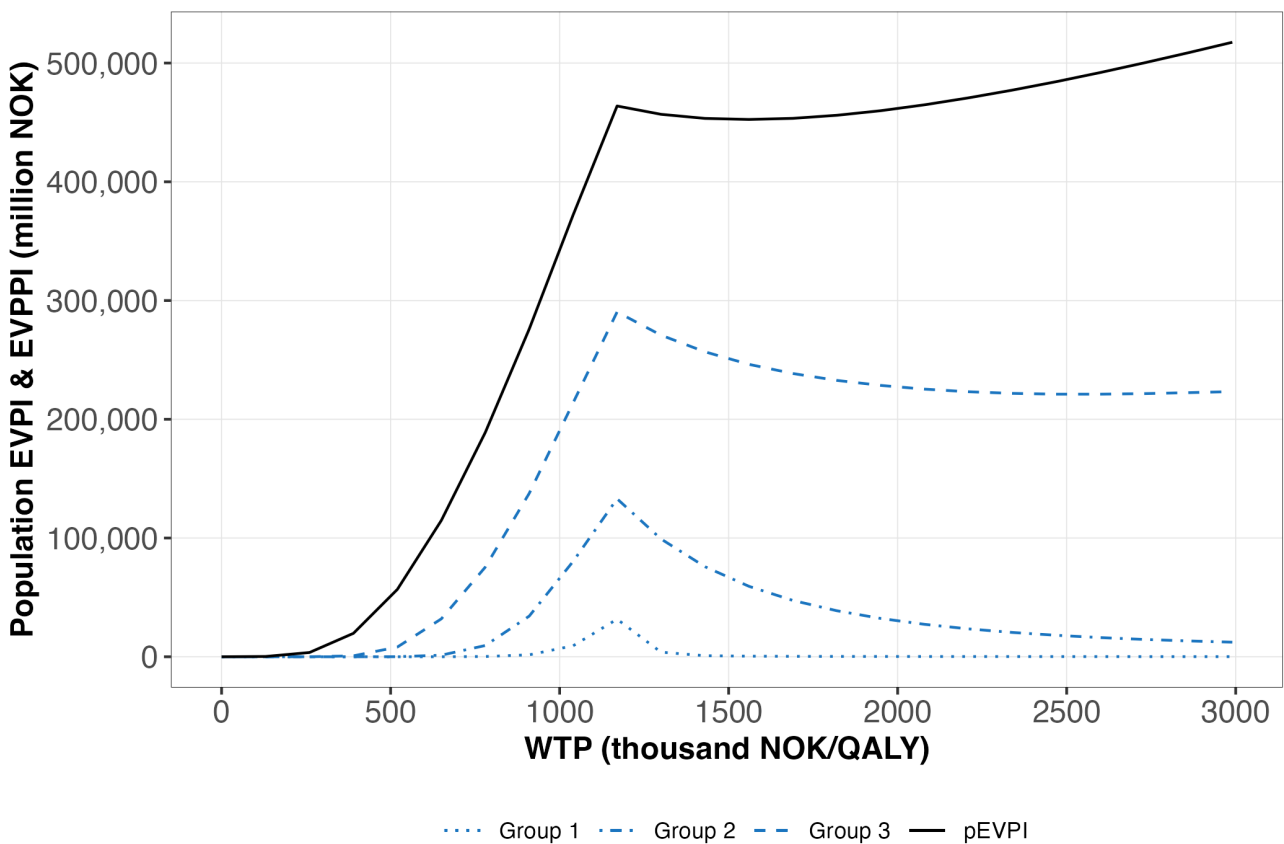
f) 3-state model

5.2.4 Value of information

We performed a value of information analysis for the base-case scenario (with a healthcare perspective). At a WTP threshold of NOK 500,000, the value of eliminating all parameter uncertainty was NOK 8,600 per patient. At that same WTP threshold, pEVPI was 57 billion NOK. The peak corresponded to the point where the optimal strategy, in terms of probability of being cost effective, switched from O-ACT to I-MORE (ICER equal to WTP). At the lower threshold of NOK 275,000, pEVPI went down to 36 billion NOK.

Figure 11 presents the pEVPI and pEVPI for the different groups of parameters at different WTP thresholds. Among the different groups of parameters studied, Group 1, which focused on cost parameters, had the lowest pEVPI. At a WTP of NOK 500,000, it was virtually 0. With that same threshold, Group 2 (state HRQoL) reached a value of 55 million NOK. Finally, Group 3, which investigated the transitions parameters associated with I-MORE (values with high uncertainties), had the highest pEVPI of 8 billion NOK. Following these results, further studies on costs would not be recommended. However, research on HRQoL and possibly a new trial, could be further explored depending on their costs.

Figure 11 Population EVPI and EVPI for each proposed group of parameters



6 Discussion and limitations

In the base-case scenario when we assumed only healthcare costs and a 2-year perspective, I-MORE was not likely to be considered cost-effective when compared to O-ACT (ICER: 1,167,887 NOK/QALY). However, by adopting a broader societal perspective and a longer time horizon, I-MORE reduced production loss and not only became cost-effective but also dominated O-ACT.

The ICER resulting from the probabilistic analysis was considerably higher than the one generated in the initial deterministic approach (1,167,887 NOK/QALY vs 870,996 NOK/QALY). The difference can be attributed to the method used to generate confidence intervals for transition probabilities. In our probabilistic analysis we used bootstrapped confidence intervals to inform the beta distributions of the transition parameters. However, if we were to use normal confidence intervals, the variation around the probability estimates would change and with them the sampled values. Running the probabilistic analysis (10,000 samples) using beta distributions informed with normal confidence intervals led to an incremental effect of 0.201 QALYs and an overall probabilistic ICER of 874,955 NOK/QALY which is in line with the deterministic results. Although bootstrapped confidence intervals might be less precise for events with few observations, they are usually expected to be more accurate than normal confidence intervals (Jackson, 2011). Given the impact of this choice on the ICER, it is important to carefully evaluate the reliability of the bootstrap method and be transparent on its repercussions. In scenario B, where more observations were available, bootstrapped and normal confidence intervals led to similar results, also in line with their deterministic counterpart.

When compared to the base-case analysis, scenario B was better aligned with the results of the trial. If our choice were to be based on this criterion alone, we should focus on the results of scenario B. However, as we presented, the clear distinction between sick leave, work assessment allowance, and the paths that workers undergo before (possibly) reaching a disability pension, make the base case a scenario an option worth considering. Even after accounting for production loss, neither base case nor scenario B turned cost saving with a 2-year time horizon. We attributed this deviation from the trial's results to how sensitive our model is, in the short term, to the initial proportion of patients in each state. As per study protocol, patients had to be sick listed for 2 months prior to inclusion. An analysis of the trial's data revealed that based on the states we defined, not all patients belonged to the SL state at time zero. Once we accounted for slightly different initial proportions (WK = 0.03, GB=0.9, DB = 0.07) scenario B became cost saving even in the short term. After the correction, the base-case incremental costs also decreased but did not turn negative. However, once we increased the time horizon to 3 years, the base-case scenario also became cost saving.

Panel data posed a further limitation. Having access to time-to-event data would have expanded our modelling options, including the possibility of fitting additional continuous probability

distributions such as Weibull, Lognormal, Logistic, etc. While *msm* can be adapted to use panel data, the limitation imposed by the exponential distribution (Markov assumption) should be carefully considered. Although fitting a piecewise constant model led to a better fit, previous research that focused on piecewise constant exponential models used datasets with more than 30,000 individuals and 10 years of observations (Kunst et al., 2020). Extending 24 months of observations (with two cut-offs) to a time horizon of 25 years (300 months) relies on a great number of assumptions and uncertainty. Specifically, regarding the consistency of trends over time. In scenario A transition probabilities from 11 to 300 months are based on a subset of the initial dataset (11 to 24 months). Considering the already limited number of transitions, especially to the DB state, outcomes for the 25 years of scenario A should be interpreted with caution. Given the better fit to the trial's data, the first 2 years of the cost-effectiveness analysis could be modelled with parameters from scenario A. Then, the remaining 23 years using the parameters of the base case model. However, given that the 2-year outcomes between the base case setting and scenario A are almost identical (NOK 156,639 base case incremental costs vs NOK 156,449 scenario A incremental costs). We expect the results of this mixed approach to be virtually the same as the ones of the 25-year base-case scenario. Determining the appropriate model/scenario to rely on is a complex task. We performed several tests (AIC and likelihood) but in the end these approaches only bring value for the time frame in which observations are available. The combination of likelihood tests, AIC, visual inspections, and comparisons with the trial outcomes suggested a preference for the base-case scenario and scenario B.

In 2011, Squires et al. developed a Markov model to extrapolate data beyond the trial's follow up and assess the cost effectiveness of three RTW interventions. HRQoL values were elicited with a different questionnaire (SF-6D) and reflected the preferences of the British population. Nevertheless, the scores for WK (0.76) and SL (0.61) aligned with the ones used in our study. Due to the design of their study, HRQoL estimates were limited to WK and SL.

Previous research already focused on the improved benefits of ACT in comparison to no treatment and treatment as usual (Finnes et al., 2022). However, ACT was implemented in the context of workers on sick leave due to mental health disorders and no physical component was included in either the intervention or the comparator. Cullen et al. (2017) carried out a more comprehensive review that included both musculoskeletal and mental health disorders. From their analysis, multi-domain interventions reduced lost time associated with musculoskeletal and pain-related conditions. Likewise, cognitive behavioural therapy (from which ACT stems) reduced lost time at work and costs associated with mental health disorders. The effects were positive only when CBT was work-focused, as no significant effect was registered in traditional CBT.

A randomized controlled trial in Germany analysed the effects of an inpatient multidisciplinary intervention on employment after sick leave. Patients in the intervention group had 3.5 times higher odds of stable employment (employment with at most 6 months of leave after the intervention). Results are interesting as both the intervention and the comparator consisted of inpatient programs, with the intervention directly focusing on work demands and abilities. Specific components of the interventions are comparable to activities carried out in I-MORE. Although stable employment increased and was significantly different, secondary outcomes such as duration of sick leave, and employment rate were not statistically significant between the interventions (Streibelt & Bethge, 2014).

Our analysis did not address any treatment as usual comparator. As we have seen, due to the great variety of disorders that qualifies for medical benefits, it is hard to identify a single comparator and even harder to define a standard treatment. A meta-analysis of RTW interventions identified visits to general practitioners and prescription of analgesics as an often-chosen comparator for RTW interventions focusing on musculoskeletal disorders (Franche et al., 2005). However, trial-specific interventions (with and without work components) were also used as comparators. Preferences in terms of study design hindered the possibility to have consistent treatment as usual comparisons (Williams et al., 2007). Nevertheless, these studies highlighted that cognitive behavioural therapy with work-related components generates better return-to-work outcomes than usual care in patients with mental health disorders, and that interventions with work-related exercises are more effective than usual care in patients with musculoskeletal pain. I-MORE included similar components which suggest that, at least in terms of effectiveness, it may lead to better outcomes than most standard approaches. O-ACT missed relevant components on work-related exercise, however, the focus on workplace topics (e.g., two individual sessions with a social worker experienced in occupational rehabilitation) suggest that O-ACT could also be more effective than standard practice in dealing with sickness absence due to psychological conditions.

To confirm these hypotheses, the next step would be to perform a subgroup analysis and determine whether a limited use criterion exists. Indeed, certain subgroups (e.g., by diagnosis or education level) might respond better to the interventions. With that in mind, I-MORE could be targeted only to those patients diagnosed with musculoskeletal disorders or psychological disorders and possibly lead to improved cost-effectiveness results. To reach statistical significance, research guidelines recommend having at least 90 patients per treatment arm (up to 100 considering dropout rates) (Sakpal, 2010). Our dataset had observations for 159 individuals, with a considerable share of missing data. Therefore, we were not able to subset patients into the three diagnosis classifications included in the trial (“L”, “P”, “A”).

The challenges posed by the sample size were not limited to the impossibility of performing a subgroup analysis. Our dataset, with 24 monthly observations, was not large enough to fully capture the transitions to the absorbing disability benefit state. Over the two years of follow-up, only seven transitions to the DB state were recorded (2 in I-MORE and 5 in O-ACT). One initial approach to expand the number of observations was to define a state (additional disability) to which patients transitioned to every time an increase in disability percentage was registered. This allowed us to observe nine transitions. We ultimately decided against using this state as it would have not been consistent with the regulatory framework and would have posed challenges in defining clear state-related costs and HRQoL.

Limited information also affected our choices regarding the 15D instrument. In the analysis, HRQoL values were estimated using imputed data from a linear regression. Still, only using the data available from the trial generated similar estimates. The main difference regarded the WAA state which was associated with a higher HRQoL (0.647) than the one used in our model (0.628). Once we incorporated this new estimate, we registered a slight increase in incremental benefits 0.153 (compared to the previous 0.145) over 25 years. This new value generated a lower ICER of 1,109,622 NOK per QALY gained.

In our VOI calculations we adjusted the estimates for a population that would benefit from the intervention. However, the uncertainty around prevalence and incidence for the two groups of disorders led to several assumptions. From the published literature we were able to elicit prevalence and incidence values that referred to the general working population but not values that took into account the smaller population of sick listed workers. Indeed, the trial protocol, in addition to a “L” or “P” diagnosis (ICPC2), required that patients had been sick listed for at least 2 months. Although patients with long term musculoskeletal and psychological disorders are likely to end up being sick listed, not all of them will. That is why, it is likely that the true pEVPI and pEVPPI are lower than the values presented in the results. Further research is needed to generate more specific values, which also account for the high levels of comorbidity between the two groups of disorders. Only then meaningful VOI conclusions can be made.

To preserve the comparability with the trial’s result, our analysis did not include transportation costs. However, Norwegian guidelines recommend the inclusion of transportation costs related to travelling to and from treatment. The Hysnes facility was located a 1-hour drive from the city of Trondheim. On the one hand, due to its inpatient design, transportation costs for I-MORE would be fairly straightforward to compute, consisting of a 2-hour return trip to the site. On the other hand, O-ACT lasted 6 weeks with a weekly meeting at the St. Olavs Hospital. The hospital is located in Trondheim’s city centre, assuming that all patients lived in the city, travel costs over the 6 weeks should be comparable to the 2-hour return trip to the Hysnes facility. Transportation costs would also affect the way we computed healthcare consumption. Building on the results of

the trial, we generated state costs for our model that accounted only for primary and secondary care consumption. It is hard to speculate on the consequences of including transportation costs for each cost item in these categories as all the values would increase depending both on where the service was provided and where the patient lived.

In the recommended extended healthcare perspective, the patient's use of time in connection with treatment should be taken into account (Norwegian Medicines Agency, 2018). The time spent at the Hysnes facility (3.5 weeks) was indeed more substantial than the time required to complete the O-ACT program (≈ 20 hours). However, keeping in mind a 25-year time horizon these differences would most likely not affect the results in any significant way. Once we considered the impact of production loss, the inclusion of patients' time dedicated to both treatment and primary/secondary care use, could potentially result in double counting costs. Specifically, we would not be able to distinguish whether the patient dedicated time to primary/secondary care during working hours (included in production loss) or during personal free time (not captured by production loss). In order to keep the model simple (but no simpler), we decided to omit this cost category.

In Norway, there is an ongoing debate regarding the adoption of a societal perspective in the assessment of health technologies. Indeed, besides considering a healthcare perspective, we also carried out the analysis of each scenario with a limited societal perspective (production loss). Undoubtedly, the results of adopting such a perspective heavily depend on the chosen cost for production loss and on the absence associated with each state. The estimation of production loss was based on monthly earnings as reported by Statistics Norway. Like in the economic evaluation of the trial, we adjusted the value to account for social costs ($\approx 40\%$) however different values are also recommended (e.g., $\approx 25\%$) (Norwegian Medicines Agency, 2012). Still, even using 25% social costs, I-MORE was cost saving (only with a 25-year time horizon). Production loss should be only linked to working days (≈ 21 days in a month). However, our absence data did not distinguish between workdays and weekends, this is possibly due to the way registries keep track of sick leave or to how data were manipulated during the trial. When a patient was on sick-leave and absent for 31 days we could easily link production loss to the 21 working days in that month. The problem occurred when the patient was not absent for the full month, as there was no way of knowing whether that number of days referred to actual workdays missed or to weekends. To partly correct for this limitation, daily loss due to absence was computed by dividing monthly earnings (social costs included) by 30.4. Again, more precise time-to-event data could improve the robustness of our results. Finally, we used the average number of absent days in a month (over 24 months) multiplied by daily loss to link each state with a value for production loss. Although we stratified by intervention to capture the better effects of I-MORE, a more precise analysis could use a time-dependent estimate of absence to incorporate production loss in a more realistic way.

7 Conclusion

In our model-based analysis, we determined that for most scenarios I-MORE led to greater benefits than O-ACT in the short and long term. However, when considering the Norwegian reference threshold of NOK 500,000, the corresponding increase in costs outweighed the benefits. Over a 25-year time horizon and with a healthcare cost perspective, only one scenario, in which we pooled data into 3 states, was cost-effective given the higher threshold of NOK 825,000. Depending on choices regarding the underlying probabilistic distributions, the base-case scenario could also be considered cost-effective under the higher threshold.

When we considered a limited societal perspective, with a proxy for production loss, I-MORE strongly dominated O-ACT. According to our model, the inclusion of production loss, made I-MORE the dominant strategy after 3 years in the base-case scenario, and after 2 years in the 3-state model structure. Implementing, to some degree, time-dependency improved the goodness-of-fit of the model. However, such a choice resulted in an inversion of the results where I-MORE was strongly dominated by O-ACT. Nevertheless, the many data-related limitations of this scenario questioned its reliability.

These findings highlighted the key roles of both the time horizon and the cost perspective in economic evaluations. Possibly prompting the debate on the inclusion of a broader societal perspective in the Norwegian guidelines. This was the first study that extrapolated data from a trial and developed a model to simulate the costs and effects of an inpatient return-to-work intervention in Norway.

8 Reference list

- Aasdahl, L., Fimland, M. S., Bjørnelv, G. M. W., Gismervik, S. Ø., Johnsen, R., Vasseljen, O., & Halsteinli, V. (2023). Economic Evaluation of Inpatient Multimodal Occupational Rehabilitation vs. Outpatient Acceptance and Commitment Therapy for Sick-Listed Workers with Musculoskeletal- or Common Mental Disorders. *Journal of Occupational Rehabilitation*. <https://doi.org/10.1007/s10926-022-10085-0>
- Aasdahl, L., Vasseljen, O., Gismervik, S. Ø., Johnsen, R., & Fimland, M. S. (2021). Two-Year Follow-Up of a Randomized Clinical Trial of Inpatient Multimodal Occupational Rehabilitation Vs Outpatient Acceptance and Commitment Therapy for Sick Listed Workers with Musculoskeletal or Common Mental Disorders. *Journal of Occupational Rehabilitation*, 31(4), 721–728. <https://doi.org/10.1007/s10926-021-09969-4>
- Al, M. J. (2012). Cost-Effectiveness Acceptability Curves Revisited. *PharmacoEconomics*, 31(2), 93–100. <https://doi.org/10.1007/s40273-012-0011-8>
- Ammendolia, C., Cassidy, D., Steenstra, I., Soklaridis, S., Boyle, E., Eng, S., Howard, H., Bhupinder, B., & Côté, P. (2009). Designing a workplace return-to-work program for occupational low back pain: an intervention mapping approach. *BMC Musculoskeletal Disorders*, 10(1). <https://doi.org/10.1186/1471-2474-10-65>
- Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J., & Paltiel, A. D. (2012). Model Parameter Estimation and Uncertainty Analysis. *Medical Decision Making*, 32(5), 722–732. <https://doi.org/10.1177/0272989x12458348>
- Briggs, A., Sculpher, M., & Claxton, K. (2011). *Decision modelling for health economic evaluation*. Oxford Oxford Univ. Press.
- Buxton, M. J., Drummond, M. F., Van Hout, B. A., Prince, R. L., Sheldon, T. A., Szucs, T., & Vray, M. (1997). Modelling in economic evaluation: an unavoidable fact of life. *Health Economics*, 6(3), 217–227. [https://doi.org/10.1002/\(sici\)1099-1050\(199705\)6:33.0.co;2-w](https://doi.org/10.1002/(sici)1099-1050(199705)6:33.0.co;2-w)
- Büyükkaramikli, N. C., Rutten-van Mölken, M. P. M. H., Severens, J. L., & Al, M. (2019). TECH-VER: A Verification Checklist to Reduce Errors in Models and Improve Their Credibility. *PharmacoEconomics*, 37(11), 1391–1408. <https://doi.org/10.1007/s40273-019-00844-y>
- Caro, J. J., Briggs, A. H., Siebert, U., & Kuntz, K. M. (2012). Modeling Good Research Practices—Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Value in Health*, 15(6), 796–803. <https://doi.org/10.1016/j.jval.2012.06.012>
- Carroll, C., Rick, J., Pilgrim, H., Cameron, J., & Hillage, J. (2009). Workplace involvement improves return to work rates among employees with back pain on long-term sick leave: a systematic review of the effectiveness and cost-effectiveness of interventions. *Disability and Rehabilitation*, 32(8), 607–621. <https://doi.org/10.3109/09638280903186301>
- Claeskens, G., & Hjort, N. L. (2008). Akaike’s information criterion. In *Model Selection and Model Averaging (Cambridge Series in Statistical and Probabilistic Mathematics)* (pp. 22–69). Cambridge University Press. <https://doi.org/10.1017/CBO9780511790485.003>
- Claxton, K. (1999). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*, 18(3), 341–364. [https://doi.org/10.1016/s0167-6296\(98\)00039-3](https://doi.org/10.1016/s0167-6296(98)00039-3)
- Cooney, G. M., Dwan, K., Greig, C. A., Lawlor, D. A., Rimer, J., Waugh, F. R., McMurdo, M., & Mead, G. E. (2013). Exercise for depression. *Cochrane Database of Systematic Reviews*, 9(9). <https://doi.org/10.1002/14651858.cd004366.pub6>
- Cox, D. R., & Miller, H. D. (1977). *The Theory of Stochastic Processes*. <https://doi.org/10.1201/9780203719152>
- Cullen, K. L., Irvin, E., Collie, A., Clay, F., Gensby, U., Jennings, P. A., Hogg-Johnson, S., Kristman, V., Laberge, M., McKenzie, D., Newnam, S., Palagyi, A., Ruseckaite, R., Sheppard, D. M., Shourie, S., Steenstra, I., Van Eerd, D., & Amick, B. C. (2017). Effectiveness of Workplace Interventions in Return-to-Work for Musculoskeletal, Pain-Related and Mental Health Conditions: An Update of the Evidence and Messages for

- Practitioners. *Journal of Occupational Rehabilitation*, 28(1), 1–15.
<https://doi.org/10.1007/s10926-016-9690-x>
- de Wreede, L. C., Fiocco, M., & Putter, H. (2011). mstate: AnRPackage for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, 38(7).
<https://doi.org/10.18637/jss.v038.i07>
- Dewa, C. S., Hoch, J. S., Loong, D., Trojanowski, L., & Bonato, S. (2020). Evidence for the Cost-Effectiveness of Return-to-Work Interventions for Mental Illness Related Sickness Absences: A Systematic Literature Review. *Journal of Occupational Rehabilitation*.
<https://doi.org/10.1007/s10926-020-09904-z>
- Drummond, M. (2015). *Methods for the economic evaluation of health care programmes*. Oxford University Press.
- E. Baca-García, M. Perez-Rodriguez, I. Basurte-Villamor, Moral, del, M. Jiménez-Arriero, Gonzalez, L., J. Saiz-ruiz, & Oquendo, M. (2021). *Diagnostic stability of psychiatric disorders in clinical practice*. British Journal of Psychiatry.
<https://www.semanticscholar.org/paper/Diagnostic-stability-of-psychiatric-disorders-in-Baca-Garc%C3%ADa-Perez-Rodriguez/1b44c133daa3e1ca320bdbec28386a977e6b2cd0>
- Fenwick, E., Claxton, K., & Sculpher, M. (2001). Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Economics*, 10(8), 779–787.
<https://doi.org/10.1002/hec.635>
- Fenwick, E., Steuten, L., Knies, S., Ghabri, S., Basu, A., Murray, J. F., Koffijberg, H. (Erik), Strong, M., Sanders Schmidler, G. D., & Rothery, C. (2020). Value of Information Analysis for Research Decisions—An Introduction: Report 1 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. *Value in Health*, 23(2), 139–150.
<https://doi.org/10.1016/j.jval.2020.01.001>
- Finnes, A., Hoch, J. S., Enebrink, P., Dahl, J., Ghaderi, A., Nager, A., & Feldman, I. (2022). Economic evaluation of return-to-work interventions for mental disorder-related sickness absence: two years follow-up of a randomized clinical trial. *Scandinavian Journal of Work, Environment & Health*, 48(4), 264–272. <https://doi.org/10.5271/sjweh.4012>
- Fleurence, R. L., & Hollenbeak, C. S. (2007). Rates and Probabilities in Economic Modelling. *PharmacoEconomics*, 25(1), 3–6. <https://doi.org/10.2165/00019053-200725010-00002>
- Franche, R.-L., Cullen, K., Clarke, J., Irvin, E., Sinclair, S., & Frank, J. (2005). Workplace-Based Return-to-Work Interventions: A Systematic Review of the Quantitative Literature. *Journal of Occupational Rehabilitation*, 15(4), 607–631. <https://doi.org/10.1007/s10926-005-8038-8>
- Garrison, L. P., Mansley, E. C., Abbott, T. A., Bresnahan, B. W., Hay, J. W., & Smeeding, J. (2010). Good Research Practices for Measuring Drug Costs in Cost-Effectiveness Analyses: A Societal Perspective: The ISPOR Drug Cost Task Force Report—Part II. *Value in Health*, 13(1), 8–13. <https://doi.org/10.1111/j.1524-4733.2009.00660.x>
- Gismervik, S. Ø., Aasdahl, L., Vasseljen, O., Fors, E. A., Rise, M. B., Johnsen, R., Hara, K., Jacobsen, H. B., Pape, K., Fleten, N., Jensen, C., & Fimland, M. S. (2020). Inpatient multimodal occupational rehabilitation reduces sickness absence among individuals with musculoskeletal and common mental health disorders: a randomized clinical trial. *Scandinavian Journal of Work, Environment & Health*, 46(4), 364–372.
<https://doi.org/10.5271/sjweh.3882>
- Gran, J. M., Lie, S. A., Øyeflaten, I., Borgan, Ø., & Aalen, O. O. (2015). Causal inference in multi-state models—sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*, 15(1). <https://doi.org/10.1186/s12889-015-2408-8>
- Hagen, K., Linde, M., Heuch, I., Stovner, L. J., & Zwart, J.-A. (2011). Increasing Prevalence of Chronic Musculoskeletal Complaints. A Large 11-Year Follow-Up in the General Population (HUNT 2 and 3). *Pain Medicine*, 12(11), 1657–1666. <https://doi.org/10.1111/j.1526-4637.2011.01240.x>
- Hagen, K., Svebak, S., & Zwart, J.-A. (2006). Incidence of Musculoskeletal Complaints in a Large Adult Norwegian County Population. The HUNT Study. *Spine*, 31(18), 2146.
<https://doi.org/10.1097/01.brs.0000231734.56161.6b>

- Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, 35(4), 639–665. [https://doi.org/10.1016/s0005-7894\(04\)80013-3](https://doi.org/10.1016/s0005-7894(04)80013-3)
- Hoefsmit, N., Houkes, I., & Nijhuis, F. J. N. (2012). Intervention Characteristics that Facilitate Return to Work After Sickness Absence: A Systematic Literature Review. *Journal of Occupational Rehabilitation*, 22(4), 462–477. <https://doi.org/10.1007/s10926-012-9359-z>
- Incerti, D., & Jansen, J. P. (2021, February 18). *hesim: Health Economic Simulation Modeling and Decision Analysis*. ArXiv.org. <https://arxiv.org/abs/2102.09437v2>
- Jackson, C. H. (2011). Multi-State Models for Panel Data: ThemsmPackage for R. *Journal of Statistical Software*, 38(8). <https://doi.org/10.18637/jss.v038.i08>
- Kaplan, R. S., Anderson, S. R., & Harvard Graduate School Of Business Administration. (2009). *Time-driven activity-based costing : a simpler and more powerful path to higher profits*. Boston, Mass. Harvard Business School Press.
- Kausto, J., Miranda, H., Martimo, K.-P., & Viikari-Juntura, E. (2008). Partial sick leave—review of its use, effects and feasibility in the Nordic countries. *Scandinavian Journal of Work, Environment & Health*, 34(4), 239–249. <https://www.jstor.org/stable/40967715>
- Kim, D. D., Silver, M. C., Kunst, N., Cohen, J. T., Ollendorf, D. A., & Neumann, P. J. (2020). Perspective and Costing in Cost-Effectiveness Analysis, 1974–2018. *PharmacoEconomics*, 38(10), 1135–1145. <https://doi.org/10.1007/s40273-020-00942-2>
- Kinge, J. M., Knudsen, A. K., Skirbekk, V., & Vollset, S. E. (2015). Musculoskeletal disorders in Norway: prevalence of chronicity and use of primary and specialist health care services. *BMC Musculoskeletal Disorders*, 16(1). <https://doi.org/10.1186/s12891-015-0536-z>
- Krause, N., Dasinger, L. K., & Neuhauser, F. (1998). Modified work and return to work: a review of the literature. *Journal of Occupational Rehabilitation*, 8(2), 113–139. <https://doi.org/10.1023/a:1023015622987>
- Kunst, N., Alarid-Escudero, F., Aas, E., Coupé, V. M. H., Schrag, D., & Kuntz, K. M. (2020). Estimating Population-Based Recurrence Rates of Colorectal Cancer over Time in the United States. *Cancer Epidemiology, Biomarkers & Prevention*, 29(12), 2710–2718. <https://doi.org/10.1158/1055-9965.EPI-20-0490>
- L Robbins. (1935). *An Essay on the nature and significance of economic science*. (p. 15). Macmillan.
- Michel, Y., Liv Berit Augestad, Barra, M., & Rand-Hendriksen, K. (2019). *A Norwegian 15D value algorithm: proposing a new procedure to estimate 15D value algorithms*. 28(5), 1129–1143. <https://doi.org/10.1007/s11136-018-2043-9>
- Moens, M. (2022). *Personalised rehabilitation to improve return to work in patients with persistent spinal pain syndrome type II after spinal cord stimulation implantation: a study protocol for a 12-month randomised controlled trial—the OPERA study*. Springermedizin.de. <https://www.springermedizin.de/personalised-rehabilitation-to-improve-return-to-work-in-patient/23790628>
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90–100. [https://doi.org/10.1016/s0022-2496\(02\)00028-7](https://doi.org/10.1016/s0022-2496(02)00028-7)
- National Institute for Health and Care Excellence. (2022). Depression in adults: treatment and management. In *PubMed*. National Institute for Health and Care Excellence (NICE). <https://www.ncbi.nlm.nih.gov/books/NBK583074/>
- NAV. (2020, December 8). *Sickness benefit (sykepenger) for employees*. Nav.no. <https://www.nav.no/en/home/benefits-and-services/Sickness-benefit-for-employees#chapter-8>
- Norwegian Institute of Public Health. (2021). *Guidelines for the submission of documentation for single technology assessments (STAs) of medical devices and diagnostic interventions*. <https://www.fhi.no/en/qk/HTA/brukermedvirkning-metodevurderinger/submission-sta/>
- Norwegian Medicines Agency. (2012). *Dokumentasjon av enhetskostnader Versjon 1.2*. <https://legemiddelverket.no/Documents/Offentlig%20finansiering%20og%20pris/Dokumentasjon%20til%20metodevurdering/Dokumentasjon%20av%20enhetskostnader%20V1.2.pdf>

- Norwegian Medicines Agency. (2018). *Guidelines for the submission of documentation for single technology assessment (STA) of pharmaceuticals*.
<https://legemiddelverket.no/Documents/English/Public%20funding%20and%20pricing/Documentation%20for%20STA/Guidelines%20151018.pdf>
- Norwegian Ministry of Health and Care Services. (2015, November 4). *På ramme alvor*.
 Regjeringen.no. <https://www.regjeringen.no/no/dokumenter/pa-ramme-alvor/id2460080/>
- Norwegian Ministry of Health and Care Services. (2017). *Principles for priority setting in health care*. <https://www.regjeringen.no/contentassets/439a420e01914a18b21f351143ccc6af/en-gb/pdfs/stm201520160034000engpdfs.pdf>
- NTNU. (2019). *HUNT Databank*. Hunt-Db.medisin.ntnu.no. <https://hunt-db.medisin.ntnu.no/hunt-db/>
- Nystuen, P., Hagen, K. B., & Herrin, J. (2001). Mental health problems as a cause of long-term sick leave in the Norwegian workforce. *Scandinavian Journal of Public Health*, 29(3), 175–182.
<https://pubmed.ncbi.nlm.nih.gov/11680768/>
- Øyeflaten, I., Lie, S. A., Ihlebæk, C. M., & Eriksen, H. R. (2012). Multiple transitions in sick leave, disability benefits, and return to work. - A 4-year follow-up of patients participating in a work-related rehabilitation program. *BMC Public Health*, 12(1).
<https://doi.org/10.1186/1471-2458-12-748>
- Petrou, S., & Gray, A. (2011). Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ*, 342(apr11 1), d1766–d1766.
<https://doi.org/10.1136/bmj.d1766>
- Powers, M. B., Zum Vörde Sive Vörding, M. B., & Emmelkamp, P. M. G. (2009). Acceptance and Commitment Therapy: A Meta-Analytic Review. *Psychotherapy and Psychosomatics*, 78(2), 73–80. <https://doi.org/10.1159/000190790>
- Raftery, J. P. (2008). Paying for costly pharmaceuticals: regulation of new drugs in Australia, England and New Zealand. *Medical Journal of Australia*, 188(1), 26–28.
<https://doi.org/10.5694/j.1326-5377.2008.tb01500.x>
- Reme, S. E., Tangen, T., Moe, T., & Eriksen, H. R. (2011). Prevalence of psychiatric disorders in sick listed chronic low back pain patients. *European Journal of Pain*, 15(10), 1075–1080.
<https://doi.org/10.1016/j.ejpain.2011.04.012>
- Rothery, C., Strong, M., Koffijberg, H. (Erik), Basu, A., Ghabri, S., Knies, S., Murray, J. F., Sanders Schmidler, G. D., Steuten, L., & Fenwick, E. (2020). Value of Information Analytical Methods: Report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. *Value in Health*, 23(3), 277–286. <https://doi.org/10.1016/j.jval.2020.01.004>
- Sakpal, T. V. (2010). Sample size estimation in clinical trial. *Perspectives in Clinical Research*, 1(2), 67–69. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148614/>
- Services, M. of H. and C. (2016, October 20). *Severity of illness and priority setting in Norway*.
 Government.no. <https://www.regjeringen.no/en/dokumenter/severity-of-illness-and-priority-setting-in-norway/id2460080/>
- Shand, F. L., Ridani, R., Tighe, J., & Christensen, H. (2013). The effectiveness of a suicide prevention app for indigenous Australian youths: study protocol for a randomized controlled trial. *Trials*, 14(1), 396. <https://doi.org/10.1186/1745-6215-14-396>
- Simoens, S. (2012). The cost-effectiveness of prevention: is an ounce of prevention worth a pound of cure? *Farmeconomia. Health Economics and Therapeutic Pathways*, 13(1), 5–6.
<http://journals.seedmedicalpublishers.com/index.php/FE/article/view/197/182>
- Sintonen, H. (2001). The 15D instrument of health-related quality of life: properties and applications. *Annals of Medicine*, 33(5), 328–336. <https://doi.org/10.3109/07853890109002086>
- Spitzer, R. L., Kroenke, K., Linzer, M., Hahn, S. R., Williams, J. B. W., deGruy, F. V., Brody, D., & Davies, M. (1995). Health-Related Quality of Life in Primary Care Patients With Mental Disorders. *JAMA*, 274(19), 1511. <https://doi.org/10.1001/jama.1995.03530190025030>
- Squires, H., Rick, J., Carroll, C., & Hillage, J. (2011). Cost-effectiveness of interventions to return employees to work following long-term sickness absence due to musculoskeletal disorders. *Journal of Public Health*, 34(1), 115–124. <https://doi.org/10.1093/pubmed/fdr057>

- St. Olavs Hospital. (2016). *Hysnes Helsefort*. St. Olavs Hospital. <https://stolav.no/hysnes-helsefort>
- Statistics Norway. (2023, April 11). *Consumer price index*. SSB. <https://www.ssb.no/en/priser-og-prisindekser/konsumpriser/statistikk/konsumprisindeksen>
- Storeng, S. H., Sund, E. R., & Krokstad, S. (2020). Prevalence, clustering and combined effects of lifestyle behaviours and their association with health after retirement age in a prospective cohort study, the Nord-Trøndelag Health Study, Norway. *BMC Public Health*, 20(1). <https://doi.org/10.1186/s12889-020-08993-y>
- Streibelt, M., & Bethge, M. (2014). Effects of intensified work-related multidisciplinary rehabilitation on occupational participation. *International Journal of Rehabilitation Research*, 37(1), 61–66. <https://doi.org/10.1097/mrr.0000000000000031>
- Strong, M., Oakley, J. E., & Brennan, A. (2013). Estimating Multiparameter Partial Expected Value of Perfect Information from a Probabilistic Sensitivity Analysis Sample. *Medical Decision Making*, 34(3), 311–326. <https://doi.org/10.1177/0272989x13505910>
- Svebak, S., Hagen, K., & Zwart, J.-A. (2006). One-Year Prevalence of Chronic Musculoskeletal Pain in a Large Adult Norwegian County Population: Relations with Age and Gender—The HUNT Study. *Journal of Musculoskeletal Pain*, 14(1), 21–28. https://doi.org/10.1300/j094v14n01_04
- van Delden, J. J. M. (2004). Medical decision making in scarcity situations. *Journal of Medical Ethics*, 30(2), 207–211. <https://doi.org/10.1136/jme.2003.003681>
- Vemer, P., Corro Ramos, I., van Voorn, G. A. K., Al, M. J., & Feenstra, T. L. (2015). AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. *PharmacoEconomics*, 34(4), 349–361. <https://doi.org/10.1007/s40273-015-0327-2>
- Versteegh, M., Knies, S., & Brouwer, W. (2016). From Good to Better: New Dutch Guidelines for Economic Evaluations in Healthcare. *PharmacoEconomics*, 34(11), 1071–1074. <https://doi.org/10.1007/s40273-016-0431-y>
- Werner, E. L., & Côté, P. (2009). Low back pain and determinants of sickness absence. *European Journal of General Practice*, 15(2), 74–79. <https://doi.org/10.1080/13814780903051866>
- WHO. (2021, February 8). *Musculoskeletal conditions*. Who.int; World Health Organization: WHO. <https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions>
- Williams, R. M., Westmorland, M. G., Lin, C. A., Schmuck, G., & Creen, M. (2007). Effectiveness of workplace rehabilitation interventions in the treatment of work-related low back pain: A systematic review. *Disability and Rehabilitation*, 29(8), 607–624. <https://doi.org/10.1080/09638280600841513>

Appendix A: Tables & Figures

Table 1 T-test for differences in total healthcare consumption between years 1-2

Two-sample t test with equal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
1	1,908	2544.775	449.7751	19646.47	1662.672	3426.878
2	1,908	1961.301	225.2686	9839.882	1519.502	2403.1
Combined	3,816	2253.038	251.5286	15537.87	1759.895	2746.182
diff		583.4745	503.0344		-402.7678	1569.717

diff = mean(1) - mean(2)

t = 1.1599

H0: diff = 0

Degrees of freedom = 3814

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

Pr(T < t) = 0.8769

Pr(|T| > |t|) = 0.2462

Pr(T > t) = 0.1231

Table 2 T-test for differences in total healthcare consumption (ln) between years 1-2

Two-sample t test with equal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
1	1,361	6.754744	.0453273	1.672204	6.665825	6.843663
2	1,158	6.69757	.048754	1.659068	6.601914	6.793226
Combined	2,519	6.728461	.0331959	1.666092	6.663367	6.793555
diff		.0571738	.066612		-.0734461	.1877937

diff = mean(1) - mean(2)

t = 0.8583

H0: diff = 0

Degrees of freedom = 2517

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

Pr(T < t) = 0.8046

Pr(|T| > |t|) = 0.3908

Pr(T > t) = 0.1954

Table 3 State-costs estimates (primary and secondary care) using random effects, fixed effects, and GLM regressions.

State	Random effects		Fixed effects		GLM	
	Cost (NOK)	SE	Cost (NOK)	SE	Cost (NOK)	SE
WK	785	198	562	539	771	86
SL	3,229	665	2,852	643	3,258	661
WAA	2,922	496	2,566	599	2,933	492
DB	1,852	480	5,647	1,736	1,770	375

Notes: GLM (family:gamma, link:log)

WK: work

SL: sick leave

WAA: work assessment allowance

DB: disability benefit

Table 4 Input parameters for scenario A

Intervention	Time (months)	Input parameter	Mean	SE ^a	Distribution ^b	Type of parameter	
O-ACT	from 0 to 5	WK-SL	0.08620	0.0363	Beta	Probability	
		WK-WAA	0.04060	0.0297	Beta	Probability	
		SL-WK	0.90600	0.0110	Beta	Probability	
		SL-WAA	0.04550	0.0095	Beta	Probability	
		SL-DB	0.00002	0.0000	Beta	Probability	
		WAA-WK	0.00368	0.0011	Beta	Probability	
		WAA-SL	0.00018	0.0001	Beta	Probability	
		WAA-DB	0.00098	0.0002	Beta	Probability	
	from 5 to 11	WK-SL	0.02990	0.0123	Beta	Probability	
		WK-WAA	0.02090	0.0127	Beta	Probability	
		SL-WK	0.66200	0.0313	Beta	Probability	
		SL-WAA	0.22000	0.0313	Beta	Probability	
		SL-DB	0.00002	0.0000	Beta	Probability	
		WAA-WK	0.00862	0.0060	Beta	Probability	
		WAA-SL	0.00014	0.0001	Beta	Probability	
		WAA-DB	0.00019	0.0000	Beta	Probability	
	11 on	WK-SL	0.05350	0.0132	Beta	Probability	
		WK-WAA	0.00170	0.0007	Beta	Probability	
		SL-WK	0.72300	0.0558	Beta	Probability	
		SL-WAA	0.03960	0.0210	Beta	Probability	
		SL-DB	0.00018	0.0001	Beta	Probability	
		WAA-WK	0.01690	0.0050	Beta	Probability	
		WAA-SL	0.00050	0.0002	Beta	Probability	
		WAA-DB	0.00835	0.0039	Beta	Probability	
			Treatment cost	11,033		Fixed	Cost
			Production loss SL	49,460		Fixed	Production loss
			Production loss WAA	51,933		Fixed	Production loss
			Production loss DB	32,149		Fixed	Production loss
I-MORE	from 0 to 5	WK-SL	0.12500	0.0442	Beta	Probability	
		WK-WAA	0.00378	0.0015	Beta	Probability	
		SL-WK	0.89100	0.0122	Beta	Probability	
		SL-WAA	0.04110	0.0104	Beta	Probability	
		SL-DB	0.00030	0.0001	Beta	Probability	
		WAA-WK	0.00372	0.0014	Beta	Probability	
		WAA-SL	0.00026	0.0001	Beta	Probability	
		WAA-DB	0.00071	0.0003	Beta	Probability	
	from 5 to 11	WK-SL	0.06610	0.0154	Beta	Probability	
		WK-WAA	0.00619	0.0015	Beta	Probability	

Appendix A: Tables & Figures

	SL-WK	0.66300	0.0289	Beta	Probability
	SL-WAA	0.13600	0.0220	Beta	Probability
	SL-DB	0.00016	0.0001	Beta	Probability
	WAA-WK	0.01850	0.0094	Beta	Probability
	WAA-SL	0.00069	0.0004	Beta	Probability
	WAA-DB	0.00028	0.0001	Beta	Probability
11 on	WK-SL	0.08220	0.0149	Beta	Probability
	WK-WAA	0.00785	0.0041	Beta	Probability
	SL-WK	0.76900	0.0311	Beta	Probability
	SL-WAA	0.02410	0.0118	Beta	Probability
	SL-DB	0.00547	0.0042	Beta	Probability
	WAA-WK	0.01460	0.0058	Beta	Probability
	WAA-SL	0.00067	0.0003	Beta	Probability
	WAA-DB	0.00247	0.0020	Beta	Probability
	Treatment cost	141,455		Fixed	Cost
	Production loss SL	44,514		Fixed	Production loss
	Production loss WAA	51,933		Fixed	Production loss
	Production loss DB	24,730		Fixed	Production loss
	c_WK	785	198	Gamma	State cost
	c_SL	3,229	665	Gamma	State cost
	c_WAA	2,922	496	Gamma	State cost
	c_DB	1,852	480	Gamma	State cost
	u_WK	0.699	0.011	Beta	HRQoL
	u_SL	0.608	0.006	Beta	HRQoL
	u_WAA	0.628	0.009	Beta	HRQoL
	u_DB	0.582	0.023	Beta	HRQoL

Notes:

All costs (and respective SE) are reported in NOK 2016 (In the analysis we accounted for a price inflator value of 23.2%)

Production loss is reported in NOK 2023

a) Standard errors for transition probabilities were computed from 95% CI following Briggs (2011)

a) Standard errors for HRQoL were computed using the Delta method (STATA)

b) "Fixed" parameters were not varied in the probabilistic analysis

WK: work

SL: sick leave

WAA: work assessment allowance

DB: disability benefit

Table 5 Input parameters for scenario B

Intervention	Input parameter	Mean	SE ^a	Distribution ^b	Type of parameter
O-ACT	WK-GB	0.06855953	0.0111	Beta	Probability
	GB-WK	0.0479891	0.0060	Beta	Probability
	GB-DB	0.0038	0.0017	Beta	Probability
	Treatment cost	11,033		Fixed	Cost
	Production loss GB	51,933		Fixed	Production loss
	Production loss DB	32,149		Fixed	Production loss
I-MORE	WK-GB	0.1013563	0.0123	Beta	Probability
	GB-WK	0.0838589	0.0082	Beta	Probability
	GB-DB	0.0017	0.0011	Beta	Probability
	Treatment cost	141,455		Fixed	Cost
	Production loss GB	49,460		Fixed	Production loss
	Production loss DB	22,257		Fixed	Production loss
	c_WK	774	258	Gamma	State cost
	c_GB	3,044	460	Gamma	State cost
	c_DB	1,851	484	Gamma	State cost
	u_WK	0.699	0.00876	Beta	HRQoL
	u_GB	0.617	0.00469	Beta	HRQoL
	u_DB	0.582	0.02041	Beta	HRQoL

Notes:

All costs (and respective SE) are reported in NOK 2016 (In the analysis we accounted for a price inflator value of 23.2%)
Production loss is reported in NOK 2023

a) Standard errors for transition probabilities were computed from 95% CI following Briggs (2011)

a) Standard errors for HRQoL were computed using the Delta method (STATA)

b) "Fixed" parameters were not varied in the probabilistic analysis

WK: work

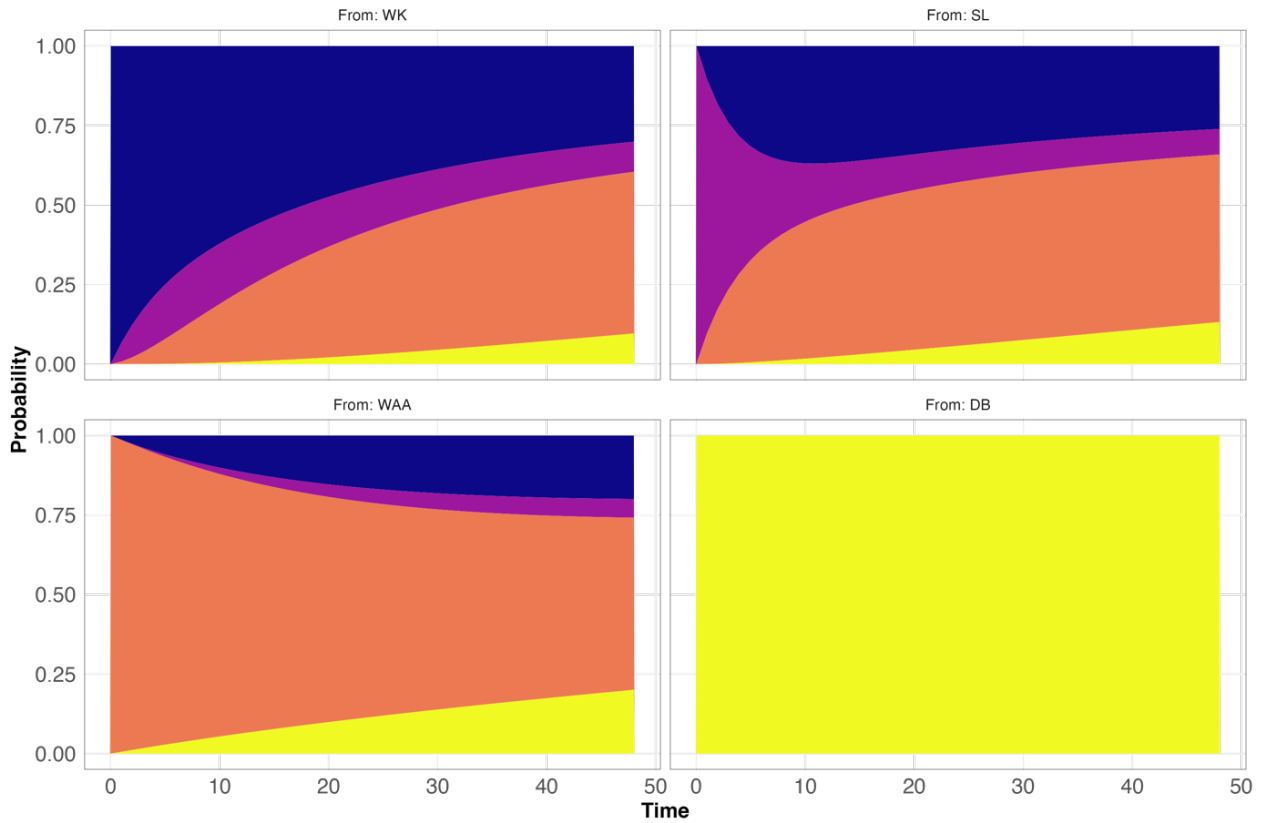
SL: sick leave

WAA: work assessment allowance

DB: disability benefit

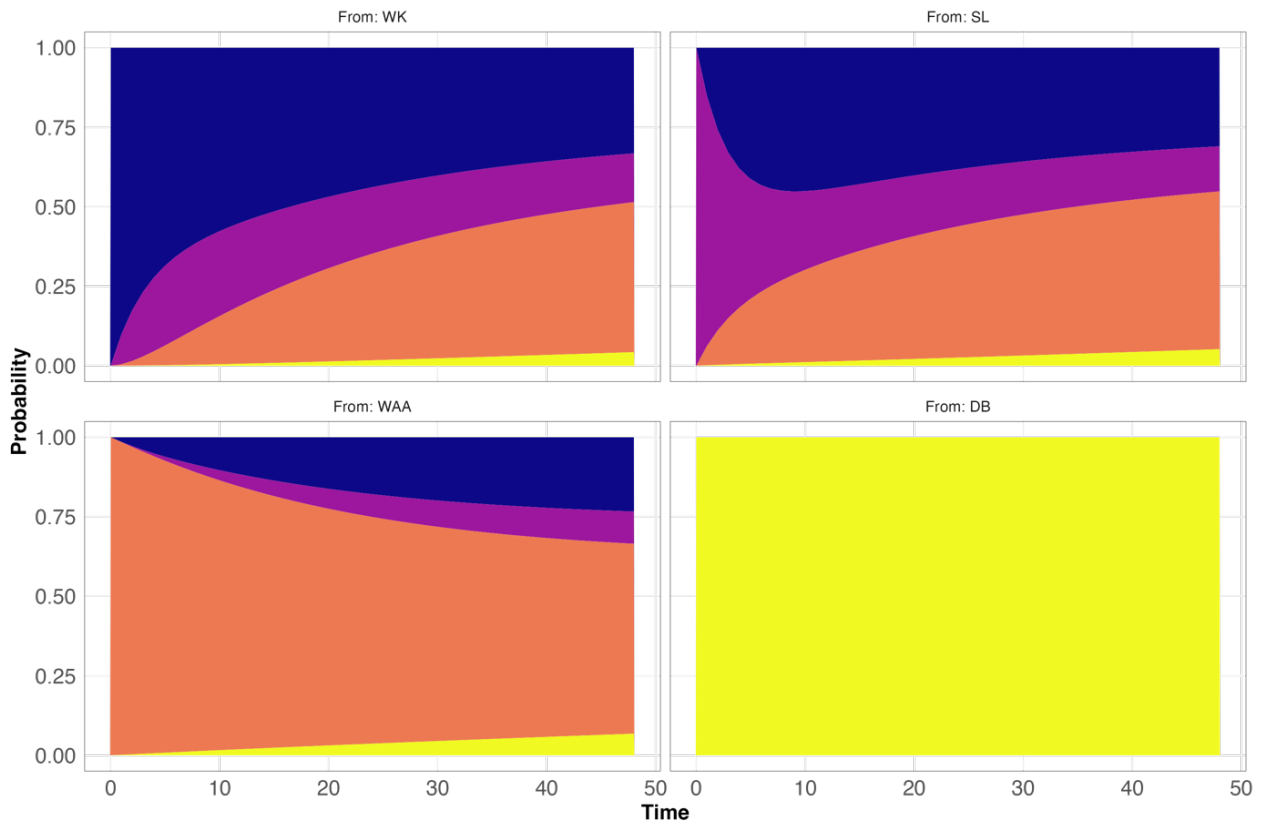
Figure 1 Stacked probabilities starting from all model states (base-case)

O-ACT



To: WK SL WAA DB

I-MORE



To: WK SL WAA DB

Figure 2 CEAC and CEAF with a limited societal perspective

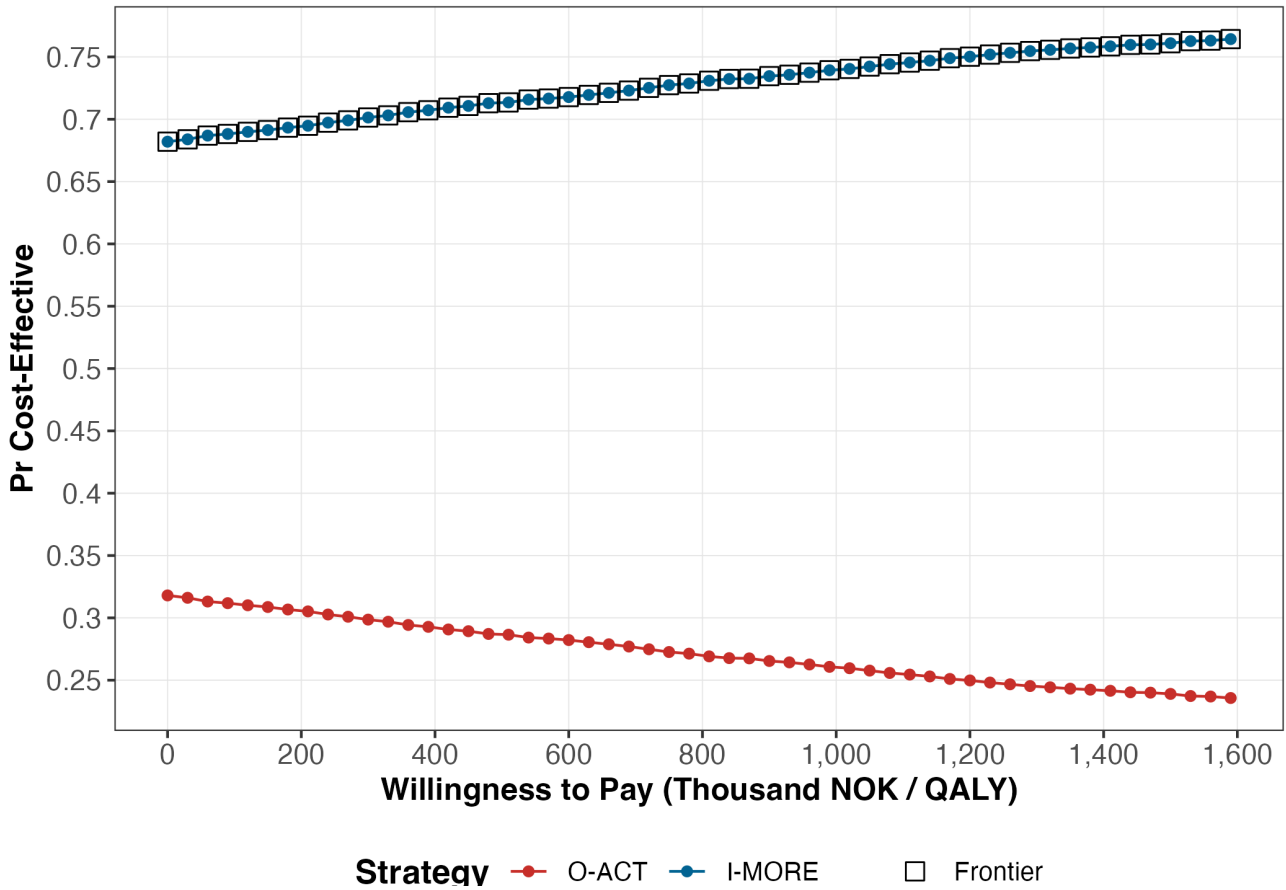


Figure 3 EVPPI for individual model parameters at a WTP of NOK 825,000.

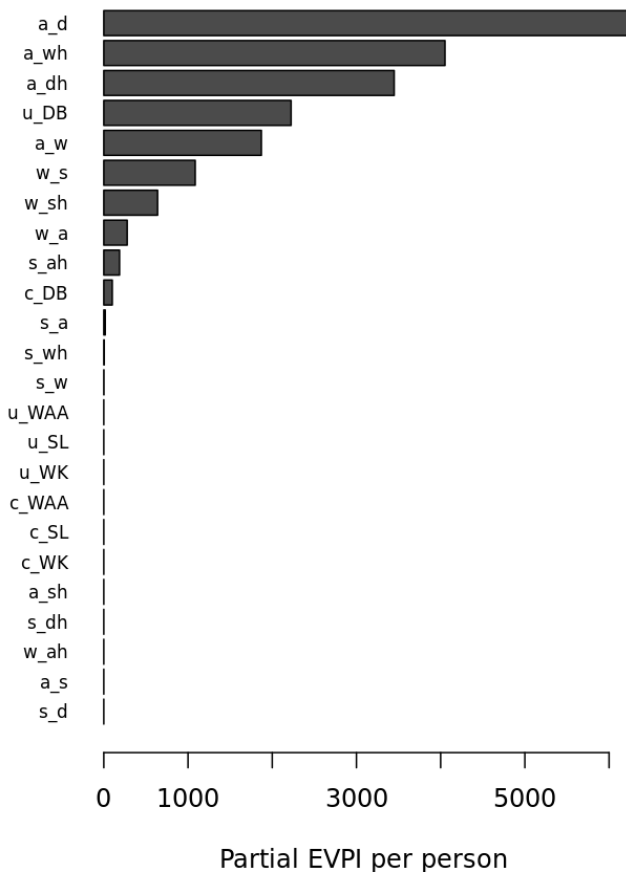


Table 6 Groups of parameters for EVPPI

Type of study	Name	Type of parameters	Parameters
Study on costs	Group 1	State costs	c_WK, c_SL, c_WAA, c_DB
Study on HRQoL	Group 2	State HRQoL	u_WK, u_SL, u_WAA, u_DB
Trial ^a	Group 3	Transitions	w_sh, w_ah, w_dh, s_wh, s_ah, s_dh, a_wh, a_sh, a_dh

Notes:

a) We only included I-MORE parameters, EVPPI can be expected to be higher as more parameters would be explored in a full trial

Appendix B: Validation checklist

TECH-VER checklist (Büyükkaramikli et al., 2019) available at:

<https://github.com/nasuhcagdas/TECHVER>

Test description (Please document how the test is conducted, as well)	Expected result of the test	
Pre-analysis calculations		
Does the technology (drug/device, etc.) acquisition costs increase with higher prices?	Yes	Yes
Does the drug acquisition cost increase for higher weight or body surface area?	Yes	NA
Does the probability of an event, derived from an odds ratio (OR)/ relative risk (RR) / hazard ratio (HR) and baseline probability, increases with higher OR/RR/HR?	Yes	NA
If survival parametric distributions are used in the extrapolations, can the formulae used for the Weibull (generalized gamma) distribution generate the values obtained from the exponential (the Weibull or Gamma) distribution(s) under some parameter transformations?	Yes	NA
In a partitioned survival model, does the progression free survival curve or the time on treatment curve crosses the overall survival curve?	No	NA
If survival parametric distributions are used in the extrapolations or time-to-event calculations, can the formulae used for the Weibull (generalized gamma) distribution generate the values obtained from the exponential (the Weibull or Gamma) distribution(s) after replacing/transforming some of the parameters?	Yes	
Is hazard ratio calculated from Cox proportional hazards model applied on top of the parametric distribution extrapolation found from the survival regression?	No, it is better if the treatment effect that is applied to the extrapolation comes from the same survival regression in which the extrapolation parameters are estimated.	NA
For the treatment effect inputs, if the model uses outputs from WINBUGs, are the OR, HR and RR values all within plausible ranges? (should be all non-negative and the average of these WINBUGs outputs should give the mean treatment effect)	Yes	
Event-state calculations		
Calculate the sum of the number of patients at each health state	Should add up to the cohort size	Yes. Proportions add up to 1
Check if all probabilities and number of patients in a state are greater than or equal to zero	Yes	Yes. Developed a function to check the trace in R,
Check if all probabilities are smaller than or equal to one	Yes	Yes
Compare the number of dead (or any absorbing state) patients in a period with the number of dead (or any absorbing state) patients in the previous periods?	Should be larger	Yes. It increases over time
In case of lifetime horizon, check if all patients are dead at the end of the time horizon	Yes	NA
Discrete event simulation specific: sample one of the "time to event" types used in the simulation from the specified distribution. Plot the samples and compare the mean and the variance from the sample	Sample mean and variance & the simulation outputs should reflect the distribution it is sampled from.	NA
Set all utilities to one Set all utilities to zero	The QALYs accumulated at a given time would be the same as the life years accumulated at that time No utilities will be accumulated in the model	QALYs turn to 0 when all utilities are 0 QALYs = LY when all utilities are 1
Decrease all state utilities simultaneously (but keep event based utility decrements constant)	Lower utilities will be accumulated each time	NA
Set all costs to zero	No costs will be accumulated in the model at any time	No costs are accumulated
Put mortality rates to 0	Patients never die	NA
Put mortality rate extremely high	Patients die in the first few cycles	NA
Set the effectiveness, utility and safety related model inputs for all treatment options equal	Same life years and QALYs should be accumulated for all treatment at any time	QALY=LY
In addition to the inputs above, set cost related model inputs for all treatment options equal	Same costs, life years and QALYs should be accumulated for all treatment at any time	Yes
Change around the effectiveness, utility and safety related model inputs between two treatment options	Accumulated life years and QALYs in the model at any time should be also reversed	NA
Check if the number of alive patients estimate at any cycle is in line with general population life table statistics	At any given age, the % alive should be lower or equal in comparison to the general population estimate	NA

Appendix B: Validation checklist

Check if the QALY estimate at any cycle is in line with general population utility estimates	At any given age, the utility assigned in the model should be lower or equal in comparison to the general population estimate	NA
Set the inflation rate of the previous year higher	The costs (which are based on a reference from previous years) assigned at each time will be higher	Yes
Calculate the sum of all ingoing and outgoing transition probabilities	Both should be one	Yes. Dedicated function.
Calculate the number of patients entering and leaving a tunnel state throughout the time horizon	Numbers entering = Numbers leaving	NA
Check if the time conversions for probabilities were conducted correctly.	Yes	Yes
Decision tree specific: calculate the sum of the expected probabilities of the terminal nodes	Should sum up to one	NA
Patient-level model specific: check if common random numbers are maintained for sampling for the treatment arms?	Yes	NA
Patient-level model specific: check if correlation in patient characteristics is taken into account when determining starting population?	Yes	NA
Increase the treatment acquisition cost	Costs accumulated at a given time will increase during the period when the treatment is administered	Partly, since treatment costs are a lumpsum in the model
Population model specific: set the mortality and incidence rates to zero	Prevalence should be constant in time	NA
Results calculations		
Check the incremental life years and QALYs gained results. Are they in line with the comparative clinical effectiveness evidence of the treatments involved?	If a treatment is more effective, it generally results in positive incremental LYs and QALYs in comparison with the less effective treatments	No previous HRQoL studies on the intervention. However the better outcomes of the trial, resulted in higher QALYs in our model
Check the incremental cost results. Are they in line with the treatment costs?	If a treatment is more expensive, and if it does not have much effect on other costs, it generally results in positive incremental costs.	Incremental cost is in line with the higher treatment cost, difference diminishes as time passes and control group consumes more care.
Total life years > total quality adjusted life years	Yes	Yes
Undiscounted results > discounted results	Yes	Costs and QALYs are higher when undiscounted
Divide undiscounted total QALYs by undiscounted life years.	This value should be within the outer ranges (maximum and minimum) of the all utility value inputs.	Within range
Subgroup analysis results: How do the outcomes change if the characteristics of the baseline change?	Better outcomes for better baseline health conditions and worse outcomes for worse health conditions are expected.	NA
Could you generate all the results in the report from the model (including the uncertainty analysis results)?	Yes	Parameter estimation was not performed with the decision analytic model. But with regressions and multi state modelling
Does the total life years, QALYs and costs decrease if a shorter time horizon is selected?	Yes	Yes they all decrease
Is the reporting and contextualization of the incremental results correct?	The use of the terms such as: "dominant" / "dominated" / "extendedly dominated" / "cost-effective" etc. should be in line with the results. In the incremental analysis table involving multiple treatments, ICERs should be calculated against the next non-dominated treatment.	Yes, although only 2 strategies are considered in the report
Are the reported ICERs in the fully incremental analysis non-decreasing?	Yes	Yes
If disentangled results are presented, do they sum up to the total results? (e.g. different cost types sum up to the total costs estimate)	Yes	Yes

Appendix B: Validation checklist

Check if half cycle correction is implemented correctly (total life years with half cycle correction should be lower than without)	The half cycle correction implementation should be error free. Also check if it should be applied for all costs, for instance if a treatment is administered at the start of a cycle, half cycle correction might be unnecessary.	Half cycle correction was not used given the short cycle length (1 month)
Check the discounted value of costs/QALYs after 2 years	Discounted value=undiscounted/(1+r) ²	Yes
Set discount rates to zero	The discounted and undiscounted results should be the same	Yes
Set mortality rate to zero	The undiscounted total life years per patient should be equal to the length of the time horizon	NA
Put the consequence of adverse event/discontinuation to zero. (zero costs and zero mortality/utility decrements)	The results would be the same as the results when AE rate is set to zero.	NA
Divide total undiscounted treatment acquisition costs by the average duration on treatment.	This should be similar to treatment related unit acquisition costs	NA
Set discount rates to a higher value	Total discounted results should decrease	Yes
Set discount rates of costs/effects to an extremely high value	Total discounted results should be more or less the same as the discounted results accrued in the first cycles	Although decreased, results are similar
Put adverse event/discontinuation rates to zero and then to extremely high level.	Less costs higher QALYS/LYs when adverse event rates are 0, higher costs and lower QALYS/LYs when AE rates are extreme	NA
Double the difference in efficacy and safety between new intervention and comparator and report the incremental results.	Approximately twice of the incremental effect results of the base case. If this is not the case : report and explain the underlying reason/ mechanism	NA
Do the same for a scenario in which the difference in efficacy and safety is halved.	Approximately halve of the incremental effect results of the base case. If this is not the case : report and explain the underlying reason/ mechanism	NA
Uncertainty analysis calculations		
Are all parameters subject to uncertainty included in the one-way sensitivity analysis (OWSA)? Check if the OWSA includes any parameters associated with joint uncertainty (e.g. parts of a utility regression equation, survival curves with multiple parameters).	Yes No	NA
Are the upper and lower bounds used in the one-way sensitivity analysis used confidence intervals based on the statistical distribution assumed for that parameter? Are the resulting ICER, incremental costs/QALYs with upper and lower bound of a parameter plausible and in line with a priori expectations?	Yes Yes	NA
Check that all parameters used in the sensitivity analysis have an appropriate associated distributions – upper and lower bounds should surround the deterministic value (i.e. Upper bound ≥ mean ≥ Lower bound) – standard error and not standard deviation used in sampling – Lognormal / gamma distribution for hazard ratios and costs/ resource use – Beta for utilities and proportions/probabilities – Dirichlet for multinomial – Multivariate normal for correlated inputs (e.g. survival curve or regression parameters) – Normal for other variables as long as samples don't violate requirement to remain positive when appropriate	Yes	Yes
Check PSA output mean costs, QALYs and ICER compared to the deterministic results. Is there a large discrepancy?	No (in general)	No. Although the use of bootstrap vs multivariate normal distribution for estimation of SE generates different ICERs (see Discussion)
If you take new PSA runs from the excel model do you get similar results?	Yes	NA
Is(are) the CEAC line(s) in line with the CE scatter plots and the efficient frontier?	Yes	Yes
Does the PSA cloud demonstrate an unexpected behavior or has an unusual shape?	No	No
Is the sum of all CEAC lines equal to 1 for all WTP values?	Yes	Yes

Appendix B: Validation checklist

Are the explored scenario analyses provide a balanced view on the structural uncertainty? (i.e. not always looking at more optimistic scenarios)	Yes	Yes (i.e, Scenario A)
Are the scenario analysis results plausible and in line with a priori expectations?	Yes	Yes
Check the correlation between 2 PSA results (i.e. costs/QALYs under the SoC and costs/QALYs under the comparator)	Should be very low (very high) if different (same) random streams are used for different arms	NA
If a certain seed is used for random number generation (or previously generated random numbers are used), check if they are they scattered evenly between 0–1 when they are plotted?	Yes	NA
Compare the mean of the parameter samples generated by the model against the point estimate for that parameter, use graphical methods to examine distributions, functions	The sample means and the point estimates will overlap, the graphs will be similar to the corresponding distribution functions (e.g. Normal, Gamma, etc.)	Yes, values were compared and aligned well with point estimates.
Check if sensitivity analyses include any parameters associated with methodological/ structural uncertainty (e.g. annual discount rates, time horizon).	No	No. But we present probabilistic results for 2-year and 25-year time horizons
Value of information analysis if applicable: Was this implemented correctly? Which types of analysis? Were aggregated parameters used? Which parameters are grouped together? Does it match the write-up's suggestions? Is EVPI larger than all individual EVPPI? Is EVPPI for a (group of) parameters larger than the EVSI of that (group) of parameter(s)? Are the results from EVPPI in line with OWSA or other parameter importance analysis (e.g. ANCOVA)?	Yes	Yes EVPI>EVPPI Parameters were grouped according to their nature and in a way that makes future studies possible. No EVSI
Did the electronic model pass the black-box tests of the previous verification stages in all PSA iterations and in all scenario analysis settings? (additional macro can be embedded to PSA code, which stops the PSA when an error such as negative transition probability, is detected)	Yes	Yes
Check the correlation between 2 PSA results (i.e. costs/QALYs under the SoC and costs/QALYs under the comparator)	Should be very low (very high) if different (same) random streams are used for different arms	NA
OWSA=one-way sensitivity analysis; ICER = incremental cost-effectiveness ratio; PSA = probabilistic sensitivity analysis; WTP = willingness to pay; CE = cost-effectiveness; CEAC = cost-effectiveness acceptability curve; LY = life years; QALYs = Quality adjusted life years; OR = odds ratio; RR= relative risk; HR = hazard ratio		