EXPLORING THE ROLE OF REPLICATION IN HEALTH ECONOMIC DECISION MODELS

Emma J. McManus

ORCID ID: 0000-0002-3442-8721

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Health Economics Group, Norwich Medical School, Faculty of Medicine and Health Sciences, University of East Anglia

December, 2023

Abstract

Several scientific disciplines have announced a “reproducibility crisis”, initiated by numerous high-profile studies being found to be unreproducible. Although health economic research has not been subject to such discussions, decision models are often termed ‘black boxes’ and there have been calls for heightened transparency in their reporting.

This thesis explores the role and value of replication within health economic decision modelling, specifically, how replicability is defined, what it means for a model to be replicable and the challenges facing modellers in incorporating replicability. This was achieved in a series of interlinked works. First, I identified studies defining replication success across all scientific disciplines. Whilst many studies discussed replicability, few defined replication success, none of which were found within health economics. Definitions ranged from subjective assessment to obtaining identical results. From these, definitions with varying specificity applicable to decision models were proposed.

Next, to examine factors influencing replication and assess the viability of the proposed definitions, five published models were replicated. This identified barriers and facilitators to replication and found that common reporting checklists were poor indicators of model replicability.

Finally, a decision model was developed with replicability in mind, to assess the feasibility of implementing the replication facilitators identified and overcoming the barriers. This highlighted the considerable time required to develop accessible models using open-source methods. These time requirements conflicted with the funded research project’s timeline, suggesting that in order to build replication into research, specific researcher time must be funded for replicability.

Overall, this thesis has shown that there is currently no consensus about how to define replication success for health economic models. Despite this, the importance of replication has been demonstrated as has the need for further work when reporting models to facilitate replication. To enable this, reforms to research infrastructure have been proposed.

Contents

Word count: 44,209

List of tables

List of figures

Abbreviations

CEAC: Cost-effectiveness acceptability curve

CHEERS: Consolidated Health Economic Evaluation Reporting Standards

COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments

DIPLOMA: Diabetes Prevention – Long Term Multimethod Assessment

DPP: Diabetes Prevention Programme

GORD: Gastro-oesophageal reflux

HbA1c: Glycated haemoglobin (A1c)

HEAP: Health economic analysis plan

HR: Hazard ratio

HTA: Health Technology Assessment

ICER: Incremental cost-effectiveness ratio

ISPOR: International Society for Pharmacoeconomics and Outcomes Research

MDS: Minimum dataset

N/A: Not applicable

NGT: Normal glucose tolerance

NHS: National Health Service

NICE: National Institute for Health and Care Excellence

NIHR: National Institute for Health and Care Research

QALY: Quality-adjusted life year

SD: Standard deviation

SE: Standard error

TAR: Technology Assessment Review

UK: United Kingdom

WHO: World Health Organization

Acknowledgements

I would like to thank my supervisors: Professor Tracey Sach, Professor Nick Levell and Mr David Turner for their unwavering support and guidance throughout the development of this thesis. I am extremely grateful for the time they have invested in me and I will truly miss our supervision meetings!

I would also like to thank my colleagues, past and present, for their encouragement and motivation in getting the PhD completed.

Author’s Declaration

Chapter 1

Excerpts of the following published paper have been used in the introduction chapter of this thesis, to describe methods of decision modelling.

> **McManus, E.**, Sach, T.H. and Levell, N.J., 2019. An introduction to the methods of decision-analytic modelling used in economic evaluations for Dermatologists. *Journal of the European Academy of Dermatology and Venereology*, 33(10), 1829-1836.

A license to reuse excerpts of this publication has been received by John Wiley and Sons and is provided at the end of this thesis in Journal permissions.

This work was also presented at the following conference:

> **McManus, E.**, McMonagle, C., Sach, T., 21st June 2016. An assessment of the transparency and quality of decision-analytic models in an entire clinical area. Discussed by Sullivan, W., at the Summer 2016 University College London Health Economists’ Study Group Meeting, Gran Canaria.

Contributions

I along with Prof Tracey Sach and Prof Nick Levell conceived the idea for the paper. I took the lead in writing the manuscript. All authors were involved in editing the manuscript.

Funding

This paper was written whilst myself, Prof Tracey Sach and Prof Nick Levell were funded by Prof Tracey Sach’s Career Development Fellowship (CDF award-2014-07-006) supported by the National Institute for Health and Care Research (NIHR). The views expressed in the thesis introduction are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Chapter 2

An earlier version of this chapter was presented at the following conference:

McManus, E., Sach, T., Turner, D., 21st June 2018. When is a model replication successful? Exploring the definition of success. Discussed by Faria, R., at the Summer 2018 University of Bristol Health Economists' Study Group Meeting.

A version of this chapter has been published:

McManus, E., Turner, D. and Sach, T., 2019. Can you repeat that? Exploring the definition of a successful model replication in health economics. *Pharmacoeconomics*, 37(11), 1371-1381.

A license to reuse excerpts of this publication has been received by Springer Nature and is provided at the end of this thesis in Journal permissions.

Contributions

I conceived the idea of this paper, with support from my supervisory team. I conducted the search strategy, searches and screening of papers described in this chapter. I was the lead author on the manuscript. Prof Tracey Sach and David Turner were co-authors on the publication that resulted from this chapter of work and commented on manuscript drafts.

Funding

This work was funded by the NIHR Collaboration for Leadership in Applied Health Research and Care, East of England (project HE7: Dec 2015). At this time, I also received funding via Prof Tracey Sach's Career Development Fellowship (CDF award-2014-07-006) supported by the NIHR. The views expressed in this thesis chapter are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Chapter 3

A version of this chapter has been published:

McManus, E., Turner, D., Gray, E., Khawar, H., Okoli, T., Sach. T., 2019. Barriers and Facilitators to Model Replication within Health Economics. *Value in Health*, 22(9), 1018-1025.

As author of this article, I retain the right to include it in a thesis or dissertation, provided it is not published commercially. Elsevier do not require me to obtain permission, as long as the journal paper is referenced as the original source.

Contributions

This chapter details model replications (case studies) conducted by myself, David Turner, Ewan Gray, Haseeb Khawar and Toochukwu Okoli. Both Haseeb Khawar and Toochukwu Okoli conducted this work as part of their MSc in Health Economics dissertation, which was co-supervised by Prof Tracey Sach and myself. The replicators' feedback on the barriers and facilitators of replication were used in the write up of the case studies, of which I was entirely responsible. I took the lead in writing the manuscript. All authors were involved in editing the manuscript.

Funding

This work was funded by the NIHR Collaboration for Leadership in Applied Health Research and Care, East of England (project HE7: Dec 2015). At this time, I also received funding via Prof Tracey Sach's Career Development Fellowship (CDF award-2014-07-006) supported by the NIHR. The views expressed in this thesis chapter are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Chapter 4

Aspects of the economic evaluation and model developed as part of this chapter were presented at the following conferences and seminars:

> **McManus, E.**, July 2021. Evaluating the effectiveness of the NHS Diabetes Prevention Programme. Health Services Research UK Conference (virtual).
>
> **McManus, E.**, November 2022. Evaluating the NHS Diabetes Prevention Programme. University of East Anglia.
>
> **McManus, E.**, September 2023. Evaluating the long-term cost-effectiveness of the NHS Diabetes Prevention Programme using a Markov Model. University of Manchester.

Contributions

The model developed in this chapter was done so as part of a funded research project. The model was developed by myself, with some preliminary input from Prof Matt Sutton and Dr Rachel Meacock who were work-package leads on the Diabetes Prevention – Long Term Multimethod Assessment (DIPLOMA) research grant.

Funding

This research was funded by the NIHR (Health Services and Delivery Research, 16/48/07 – Evaluating the NHS Diabetes Prevention Programme (NHS DPP): the DIPLOMA research programme). The views and opinions expressed in this thesis chapter are those of the author and do not necessarily reflect those of the NHS, the NIHR or the Department of Health and Social Care.

Chapter 1 Introduction

1.1 Introduction to health economics

Health economics is the application of economic theory in the context of health and health care (Morris et al., 2007). At its cornerstone, is the premise that in society, there are limited resources to provide health care, but there is unlimited health need. Health economics studies the production of health and health care, and the allocation of these scarce resources, taking into consideration the efficiency and equity of allocations. There are different types of efficiency which include: technical, productive and allocative efficiency (S. Palmer & Torgerson, 1999). Technical efficiency is defined to be obtaining the greatest output for a given unit of resource (Knapp, 1984) and productive efficiency relates to the maximisation of output for a given cost. Allocative efficiency considers both productive efficiency and the optimal allocation of goods and services across society so as to maximise the welfare of that society. Equity on the other hand, is more complex and refers to the fair distribution of that output throughout society. The World Health Organization (WHO) refers to equity as “the absence of unfair, avoidable or remediable differences among groups of people” (World Health Organization, 2023), and therefore may encompass a range of measures such as access to health care, health outcomes or allocated resources. In making choices to allocate resources, there is an inherent trade-off in that the same resources cannot then be allocated elsewhere. In allocating resources to one area instead of another, there is also the lost opportunity to provide benefit. This is referred to as the opportunity cost.

Health and health care are unique to other conventional economic markets, in several respects (Arrow, 1963; Olsen, 2022). The first is that access to health care is considered to be a basic human right (World Health Organization, 1948), and so unlike traditional markets, it is considered inhumane to let the market decide who can (those who are willing and able to pay) and those who cannot consume it. The demand for health care is also a derived demand, in that it derives from the fundamental demand for good health (Santana et al., 2023). There is also information asymmetry, in that a healthcare provider typically has more knowledge of the service being provided than the patient (consumer) and therefore they may be unable to determine alone which strategy to choose or the quality of the health care they are receiving, creating an imbalance which is not usually observed in an economic market. As well, the concept of health itself is intangible, in that it cannot be traded or inherently passed on from one person to another, but it can generate positive externalities to the wider society. Positive externalities refer to the positive spill-over effect from the consumption or action of an individual on a third-party, an example of which would be herd immunity from vaccination programmes (Mwachofi & Al-Assaf, 2011). There are also multiple factors than can affect health, beyond just the provision of health care.

1.2 Economic evaluation

Given the scarcity of resources, any health care intervention to be implemented needs to be demonstrated as a good use of funds, or value for money. Demonstrating value for money is now considered the fourth hurdle of technology approval, along with quality, safety and efficacy (Paul & Trueman, 2001).

To investigate whether an intervention is value for money, an economic evaluation can be performed. The aim of an economic evaluation is to identify which treatment (or prevention strategy) represents the most effective use of resources, commonly referred to as cost-effectiveness.

Economic evaluation is a systematic approach, involving the identification, measurement and valuation of inputs and outcomes of two or more alternative health care interventions (Drummond et al., 2015), measured in terms of their costs and health benefits. The methods used to conduct such evaluations may vary, for example, in terms of the perspective taken, the costing methods used, the measure of benefit and the time horizon taken. There are also different types of evaluations, such as cost-minimisation, cost-effectiveness, cost-utility and cost-benefit analysis. Whilst all of these evaluation types measure costs in monetary terms, they vary in how they value outcomes. In cost-minimisation analysis, it is assumed that the health outcomes of different interventions are identical and so the focus is on finding the intervention with the lowest costs. In cost-effectiveness analyses outcomes are measured in terms of a health outcome or unit measure. Cost-utility analyses use quality-adjusted life years (QALYs) as their measure of outcome (there is more on this concept below). In cost-benefit analyses, health outcomes are valued in terms of a monetary value, assessing whether the monetary value of benefits is greater or less than the costs of obtaining them (Drummond et al., 2015). These methodological variations may lead to different results being obtained. In an attempt to standardise this, the National Institute for Health and Care Excellence (NICE) who produce guidelines and recommendations for England, Wales and Northern Ireland, developed a reference case, outlining the desired methods for conducting economic evaluations (National Institute for Health and Care Excellence, 2013). In this, NICE stated that economic evaluations should be undertaken using cost-utility analyses and incremental analysis. They also stated the primary perspective to be used and advocated for the EQ-5D to be used as its preferred measure of health-related quality of life in adults (National Institute for Health and Care Excellence, 2022), although NICE accepted that other instruments might be used if the EQ-5D is unavailable or unsuitable. Preference weightings used alongside EQ-5D responses generate utility values, which when paired with the length of time in that state can be converted into QALYs (Whitehead & Ali, 2010). QALYs capture both the effect on survival and the quality of life.

Measuring outcomes with a generic measure, such as QALYs, facilitates the cross comparison of interventions for different disease areas. A health care intervention would be determined as cost-effective if the costs for a unit of benefit were below a given willingness to pay threshold. NICE currently (in 2023) states that the willingness to pay for an additional QALY generated is between £20,000 and £30,000 (National Institute for Health and Care Excellence, 2013). Academic debate has challenged this figure, with some suggesting this threshold ought to be considerably lower, with Claxton et al. (2015) estimating £13,000 per QALY and S. Martin et al. (2023) suggesting between £6,000 to £8,000. In contrast, others have suggested it should instead be raised (Low & Macaulay, 2022).

Economic evaluations are often conducted alongside clinical trials (known as within-trial economic evaluations); however, there is often conflict between the clinical and economic objectives (Raftery et al., 2020). Usually, the sample size is calculated with the intention of demonstrating clinical efficacy rather than for economic outcomes. As is the timeframe of trials, and thus trials may end before the outcomes of economic interest have been fully observed and measured. The setting of the trial may also influence the economic analysis, due to strict treatment protocols designed to improve treatment adherence. Therefore some of the findings of the within-trial economic evaluation may not be transferable to real-life practice (Morris, 1997).

1.3 Decision modelling

To consider a longer time horizon, a decision-analytic model (from herein referred to as decision model) can be used to extend an economic evaluation, building upon the findings of short-term trial analyses and allowing them to be extrapolated. Models may also facilitate the comparison of multiple treatment options. A decision model is defined as:

> “An analytic methodology that accounts for events over time and across populations, that is based on data drawn from primary and/or secondary sources, and whose purpose is to estimate the effects of an intervention on valued health consequences and costs” (Weinstein et al., 2003).

The benefits of using a modelling approach are that multiple sources of evidence can be used, an extended time horizon can be considered, the effect of changing parameters can be explored and perhaps most importantly, the uncertainty surrounding the long-term result can be assessed. Whilst decision models have many associated advantages, it should be acknowledged that they are not complete alternatives to within-trial economic evaluations, as the economic data from these trials is often used within modelling studies.

The most common modelling approaches used within economic evaluations (sourced from McManus, Sach, et al. (2019)) are described below.

1.3.1 Common Modelling Approaches

1.3.1.1 Decision Tree Model

The decision tree is often the simplest modelling method available and may be used to model one-off decision processes (P. Barton et al., 2004). To produce a decision tree model, the tree must begin with a decision node, which is a point where a choice is made. Importantly, the choice options branching from the decision node must be mutually exclusive, meaning if one is chosen then the other is not.

Along each branch there may be further nodes (referred to as event or chance nodes), which represent points at which different events can arise (for example switching to a second-line antibiotic or not). As with the decision node, the events represented by the chance node must also be mutually exclusive, as well as being collectively exhaustive, meaning that all possible patient pathways are shown.

Alongside each of these branches, probabilities are displayed which show the likelihood of the event occurring, and at the end of each patient pathway or branch, the resulting outcome measures are displayed, such as effect on utility value and cost.

A decision tree is an appropriate choice of model when the time horizon is short, the individuals represented in the tree can be thought of as independent from one another and the number of events spanning from the decision node are manageable. If these criteria are satisfied, then decision trees are usually simple to produce and make calculations with.

However, due to the general probabilities applied, decision trees represent an aggregate (population) level approach and therefore do not consider individual-level attributes. They may ignore characteristics of the patient that may make certain events unlikely (such as antibiotic allergy). Furthermore, decision trees do not demonstrate the passage of time, only that at some point an event will occur. This is why they may be regarded as only suitable to model events over a short time horizon. Thus, they are not appropriate to model chronic illnesses or choices that may vary greatly depending on individual attributes.

1.3.1.2 Markov Cohort Model

A Markov model (also referred to as a state-transition model) comprises a finite number of mutually exclusive and collectively exhaustive disease or treatment states. These states aim to represent the consequences of treatment options under analysis (Sonnenberg & Beck, 1993). Attributed to each disease state is a cost and associated utility value for being in that state. It is possible to transition between these disease states, which allows the Markov model to deal more succinctly with disease recurrence and flare up than the growing number of branches that would be seen within a decision tree. The likelihood of the patient moving from one state to another is defined by transition probabilities. The main advantage of using Markov models is their ability to deal easily with recurrent events (P. Barton et al., 2004).

Time is represented in the model using unit cycles. It is assumed that only one state transition (e.g. moving from remission to eczema flare up) can be made during each cycle. The cycle length must be short enough so that events that change over time can be represented by individual, successive cycles (Briggs & Sculpher, 1998).

A great benefit to introducing time cycles into the model is that the transition probabilities between states, as well as the cost and health utilities experienced can vary with time. This, for example, allows the transition probability from any state to the “Dead” state to increase over time, representing either general or disease specific mortality. During each time cycle, the various costs and utilities attributed to being in each disease state can be totalled. This gives a different cost and overall health utility output dependent on the pathway taken (representing the course of the disease) and the number of cycles spent in each state.

All of this is represented in a state transition diagram, where disease states are represented as circles, and arrows from these circles represent the possibility and direction of transition to a different disease state. It is possible to remain in the same transition state for consecutive cycles (P. Barton et al., 2004), which is represented by circular arrows going and returning to the same state. It is also possible to construct an absorbing state, one that a cohort (or individual) can enter but not leave. The most common example of an absorbing state would be death.

One fundamental restriction that must hold in a Markov model is known as the Markovian assumption (Briggs et al., 2006, pp. 36-37). This specifies that the probabilities that govern how an individual stays in, or moves from, any given disease state are not affected by the previous disease states or the duration spent in such states. In this sense, the Markovian assumption means that the process has no memory and that all individuals within any given state are treated in the same way (homogeneity).

This inherent “lack of memory” is a disadvantage of Markov Cohort models. However, the severity of this limitation can be reduced by creating additional states that take into account the history of the individual. As well, a series of “tunnel states” could be introduced. Tunnel states allow transition through one state directly into another, which might allow for an extended treatment time.

Despite there being potential to build in some form of memory into the Markov model, as with the decision tree, the model may quickly become complex with a cumbersome number of disease states.

1.3.1.3 Markov Monte Carlo Simulation

Instead of assuming patients can be grouped into homogenous cohorts as is done in the Markov approach above, it is also possible to simulate patients with individual level attributes, using Monte Carlo Simulation (Brennan et al., 2006). In this process, each patient begins in a given starting state. At the end of each cycle, a random number generator (see section 4.4 of Briggs et al. (2006)) produces a value, from which this and the predetermined transition probabilities determine which state the individual will move to for the beginning of the next cycle (P. Barton et al., 2004). In a simplistic example with only two states: alive and dead, where the transition probability of staying alive over a twelve month period is 0.7 and the probability of death is 0.3, a random number can be generated between 0 and 1. If the random number is between 0 and 0.7, the individual will remain in the well state, if the number is between 0.7 and 1, the individual will move to the dead state. This process is repeated over a finite number of cycles, defined as the time horizon of the model, or until the individual has reached the dead state (which is an example of an absorbed state). As with the Markov Cohort model, each respective state has associated utility values and costs, able to vary with time, which accumulate over the number of cycles. This process can be repeated to simulate a large number of individuals. The Markov Monte Carlo simulation gives a measure of variability that is not possible with the previously described Markov Cohort approach (Briggs & Sculpher, 1998).

1.3.1.4 Discrete Event Simulation

Discrete Event Simulation is a method primarily used for modelling queue systems or processes (Caro, 2005), an example might be to look at the effects of changing a particular health service pathway (Vahdat et al., 2018). This is achieved by allocating each individual their own attributes, which may then affect their progression through the model and the events that occur.

The discrete event simulation model structure comprises of entities, events, resources and time (Brennan et al., 2006). Entities are the items (usually, but not always; patients) that proceed through the simulation. Each entity can be given different attributes, such as age, sex or duration of disease, and these can be updated as the entity progresses through the simulation. Events refer to any defined diseases or treatments that may occur during that patient's lifetime. Events may occur simultaneously and future events may be determined by previous event history. The occurrence of an event in the model does not necessarily imply that the patient has changed disease state. This approach allows patients to experience competing probabilities of risks; in which the experience of one event, may influence subsequent risks to both the individual and other entities within the simulation (for example, possible reduced access to a rationed drug).

Timing within discrete event simulation is based on an events list; all events that take place are listed in the model in a way that allows them to be processed in a chronological order. In contrast to the Markov process which focuses on the probability of transitioning to another state, discrete event simulation is focused on the events an entity has experienced and the decision about what the next event will be and for how long until it occurs (Briggs et al., 2006, p. 59). By having an events list, the idea of a queue system (e.g. patients waiting for a referral to secondary care) can be introduced into the model. With discrete event simulation, it is unnecessary to specify the unit of time, as patients move through the model and can experience events at any discrete point. This means that discrete event simulations can proceed very efficiently, as the simulation clock can advance to the time when the next event will occur, without conducting the interim computations required in models that utilise unit cycles (Caro, 2005). Resources are incorporated directly, and entities are able to consume a resource at any appropriate time, it is also possible for entities to consume more than one resource (e.g. multiple medications) at a time.

Overall, discrete event simulation provides greater flexibility than a Markov process and it may also add a greater sense of realism to the model than the use of disease states and transition probabilities (Karnon, 2003). However, to achieve this, discrete event simulation requires a large volume of clinical data to populate parameters, access to specific software, specialist programming knowledge as well as the need for greater computational power. As well as this, due to the complexity of discrete event simulation, it is often difficult to thoroughly and transparently report the methods and data sources used within the model within the confines of a published manuscript.

1.3.2 Model transparency

Regardless of the modelling approach chosen, in order for a model to be informative and usable, model developers must endeavour for it to be transparent, reliable and valid. Model validation is a form of quality assurance and comes in many forms, including but not limited to: face validity, internal validity, cross-validity and external validity (Eddy et al., 2012). Face validity relates to ensuring that the structure of the model makes intuitive sense, and also that it is clinically valid, in that it accurately captures the condition represented and the clinical pathway. Internal validity relates to the mathematical equations and coding used, ensuring that it is free from errors. Cross-validity compares the results of a model to those of existing models answering the same research question. External validity relates to how the model output compares with observed real-world data and how its predictions may compare to observed outcomes. It is also important for models to adequately capture the uncertainty surrounding a decision, this can be achieved by conducting sensitivity analyses. Sensitivity analyses may involve the variation of parameters within a certain range, using different sources of parameters or changing of the model structure and any underlying assumptions. All of these aspects are important in their own right and contribute to a transparent and valid model. This is important as models are often considered to be ‘black boxes’ (Birkmeyer & Liu, 2003) by those who make decisions using them, meaning that the mechanisms by which the model makes conclusions are not clear and with many commenting that they lack adequate transparency (Sampson et al., 2019). Decision models are also particularly vulnerable to manipulation, given that researchers often have free license on how the model is developed and the parameters that are chosen, which may impact the results and the overall cost-effectiveness estimates obtained. This is especially true if the model developers have ties to the intervention being evaluated, where there may be a perceived pressure for the model to derive favourable results. This idea of decision models lacking transparency is the central theme of this thesis. The concept of research transparency and how it relates to health economics, specifically decision models is explored in more detail below.

1.4 Introduction to research transparency

1.4.1 Introduction to transparency

Alongside the need for research to be ethical, relevant, valid and reliable (Jansen & Warren, 2020), a fundamental concept of open research is that it should be transparent. Research transparency involves the detailed reporting of methods, datasets and any statistical analyses conducted so that others may inspect the scientific rigour and understand what was conducted (Prager et al., 2019). Truly transparent research should be reported with sufficient detail so that it may be repeated and the results replicated by a third party. Research transparency is essential for building trust in the scientific community, promoting research integrity, and advancing knowledge by increasing the reliability and verification of results as well as helping to identify errors. It is also considered a part of good, ethical research practice and a sign of academic rigour.

There are numerous reasons why research transparency is essential. Firstly, it may help to facilitate research development. With increased transparency, researcher effort can be focused on developing new ideas rather than reinventing aspects of previous research due to a lack of reporting transparency and the need to repeat processes, hence reducing research waste (Chan et al., 2014). Increased transparency is likely to speed up the research process and facilitate the spread of knowledge. Other reasons may be more malign. For example, researchers face increased pressures to publish, which may be referred to as “publish or perish” (Van Dalen & Henkens, 2012). These pressures, paired with academic journals’ tendency to publish positive, novel results, known as publication bias (Franco et al., 2014), may incentivise researchers consciously or unconsciously to manipulate or falsify results. There may also be commercial interests related to the research being conducted, which may influence the methods or reporting of results (Goldacre, 2014), and thus reduce transparency (Moynihan et al., 2019). In a meta-analysis of researcher surveys exploring misconduct, Fanelli (2009) found that on average 1.97% of scientists admitted to having “fabricated, falsified or modified data or results at least once” and up to 33.7% reported “questionable research practices” (Fanelli, 2009). This highlights the need for a research system whereby other researchers can examine the results of others, and thus a need for transparency. Another argument in favour of increased transparency relates to the possibility of errors within the research, which may mean that the results of studies are not replicable. One study examined the reason for article retractions, and found that 7% were due to data fabrication and 13% related to “honest error”, showing that such mistakes do occur. The average time between publication and retraction for these articles was 337.5 days, which implies that the results may already have been used in other studies or to inform policy decisions (Moylan & Kowalczuk, 2016).

There have been calls for greater research reproducibility and research integrity. Most recently (2023), the UK government commissioned a report on the reproducibility of research (Science Innovation and Technology Committee, 2023), citing the increased public investment in research and development (with the Government Research and Development budget set to reach £20 billion annually by 2024/5) and the fact that some research has been found to not be reproducible, with some scientific disciplines going as far to suggest that there is a “reproducibility crisis” (Baker, 2016; Chang & Li, 2015; Ioannidis, 2005; Maxwell et al., 2015).

1.4.2 Transparency and health economics

Whilst transparency should be a goal of all scientific disciplines, it is especially important within health economics, given that research findings may have direct implications on health policy, such as the commissioning of services or provision of treatments. In fact, it is stated within NICE guidance that researchers conducting economic evaluations should have the “highest level of transparency in reporting methods and results” in order to ensure appropriate, transparent and robust decision making (National Institute for Health and Care Excellence, 2013). This is particularly important given that research is often funded by public money (such as with the NIHR) and as such should be openly accessible, to allow for reuse and scrutiny. With regards to decision models, publications by the International Society for Pharmacoeconomics and Outcomes Research & The Society for Medical Decision Making (ISPOR-SMDM) Task Force provide an explicit definition of a transparent decision model. Significantly, this definition cites the process of replication, suggesting that the ability to replicate a model may indicate that it is transparent:

> “Transparency serves two purposes: 1) to provide a non-quantitative description of the model … and 2) to provide technical information to readers who want to evaluate a model at higher levels of mathematical and programming detail, and possibly replicate it” (Eddy et al., 2012).

In another publication by the Task Force, the importance of transparency is repeated, stating that “a model should not be a ‘black box’ for the end-user but be as transparent as possible, so that the logic behind its results can be grasped at an intuitive level” (Weinstein et al., 2003). This is echoed in a study focusing on models used within oncology, which concluded that, “there is a need for elevated rigor and transparency of reporting” (Beca et al., 2017) although exactly how this might be achieved was not discussed.

1.4.3 Transparency initiatives

Several initiatives have been developed within health economics to try and improve research transparency. Applicable to all types of economic evaluation are initiatives such as reporting checklists and Health Economic Analysis Plans (HEAPs) (Thorn et al., 2021). Several checklists have been developed to facilitate thorough reporting of methods and critical appraisal. In a systematic review of checklist use within economic evaluations, a study identified 18 different checklists used between 2010 to 2018 (Watts & Li, 2019). One of the most commonly used is the Consolidated Health Economic Evaluation Reporting Statement (CHEERS) (Husereau et al., 2022; Husereau et al., 2013) which is a relatively simple, 24 point (28 point in the 2022 updated version), checklist designed to ensure thorough reporting of economic evaluations, both those conducted alongside trials and using a decision model. Given the generalisability and relative shortness of the CHEERS checklist, it is this checklist that is often required when submitting research to peer-reviewed journals. This checklist was updated in 2022, and now includes wording within the “if modelling is used, describe in detail and why used” item to also “report if the model is publicly available and where it can be accessed”. There is also a decision modelling specific checklist, the Philips checklist (Philips et al., 2004), which is more commonly used by health economists to critically appraise decision models. This checklist is more comprehensive, comprising of over 50 items designed to evaluate if the decision model and any underlying assumptions have been thoroughly reported.

Researchers undertaking economic evaluations have also been pushed towards developing HEAPs, which is a guidance document or standard operating procedure which outlines how an economic evaluation will be conducted in predetermined steps. This is to avoid selection of analyses or parameters once data is available such that decisions about methods could then be based on the results they obtain. Such plans are not yet routine practice, one paper by Dritsaki et al. (2018) evaluated the usage of HEAPs across clinical trials units and found that they were used as standard in only one third of clinical trial units. Although this paper did not consider the use of HEAPs when developing decision models.

Other transparency initiatives include the calls for modelling registries and the push to publish models as open-source (Sampson & Wrightson, 2017). These calls have, so far, had limited impact on how decision models are published, with Sampson et al. suggesting that whilst calls for model transparency have been numerous, “there are few signs of improvement in practice” (Sampson & Wrightson, 2017). This was also highlighted in a study by Emerson et al., which conducted a survey of US authors who had published papers describing decision models from 2010 to 2017 to investigate their willingness to share their model source code. Of the 248 distinct authors surveyed, 7.3% responded to the request for model code (n=18), of which five said they would share code, of whom only four actually did (Emerson et al., 2019). Notably, there have also been arguments against open-source modelling, with Padula et al. (2017) suggesting that there may be issues surrounding the retention of individuals’ intellectual property and also that open-source models do not necessarily facilitate clinician access or interpretation. Instead, the authors argued for the journal peer-review process to become more rigorous by providing all technical details, including the model, to reviewers who had signed a confidentiality agreement. Other modelling initiatives to increase transparency and the betterment of modelling methods, include the Mount Hood diabetes modelling group, who have developed guidance on developing diabetes simulation models and also developed model validation methods. One of these, uses a set of reference input parameters which can be used across different simulation models to see the variation in results, in what is known as “comparative modelling” (Kim & Neumann, 2019). In another disease area, there is the Birmingham Rheumatoid

Arthritis Model (BRAM). This initiative brought together different modellers in order to develop a consensus model that is continually developed and updated, and has now been used for several health technology appraisals, rather than developing multiple piece wise models separately (Pelham Barton, 2011).

1.5 Replication

One way of exploring how transparently research is reported is to conduct a replication study. A replication study is an independent duplication of a published study using (in the most basic sense), the same methods and data, with the intention of recreating the reported results. Aside from evaluating transparency, there are several other motivations for conducting a replication study, these range from checking for calculation errors (McCullough et al., 2008), to demonstrating understanding of the original study (Duvendack et al., 2015) and to then improve and extend upon existing research (Chang & Li, 2015). Replication studies are also valued as a learning tool. For example, in the wider economic disciplines, there have been instances where replications are commissioned as coursework for students as a means of gaining practical experience of the techniques and theory or modelling that they have learnt. In the infamous case of Reinhart and Rogoff (Herndon et al., 2014), such a coursework replication study resulted in the identification of serious calculation errors.

There is growing consensus that an independent modeller should be able to reproduce the results of a model using only the published information (Eddy et al., 2012; Weinstein et al., 2003). Replicable models have practical benefits, in terms of potentially reducing research waste as well as reducing researcher time spent developing future models, as advocated by the REWARD Alliance (The REWARD Alliance, 2016). For example, if an existing model was easily replicable, this could mean future modellers may be able to use this model as a springboard for the development of another, leaving more time to devote to validation work. Chilcott et al. (2010) discuss the concept of replication as a method to check the face validity of models in the development process, as well as citing the potential benefits of replicating a model using different software.

1.6 Aim and research questions

The concept of transparency is of great importance to research, as highlighted above, and is especially needed within health economics to ensure evidence-based decisions are made. This thesis will focus on exploring the role and value of replication within health economic decision models, as an indicator of overall transparency. Given the remit and scope of the thesis, it will focus on the replicability of decision models, rather than within health economics as a whole. This thesis therefore aims to address the following research questions:

- RQ1. Why is replication needed?
- RQ2. How do other scientific disciplines approach replication?
- RQ3. What is the role of replication in decision models within health economics research?
- RQ4. How could a successful replication be defined?
- RQ5. What are the barriers and facilitators to replicating decision models?
- RQ6. What are the implications of a model being replicable (or not)?
- RQ7. What are the implications and challenges for modellers trying to incorporate replicability?
- RQ8. Does the ability to replicate lead to greater transparency?

1.7 Thesis structure

This thesis consists of three substantive chapters (Chapters 2, 3 and 4), presented in the style of journal articles. Below, I outline the role of each of these chapters and the research question(s) that they aim to address. As this thesis represents an iterative body of work, some research questions are revisited at multiple points throughout the thesis.

Chapter 2 begins by reviewing the existing replication literature. It aims to identify how a successful replication has been defined both outside of health economics and, if at all, within health economic research, to address questions RQ2 and RQ3. Based on these findings, it then proposes several definitions with variable levels of specificity that could be used when replicating decision models within health economics, addressing RQ4.

Chapter 3 attempts to replicate five published decision models, with the intention of identifying common facilitators and barriers to replication. This chapter addresses questions RQ5 to RQ6. Within this chapter, the definitions proposed in Chapter 2 are applied, to evaluate their usability, revisiting question RQ2.

Chapter 4 focuses on the development of a decision model with the intention of future replicability to address RQ7. A list of items identified from the previous chapters work are compiled and evaluated to see if and how they can be incorporated into the model development process.

Chapter 5 provides the final discussion, which summarises the key results from the chapters above, as well as discussing the overall implications of this research, addressing the overarching research questions of RQ1 and RQ8. Finally, areas for further research are proposed.

Chapter 2 Defining a successful replication

2.1 Introduction

The concept of replication is widely discussed across scientific disciplines, including but not limited to: biomedicine (Iqbal et al., 2016), computational science (Peng, 2011; Rougier et al., 2017), psychology (Makel et al., 2012) and epidemiology (Peng et al., 2006). Replication has also been explored in other economics disciplines such as development and strategic economics (Bettis et al., 2016; Brown et al., 2014; Duvendack et al., 2017; Duvendack et al., 2015; Höffler, 2017).

Publications from various scientific disciplines have announced a “reproducibility crisis” (Baker, 2016; Chang & Li, 2015; Ioannidis, 2005; Maxwell et al., 2015), initiated by several high profile studies being found to contain errors or to not be reproducible. In response, numerous replication initiatives (Berkeley Initiative for Transparency in the Social Sciences, 2017; International Initiative for Impact Evaluation, 2017; Pashler et al.; The Replication Network, 2017) have been formed, with the aim of maintaining research integrity and transparency. One of these, The International Initiative for Impact Evaluation (3ie), commissions replication studies in development economics, to improve research quality and promote good research practices (International Initiative for Impact Evaluation, 2017).

A significant amount of research has explored what constitutes a replication, resulting in the development of various taxonomies classifying the different types and definitions of replication (Bettis et al., 2016; Duvendack et al., 2017; Schmidt, 2009). Most often, these distinguish between broad and narrow replications (Bettis et al., 2016; Clemens, 2017; Goodman et al., 2016). A broad replication is defined as using data from other “periods, countries, regions or other entities as appropriate” to see if the empirical finding can be repeated (Pesaran, 2003). Broad replications can improve the understanding of a concept, test the robustness of results, as well as show how generalisable the results might be (Bettis et al., 2016). In this chapter, I consider replication in the narrow sense, that is: a replication attempt conducted using the exact same methods and data as in the original study, with the intention of regenerating the original results (Bettis et al., 2016).

As well as the distinction between narrow and broad, it is also notable that an innate difference exists within replication across different disciplines, regarding what is being replicated. For example, replication is considered in experimental disciplines (whereby one seeks to replicate an experiment observing an effect) and by contrast in analytical disciplines, which involves coded analysis of a dataset. To demonstrate, authors of a study evaluating replication in computer science remarked that: “In theory, computation is a deterministic process and exact reproduction should therefore be trivial” (Rougier et al., 2017), meaning that it should be as easy as rerunning lines of code. In comparison, a replication of an experiment or clinical trial may have inherent variability due to variation in subjects or other external factors, which may prevent the original results from being exactly reproduced.

In this chapter, I explore how other scientific disciplines have defined replication success, and whether any such definition has been proposed in the health economic literature. This is a logical starting point. Whilst there may be value in conducting replications of decision models within health economics – it is first important to define what the aim of the replication is and therefore what constitutes replication success. Without such a definition, researchers would likely have to make a subjective assessment.

From this, I propose how the definitions identified might be tailored to health economic models. This work was motivated by the COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments) initiative (Mokkink et al., 2010) founded in 2005. This initiative sought to standardise terminology and definitions within measurement properties, with the aim of improving the quality of studies.

2.2 Methods

A non-systematic literature review was conducted with the intention of identifying how a ‘successful replication’ was defined, firstly within scientific disciplines outside of health economics, and secondly to explore if the concept had been defined within health economic research. Whilst systematic literature reviews are considered to be the gold standard, this approach was not used due to the broad range of studies this review sought to identify, the nebulous nature of the search terms and the frequent use of ‘replication’ within basic science (for example relating to DNA or virus replication). Therefore, it was not possible to develop a search strategy with sufficient specificity and sensitivity such that a systematic literature review could be conducted within the scope of this thesis. Instead, a simple search strategy was designed with the intention of identifying a subset of the relevant literature, which could act as a foundation to identify further relevant studies through the use of snowballing and citation tracking (Bakkalbasi et al., 2006; Wohlin, 2014). Snowballing refers to using the reference list of a relevant study in order to identify any additional relevant publications, it therefore only identifies retrospectively published studies to the study searched. Citation tracking on the other hand is used to identify which papers cite the study, in order to find relevant papers that have subsequently been published.

Searches were conducted of the following databases: Web of Science, PubMed and the Cumulative Index of Nursing and Allied Health Literature (CINAHL). Searches were conducted originally from the date of their inception until July 2017, but were then updated in November 2023. The search terms used for each of these databases are shown in Table 2.1.

Table 2.1: Search terms used for the literature search.

Database	**Search Terms**	**Scientific discipline**
Web of Science	1. TITLE:(econ* AND replica*)	Studies specific to economics
PubMed	1. (reproducible[Title] OR transparen*[Title]) 2. research[Title] 3. 1 and 2	All studies
CINAHL	1. (AB replica*) AND (MH “Reproducibility of Results”)	All studies

Eligible studies were those that explicitly discussed what it meant for a replication to be successful, regardless of scientific discipline. Information was then extracted from the relevant papers. This included the general characteristics of the study (such as title, journal, publication year, funding source) as well as the scientific discipline, type of study (for example commentary, discussion on the concept of replication or if a replication attempt was conducted), the stated purpose of replication and how a successful replication was defined. Where the paper was related to health economics, additional information was extracted. If a replication was conducted, this included the type of study replicated, and if it were a model, the model type, modelling software used, if contact with the original authors was made, and the outcome of the replication attempt and whether the replication was defined as a success.

Using the results from the literature review, a variety of definitions were then proposed, constructed to be specific to health economic modelling. Importantly, such definitions needed to reflect the unique aspects of decision modelling. For example, models commonly evaluate multiple treatment or disease pathways. Therefore a definition would need to specify whether the replication should refer only to the base case of the analysis or if it should also extend to the sensitivity analyses carried out. Another aspect to consider is that there are multiple types of models used within health economics, which can vary in complexity. It may be reasonable to accept a lower standard of replication success if the model is more complex, which would require that the definition of replication success be proportionate to the complexity of the original model, whilst also taking into account the motivation for the replication. These motivations may range from attempting to replicate the structure of an existing model to facilitate the development of a new model, or for replication's sake, where the definition of success is likely to be a lot stricter. In this chapter replication is considered for replication's sake, that is, as a means to assess the transparency of model reporting.

2.3 Results

2.3.1 Definitions of 'Successful Replication' from other scientific disciplines

The literature review yielded many studies discussing the concept of replication, however substantially fewer were found exploring the concept of what makes a replication *successful*. Indeed, it was reported by the Open Science Collaboration that there is "no single standard for evaluating replication success" (Open Science Collaboration, 2015). This statement is supported by the definitions found within this review, detailed in Table 2.2, which range from subjective assessments to expecting exactly the same results to be reproduced. The definitions found, along with the objectives for completing the replication, were split according to whether they were replications of data analyses or seeking to replicate an observed effect (primarily conducted within psychology), the latter of which appeared to focus more on statistical significance. All definitions referred to a narrow replication, that is, they were seeking to reproduce the original results using the same methods and data.

Although not a systematic literature search, 13 definitions were found, with most including some form of subjective assessment to determine success and only one suggesting use of a statistical test. The majority of these definitions were considered opaque, with sufficiently loose wording to allow most studies to be considered successful if the right interpretation was taken. Indeed, Chang et al, stated that they deliberately chose a loose definition to determine an “upper bound on what the replication success rate could potentially be” (Chang & Li, 2015). A lack of formal definition may be due to the simplicity of the term itself and perhaps the perception that it is obvious what a ‘successful replication’ is.

Several of the replication studies relied on data or code provided by the original authors (Chang & Li, 2015; Hardwicke et al., 2017; Peng, 2009). This may have been due to convenience (in that they sought to replicate a large number of studies), but it might also suggest that the replication was more a test of open data policies and the usability of provided materials, rather than testing how thoroughly the methods were reported within the manuscript.

As well as the studies that explicitly defined success, several replications were found that made a judgement as to whether the replication was a success or failure. Whilst these studies failed to articulate exactly how a successful replication was defined, inferences can be made from the concluding judgements about aspects of the definition used. The first study conducted by Jones and Ziebarth (2016), reported a successful replication within applied econometrics, stating: “We were able to replicate Levitt’s (2008) findings *almost exactly*” (Jones & Ziebarth, 2016). The second, detailed three independent replications of a study within psychology (thus looking to replicate an observed effect), stated that their replication was a failure, given that, “The difference is in the *opposite direction* to that predicted” (Ritchie et al., 2012).

Table 2.2: Definitions of 'Successful Replication' found.

Study	**Scientific Discipline**	**Objective**	**Methods**	**Definition of ‘Successful Replication’**
DATA DRIVEN – REPLICATING ANALYSIS				
McCullough et al. (2006)	Economics	Exploring whether data and code depositing requirements of the Journal of Money, Credit, and Banking are being followed, and therefore allow replications.	Replications conducted using the original study’s data and code.	“Successful replication … refer to duplicating *all the results*.”
Peng (2009)	Biostatistics	To establish a reproducible research policy for the journal Biostatistics.	Replications conducted using the data and code provided by the original authors.	“An article is designated as reproducible if the AER (Associate Editor for reproducibility) succeeds in *executing the code* on the data provided and *produces results matching* those that the authors claim are reproducible. In reproducing these results, reasonable bounds for numerical tolerance will be considered.”
García (2014)	Economics	To verify the results of a previously published study. The original article was referred to as highly cited and relevant to ongoing policy debate.	The replication used the original study’s own data and the methods reported in the manuscript.	“A pure replication is successful if *the exact same results* reported in the original study, including any errors and omissions, can be reproduced using the inputs in the replication file.”

Table 2.2: Definitions of 'Successful Replication' found. *(Continued)*

Study	**Scientific Discipline**	**Objective**	**Methods**	**Definition of ‘Successful Replication’**
Chang and Li (2015) *Preprint* Chang (2017a) † *Published*	Economics	To broadly evaluate the state of replication within economics.	Where possible replications were conducted using data and code provided by the original authors. If not provided, they checked the personal websites of each of the authors for replication files, and failing this contacted authors via email.	“We define a successful replication as when the authors or journal provide data and code files that allow us to *qualitatively reproduce the key results* of the paper.” “For example, if the paper estimates a fiscal multiplier for GDP of 2.0, then any multiplier greater than 1.0 would produce the same qualitative result” “Defining replication success as our ability to use the author-provided data and code files to produce *the key qualitative conclusions* of the original paper”
Chang (2017b) *Preprint*	Economics	A pre-analysis plan detailing the steps to be carried out during a planned replication.	The replication would use the original, raw data and the replicator would produce their own code with the intention of reproducing the analysis reported.	“I would be “successful” if I was able to *replicate the Figures* that I pre-specified (1, 2, 3, 7, and 8) *to a reasonable degree of accuracy*.”

Table 2.2: Definitions of 'Successful Replication' found. *(Continued)*

Study	**Scientific Discipline**	**Objective**	**Methods**	**Definition of 'Successful Replication'**
Chang (2018)[†] *Published*				"I would be successful with this replication if there were *no visual difference* between any replicated figure and published figure and if the observation count cited on page 424 (190,778 observations) *matched the observation count* in the replication."
Hardwicke et al. (2017)	Cognitive Science	To more broadly examine whether open data policies are being adhered to such that: data is available, it is in a usable form and if so, that reported outcomes can be reproduced.	The replication used author provided datasets and any computer code that were used to produce published results.	"If there are only *Minor Numerical Errors*, or *no discrepancies*, then the reproducibility check is considered an overall success"
ReplicationWiki (Höffler, 2018)	Economics	To facilitate replications via an online database. Suggests studies for replication and details the results of replication attempts.	No formal methods given other than that replications should aim to repeat the reported analyses.	"Successful: *Results could be replicated without major deviations* from the published results Partially Successful: Key results could be replicated but some deviations from published results

Table 2.2: Definitions of 'Successful Replication' found. *(Continued)*

Study	Scientific Discipline	Objective	Methods	Definition of ‘Successful Replication’
				Failed: Key results could not be replicated, significant deviations from the published results.”
EXPERIMENTAL – REPLICATING AN OBSERVED EFFECT				
Brunner and Schimmack (2016)	Psychology	To explore possible proxy measures to conducting full replications.	Not applicable (looking at alternatives to full replications).	“Show that the description of the original study was *sufficiently precise to reproduce the study* in a way that it successfully *replicated the original result*.”
Patil et al. (2016)	Psychology	Exploring the definition of replication success, to take into account that observed effects will have natural levels of variation.	Not applicable (looking at statistical likelihood of replication results).	When the *95% prediction interval* for the effect size estimate of the replication study (calculated using the original study effect size) *includes the actual point estimate from the replication*.
Cova et al. (2018)	Philosophy	To broadly evaluate the reproducibility of studies within experimental philosophy.	Replications followed the design and methods of the original studies as closely as possible.	“Three different methods for designating a replication attempt as a success or a failure… (a) Were the replication results *statistically significant**?

Table 2.2: Definitions of 'Successful Replication' found. *(Continued)*

Study	**Scientific Discipline**	**Objective**	**Methods**	**Definition of ‘Successful Replication’**
				(b) *Subjective assessment* of the replicating team. (c) Comparison of the *original and replication effect size*.” *Statistical significance was defined as a p-value less than 0.05.
PsychFileDrawer, (Pashler et al.)	Psychology	An archive of replication attempts, intended to facilitate the reporting of replications and the discussion of their results.	All replications sought to reproduce the original results. However it is not clear if the methods used were exactly the same, with reported replications varying from “Highly exact/direct replication” to “Fairly exact /direct replication”.	“A very *pronounced trend or a significant difference* in the *same direction as a published* result deserves the characterization of a "successful replication".”

Note:

† These papers are published versions of the preprints highlighted in the table and are included here due to the slight change in definition of replication success.

2.3.2 Definitions of ‘Successful Replication’ used in Health Economics

A small number of studies reporting replication attempts of decision models within health economics were found (Bermejo et al., 2017a, 2017b; McManus & Sach, 2017; McManus, Turner, Gray, et al., 2019; A. J. Palmer et al., 2018; Schwander et al., 2021; Smolen et al., 2015). Notably these studies are those that have conducted replications and published them for replications sake. It is acknowledged that there are likely to be a number of replications conducted pragmatically and not reported, instead being used as a stepping-stone to inform the development of new models.

The concept of a successful replication was explored at the Eighth Mount Hood Diabetes Modelling Conference (2016), where modelling teams were set the challenge of replicating two published simulation models. Here, the replication challenge was used to indicate transparency and therefore in their proposed definition of replication success, Palmer et al. speak in terms of model transparency:

> “A simulation would be regarded as transparent if one of the users was able to produce a set of instructions of the simulation they undertook that was sufficiently detailed and comprehensive to allow the other user to implement them and *produce identical results* using the same model” (A. J. Palmer et al., 2018).

In contrast to the above, Bermejo et al. (2017a) conducted a study evaluating the replicability of five decision models within health economics, varying from Markov models to simulations. These replications were conducted using only the information presented in the original publications and any online supplementary materials. Whilst in the original manuscript an explicit definition of replication success was not given, when answering “Was full replication successful?” two out of five of the models were labelled with “yes”. In response to further academic debate (McManus & Sach, 2017), Bermejo et al. subsequently stated that they defined a successful replication as the following:

> “We considered that the replication was successful if: (a) *all the necessary information to replicate the model was available*, and (b) if *the results were not significantly dissimilar* from the original reported model results” (Bermejo et al., 2017b).

The latter clause of this definition allows some variation between original and replicated results, which might, inherently be expected, for example with probabilistic modelling, or issues with rounding error or a different software being used. However, it brings further questions as to how much variation should be considered and what should be deemed a significantly different result (it was not stated in the paper whether the term 'significance' related to statistical significance or otherwise). Whilst it is easy to suggest that if an alternative intervention was found to be cost-effective this would be considered significantly dissimilar, there is less clarity, for example, about how much costs or outcomes would need to vary before a result was considered significantly different. The two models considered to be successfully replicated by Bermejo et al. varied from the original incremental cost-effectiveness ratios (ICERs) by "all around 17%" within one and by "9.6% and 1% lower than the reported ICERs" (Bermejo et al., 2017b) in the other, whilst unfortunately the variation of the results from the failed replications were not reported.

Another study found within the literature review that specifically considered replication within health economic decision models, was conducted by Smolen et al. (2015) who replicated a Markov model. Whilst this study did not provide an explicit definition of what they considered a successful replication to be, they did state that their replication attempt was a success citing the small percentage differences obtained between the original results and those replicated, "Table 1 results *indicate success*" (percentage difference varied between -1.92% to 1.15%). This suggests that to those authors, the concept of successful replication may be closely aligned to the definition proposed by Bermejo et al. (2017b). Interestingly, when the authors were contacted about their replication attempt, they stated that they had used a pixel counting software to derive the transition probabilities used in the model which they described as a 'painstaking process', as they had not been conventionally presented in a table, but had instead been presented graphically (Figure 2, "Visual representation of transition probabilities" Batty et al. (2013)). When considering the definition proposed by Bermejo et al., outlined above, it is questionable if by presenting the transition probabilities in this way, others would have perceived this as providing all the necessary information.

Although only a limited number of studies within health economics were found to propose an idea of what constitutes a successful replication, commonalities existed in the expectation that all the necessary information was provided and that the original results were regenerated to some degree. There is need to develop a consensus on how much variation should be permitted. Whilst it may be easier to expect an exact replication of results (as with the A. J. Palmer et al. (2018) definition), it is also pragmatic to expect minor variations between replicated results and those reported in the original manuscript. These variations could be purely due to potential rounding of key results, probabilistic terms within the model or replicator error; and it seems reasonable that these should be accounted for within any definition proposed. Furthermore, it is practical to expect that the complexity of the model is also likely to impact on how exactly the results can be replicated, for example comparing the expectations of a successful replication of a simple decision tree to the results of a complex Markov model which employs a lifetime time horizon.

Finally, another study by Schwander et al. (2021) was found, however this was published after both chapters 2 and 3 were published and it therefore uses the definitions proposed within this thesis. This paper is discussed further in Chapter 3.

2.3.3 Proposed definition of 'Successful Replication' for health economic models

Given the results of the literature review, it is evident there is no consistent definition of what constitutes a 'successful replication' in other disciplines, nor has a definition been found that could directly apply to decision modelling. Therefore, there is a need to construct a definition that can be used to evaluate the replicability of decision models in health economics, holding models accountable to a set standard of reporting and transparency. Importantly, such a definition should be usable and not deliberately opaque.

Due to the above, it may not be appropriate to have a single 'one size fits all' definition, but instead to have a staged definition with levels of success, like the definition proposed by PsychFileDrawer (Pashler et al.) which include: successful, partially successful and failed. Although as of 2023, this website is no longer operational. By using staged definitions, this may also allow the definition to change depending on the different motivations for conducting the replication. For example, to evaluate the overall reporting transparency (and thus conduct a replication for replication's sake) it may be that a stricter definition is enforced. Whereas with pragmatic replications, carried out to extend the applications of an existing model, it may be acceptable for the replicated results to only loosely match those of the original, such that they have the same qualitative result (i.e. cost-effective or not).

In Table 2.3, several definitions are proposed which give a broad to narrow definition of success. These definitions were informed by the definitions found within the literature review (detailed in Table 2.2) and amended to be specific to decision models. Alongside each of these are the potential strengths and weaknesses of using such a definition.

Table 2.3: Proposed definitions of replication success for decision models, ranging from broad to narrow.

Example definition	Strengths	Weaknesses
The same conclusions for intervention cost-effectiveness were reached.	• This broad definition allows a lot of variability, allowing for potential rounding errors, different software and so on (which may naturally be encountered within a replication). • This definition is not dependent on the outcome of the original model producing an ICER.	• Does not necessarily reflect on reporting quality of the model. • Models reporting results close to the willingness to pay threshold may be subject to tighter constraints. For example if the original model reported a cost per QALY way above the threshold (e.g. £1 million) then any replication result greater than £30,000 would be equivalent to success, whilst potentially allowing a great deal of variance. In comparison, a model reporting a cost per QALY of £29,000 would be permitted significantly less variability for the replication to be deemed successful.
The calculated ICER varies by only XX% compared to the original.	• This would be useful, as long as there was an ICER to compare to. • As above, the definition also allows some inherent variability.	• ICERs may not always be reported, if for example an intervention is dominant. • Permitting ICER variation may allow for contradictions to the original study's conclusions if there was variation close to the cost-effectiveness threshold.

Table 2.3: Proposed definitions of replication success for decision models, ranging from broad to narrow. *(Continued)*

Example definition	**Strengths**	**Weaknesses**
Costs and outcomes replicated for some treatment pathways/model scenarios and not others.	• This definition incorporates the idea of a 'partial' success, and is probably the most likely outcome from attempted replications within modelling. • If the replicator is able to identify why some pathways/scenarios were replicable and not others, it may help to inform reporting guidelines.	• This definition may raise more questions than it answers, for example how many pathways/scenarios would need to be replicated for it to be considered a success?
Cost-effectiveness figures could be reproduced to a reasonable degree of success (for example the cost-effectiveness acceptability curve)	• Allows for some variability, but ensures that the general trends of the results are the same.	• Figures may not be produced, or provided in sufficient resolution to facilitate proper comparison. • Some figures, such as the cost-effectiveness plane, might be difficult to visibly see the scale of differences between the original and replicated bootstrap pairs.
Results for the costs and outcomes vary by only XX% compared to the original, *AND* are consistent with the original conclusions.	• This definition would allow for some inherent variability, whilst still being strict enough so that a replication deemed as 'successful' would be informative. • Similar to the definition proposed by Patil et al. (2016), the variability could be set so that the replicated results lie within the xx% confidence	• The variation permitted is likely to be arbitrary.

Table 2.3: Proposed definitions of replication success for decision models, ranging from broad to narrow. *(Continued)*

Example definition	Strengths	Weaknesses
	intervals or in turn x number of standard deviations from the original result.	
Identical results are produced.	- A very narrow definition, but one that suggests, if a replication is successful, that the original model is well reported and free of calculation errors.	- This definition is less informative if the replication fails, as it does not account for replicator error, probabilistic modelling or potential rounding errors. - Whilst holding the highest of standards, it is not pragmatic.

2.4 Discussion

This is the first literature review to identify how replication success has been defined across all studies considering the concept of replication, regardless of scientific discipline. It is also the first to present a review of the replication literature, tailored to a health economic audience, going on to consider how such definitions may apply to replications of health economic decision models.

This review has provided examples of how ‘successful replication’ is defined in other disciplines, but has also highlighted the lack of workable definitions that can be applied to health economic models. Given the extensive literature discussing replication and the relatively smaller number of definitions proposed, it seems there may be a reluctance to label a replication as a success or failure. This may be due to a fear of how the verdict will be received, potentially damaging research reputation and alienating colleagues. The reputation of the replicator is also at stake, given that an inability to replicate might reflect their modelling skill. It is also important to clarify what the implications of a successful or failed replication are. For example, whilst being able to replicate a model may imply that it is transparently reported and usable, it does not necessarily imply that it is accurate to the clinical condition it represents or that the underlying assumptions are valid. Likewise, whilst a failed replication may indicate a lack of reporting clarity or the use of proprietary data, it does not implicitly suggest that the model is completely invalid, or that there is a deliberate intent to mislead.

With these consequences clearly outlined, any definition constructed requires specificity and exactness to allow the definition to be informative, but should also be pragmatic, allowing for marginal variation between the replicated and original results, so that it is usable. The definition may also depend on the original purpose of the replication, or the context. For example, there are a number of reasons why models might be replicated, whether it is to be used as a springboard for developing a new model, to ensure transparent reporting or indeed to check that the results are free from calculation errors. The process of carrying out a replication and interrogating published models may also be a valuable learning exercise in itself for the modeller. How a successful replication is then defined, for example if the purpose of the replication is to check another’s work (where little variation in results would be expected), may vary in comparison to if the aim of the replication was to understand and re-purpose an existing model (where differences in results would be less of an issue). The importance of context on how a successful replication is defined is also echoed by Chang (2018), who suggest that the replication context may influence the amount of money and time a researcher is willing to devote to a replication effort.

Whilst several definitions have been proposed within this paper (informed by the breadth of definitions found within the review), it should be the wider health economics community that dictate the final definition and the standard that should be expected within economic models, to facilitate buy-in and ensure that the standards are practicable. Particularly, there is need to consider how much variation is permissible between the replicated and original models.

The role of journals in facilitating replication also needs to be considered, as there are often strict word limits, which might restrict the amount of detail than can be given about modelling methods, although it is acknowledged that these restrictions may be offset with allowances for supplementary materials. The importance of journals facilitating and advocating the inclusion of supplementary material is highlighted in the recent transparency guidelines relating to the Eighth Mount Hood Challenge. In this paper, which sees the development of a checklist for reporting the inputs of diabetes simulation models, the authors argue that journals could go beyond merely allowing the checklist to be published alongside the manuscript, but instead to mandate the populated checklist in order for the modelling study to be published (A. J. Palmer et al., 2018).

Journals may also be reluctant to publish replication attempts (as highlighted by other disciplines (Duvendack et al., 2017; G. Martin & Clarke, 2017; Ryan-Wenger, 2017)), however this is currently under explored within health economics, given the lack of formal replications being conducted. Although there are some journals that have stated replication studies will be considered for publication and others, such as the American Economic Review journal who have even published statements dedicated to replication and the standard it expects its manuscripts to uphold (American Economic Association, 2020).

It is also important to state that not all researchers consider replication as a valuable endeavour. In example of this, Bermejo et al. (2017b) stated that replication of health economic decision models is an “inefficient use of time”, although these authors considered replication mainly for the purpose of quickly developing a new model, as opposed to a means to evaluate transparent reporting. Whilst conducting a replication can be a time consuming process, and as acknowledged above there are limitations to what the results of a replication can tell researchers, it is believed that replication studies are able to provide invaluable insights into how models should be reported to enhance transparency and to improve future methods. Furthermore, whilst the concept of transparency and model replicability is a much broader topic than simply replicating published results, it is argued that it is a logical starting point for a body of further research in this area. It is hoped that through the development of an accepted standard, replication testing could act as a catalyst to promote changes in the way modelling studies are presented. Importantly this will also indicate how well currently used checklists (such as CHEERS (Husereau et al., 2022; Husereau et al., 2013) and Philips (Philips et al., 2004)) are at identifying reporting thoroughness and may also provide further evidence to support the multiple calls for modelling registries (Arnold & Ekins, 2010; Sampson, 2012; Sampson & Wrightson, 2017). There may also be scope to further the application of such standards.

2.5 Strengths and limitations

This is the first review which sought to identify definitions for replication success, and the first to consider the role of replication within health economics. The fact the literature search was not restricted to any scientific discipline may mean that a greater breadth of literature was captured than what would otherwise have been captured in discipline specific reviews. In finding explicit definitions of replication success it is possible to consider how replication exercises have been judged in the past. Furthermore, in adapting these definitions in a manner that can apply to health economics, these can be used to inform and encourage debate of how replication within health economics might be judged going forward.

With that being said, there were also some limitations. Firstly, this was not a systematic literature review, which may mean that some relevant publications were missed from inclusion. This was due to the fact that the search strategy was difficult to refine due to the nebulous search terms, and the wide scope of disciplines searched. The search was also restricted to English language articles, which again may have resulted in some relevant studies being omitted. As well, this chapter only considered replication in the context of decision modelling, and therefore definitions applicable to other areas of health economics such as economic evaluation alongside clinical trials, have not been proposed.

2.6 Conclusion

This review has shown that there is currently no universal definition of a ‘successful replication’ being used, or a definition that could directly be applied to health economic models. Community buy-in is needed to develop an accepted definition and therefore to establish a standard of transparent reporting that can be adhered to. Without which, it is difficult to assess the replicability of decision modelling within health economics and therefore to measure how well the discipline is doing in meeting transparency and reporting standards.

Chapter 3 Barriers and facilitators to model replication

3.1 Introduction

The previous chapter explored how replication success is currently defined, and used these definitions to suggest how they might in turn be applied to health economic decision modelling. Now that what researchers may strive to do in order to achieve replication has been evaluated, it is time to appreciate the practicalities of carrying out a model replication, and to explore the barriers and facilitators that may help or hinder future model publications from being replicated.

Currently, there are few published studies looking at the actual replicability of models (Bermejo et al., 2017a; A. J. Palmer et al., 2018; Schwander et al., 2021; Smolen et al., 2015). One of which, was published by the collaborative diabetes modelling group, Mount Hood. This publication detailed the results of their 2016 conference which focused on ‘Research Transparency’ and set modellers the challenge of replicating two diabetes simulation models (Mount Hood Diabetes Challenge, 2016). In response to the difficulties encountered during this challenge, Mount Hood published ‘the Diabetes Modeling Input Checklist’ which was designed to facilitate the reporting of diabetes model inputs (A. J. Palmer et al., 2018). This has since been used to varying degrees in publications detailing diabetic simulation modelling studies: one paper stated its use to inform the write up of the study, but did not include it as supplementary material (Dakin et al., 2020), whilst others have provided the completed checklist in supplementary material to the publication (Johansen et al., 2019; Shao et al., 2019). However, there have yet to be any publications evaluating the effectiveness of this checklist in increasing transparency, or commentaries on its ease of use. More generally, decision models are subject to quality and reporting checklists. A review of the use of checklists within systematic reviews of economic evaluations identified 18 unique checklists that had been used in the literature since 2010 (Watts & Li, 2019). Amongst the most commonly used was the CHEERS checklist (Husereau et al., 2022; Husereau et al., 2013) and the Philips criteria (Philips et al., 2004).

The CHEERS checklist was developed by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force, which was established in 2009 with the purpose of developing guidance to improve the reporting of economic evaluations. The task force conducted a systematic review of existing reporting guidelines and used the results of this review along with a two-round modified Delphi panel to identify the minimum set of items required when reporting economic evaluations (Husereau et al., 2013). After which, a 24-point item checklist was developed. Items relate to the reporting of elements within the study design as well as sources of funding and conflicts of interest. Importantly, this checklist is not specific to publications describing decision models, and as such only a small number of the checklist items directly relate to modelling components (items: 13b, 15-18 and 20b in the 2013 version). The CHEERS statement was originally published simultaneously by 10 journals when it was first introduced, although it is now more widely endorsed (Equator Network, 2020). Since its original conception, the CHEERS statement has now been updated (Husereau et al., 2022) with an added focus on transparency and the incorporation of public involvement in the research design. Most notably, the CHEERS checklist now includes a sentence within the modelling item relating to reporting if the model is publicly available and where it can be accessed.

In contrast, the Philips criteria was developed specifically to provide a consistent framework to assess the quality of decision models. The framework was developed following a review of existing guidelines for modelling studies, with three key themes: Structure, Data and Consistency, incorporating in total around 56 reporting criteria (note that some criteria incorporated multiple questions within one).

Elsewhere, Bermejo et al. sought to reproduce five modelling studies published in the journal of Pharmacoeconomics between the period of August and November 2016. This paper has been discussed in Chapter 2, but they concluded that the majority “could not be fully replicated” (Bermejo et al., 2017a). This paper had several limitations, in that they failed to provide the exact results of their replications, failed to define how they judged a ‘successful’ replication and also failed to provide clarity on how the models replicated were chosen, beyond being published in the stated time period.

Smolen et al. (2015) detailed a replication attempt in a conference proceeding presented at ISPOR. They replicated a Markov model, which evaluated treatments for chronic migraine from the perspective of the NHS. This model had a time horizon of two years, with 12 week cycle lengths (Batty et al., 2013); and unlike the replications detailed by Mount Hood and Bermejo et al., their motivation was pragmatic, in that they wanted to build upon the existing model in order to develop a model for a similar disease area. Smolen et al. (2015) concluded that their replication was successful, reporting differences in terms of incremental costs and QALYs of between 0.00% and 1.92% for the replication of the base case analysis.

Informed by these existing studies, this chapter seeks to explore how replicable existing published decision models are, and to identify any barriers and facilitators that may make replication more or less feasible. In doing so, it is hoped that the future reporting of modelling studies can be adapted to facilitate replication. This chapter also seeks to evaluate the success of the replication case studies, using the definitions proposed in the previous chapter. This will inform the workability of the definitions proposed.

3.2 Methods

Five published modelling studies were selected to be replicated, with each of the replication attempts detailed as case studies. The justification for selecting the original models varied. For case studies 1 to 3, each of the models were identified through a systematic review of models within that disease area. The models selected for replication were judged to be the most thoroughly reported (in terms of the most number of Philips checklist criteria satisfied (Philips et al., 2004)) and were therefore considered to be the most likely replicable. The final model replications, case studies 4 and 5, detail replications that were conducted due to needing to develop a model for a similar decision problem. The original model detailed in case study 4 was chosen as it had characteristics and health states that made it suitable for adaptation to address a specific research question. Whereas the model detailed within case study 5 was selected as it was the most up to date and contextually relevant at the time of replication.

The primary focus of the replication attempts was to recreate the key findings published, such as total costs, outcomes and if applicable, the ICER, for the base case analyses. Replications were conducted using only the information presented in the referenced publications, and where possible, the same modelling software as in the original publication was used. Replication attempts did not extend to any sensitivity analyses reported, however for case study 5 (one of the pragmatic case studies), probabilistic sensitivity analysis was replicated, and the results of this are discussed. The degree to which the original results varied in comparison to the replicated results were calculated as a percentage difference on a per patient basis, to facilitate comparison across the case studies and to an existing replication study, who also reported results in this manner (Smolen et al., 2015).

To ensure that the results of the replications conducted were generalisable and not merely reflective of one individual’s modelling expertise (or lack thereof), each replication was conducted by a different analyst; case study 1: was completed by myself, case study 2: Toochukwu Okoli; case study 3: Haseeb Khawar; case study 4: David Turner; and case study 5: Ewan Gray. Both Toochukwu Okoli and Haseeb Khawar completed these replications as part of their dissertation in MSc Health Economics, which were supervised by Prof Tracey Sach and myself. In contrast, both David Turner and Ewan Gray were contacted to be included in this chapter of work as I was aware that they had both previously attempted to replicate a decision model. It is important to note that the publications chosen for replication were not selected to intentionally single out individual authors, journals, or institutions.

It was considered that the five replicators covered a range of modelling expertise, from newly graduated in health economics, to being fully established within the modelling field. Although four of the replications were conducted by other researchers, the case studies presented are written up in their entirety by the doctoral candidate, using responses from the replicators to set questions, which are shown in Table 3.1. The type of model replicated, the journal the research was published in and the funding source of the research is also noted.

Table 3.1: Questions asked to inform the write up of the replication case studies.

1. Please give the reference to the model you replicated and a brief description of what it was evaluating.
2. Why did you choose the treatment pathway(s) you replicated?
3. Did you use any supplementary material or contact the author for clarifications when conducting the replication? If yes, please provide details.
4. Would you consider your replication attempt to be successful? Why/Why not?
5. Were there any difficulties in replicating the model?
6. If yes, what were they and what could have helped to have alleviated these? (E.g. in the presentation of the model, or providing supplementary material)
7. What things helped you in your replication attempt? (E.g. example diagram of the model, or clear table of inputs)
8. Did you have to make any assumptions when you replicated the model? If so what were they and how did you inform these?
9. Looking back on the Philips criteria, and aspects that your model satisfied/did not satisfy, do you think it is a good indicator of whether the model is likely to be replicable or not?

As with the manuscripts in case studies 1 to 3, the Philips checklist was also completed for case studies 4 and 5, with the intention of exploring whether any barriers and facilitators identified in the replications could be related to items within the checklist. For each of the criteria, a subjective response of ‘yes’, ‘no’, ‘partial’ or ‘not applicable’ was given. I completed the checklist items for all case studies along with the case study replicator and in case study 1 with Prof Tracey Sach. Where any disagreement occurred, this was resolved by reviewer discussion, although for any item where it was unclear if the criteria was satisfied, the reviewers erred on the side of caution and responded ‘No’. The Philips checklist responses were then used in conjunction with the barriers identified in the model replications. It was reasoned, that if the barriers to model replication were not picked up within the checklist or indeed that they were satisfied, this might suggest that the existing reporting criteria are insufficient to ensure research transparency and reproducibility. To facilitate this, items of the Philips checklist thought to be particularly relevant to reporting thoroughness were highlighted and the responses to these items were focused on.

The results of the replications were then compared to the definitions proposed in Chapter 2, to explore the workability of such definitions and to gauge how successful each of the replications were.

Since publishing the work documented in both this chapter (McManus, Turner, Gray, et al., 2019) and that of Chapter 2 (McManus, Turner, & Sach, 2019), a subsequent paper was published by Schwander et al. replicating four decision models within the clinical area of obesity, with the aim of identifying replication facilitators and barriers, as well as to evaluate the definitions of success proposed in Chapter 2 (Schwander et al., 2021). All of the replications within this paper were conducted by one researcher using the modelling software, TreeAge (TreeAge Pro, 2021). In the following section, I also discuss and examine the findings of this paper, to identify any areas of similarities and differences, as well as to highlight another author’s perception of applying the ‘successful replication’ definitions.

3.3 Results

The general characteristics of the models selected for replication can be found in Table 3.2, along with the results of the replication attempts compared to the published results, in Table 3.3.

Table 3.2: Characteristics of the models replicated.

Case Study	Selected by	Model Type	Disease	Population	Perspective	Software	Time Horizon (Cycle Length)	Health Outcome	Results
1 Garside (2005)	Systematic review and Philips Criteria	State-transition, Cohort	Eczema	Adults with mild to moderate facial eczema	NHS	Microsoft Excel	1 year (4 weeks)	QALY	Corticosteroid dominant
		State-transition, Cohort	Eczema	Children with moderate to severe body eczema	NHS	Microsoft Excel	14 years (4 weeks)	QALY	ICER: £14,175 for tacrolimus 2nd line treatment compared to no tacrolimus
2 Dean (2001)	Systematic review and Philips Criteria	Decision Tree	Erosive Reflux Oesophagitis	“Ambulatory care patients”	Third-party Payer	Data TreeAge	1 year (Not applicable)	Percentage of symptomatic recurrences prevented	Rabeprazole dominant

Table 3.2: Characteristics of the models replicated. *(Continued)*

Case Study	**Selected by**	**Model Type**	**Disease**	**Population**	**Perspective**	**Software**	**Time Horizon**	**Health Outcome**	**Results**
3 Affleck (2011)	Systematic review and Philips Criteria	State-transition, Cohort	Psoriasis	Adults with moderately severe scalp psoriasis	NHS in Scotland	Microsoft Excel	1 year (4 weeks)	QALY	TCF gel dominant
4 Chambers (1999)	Model development	State transition, Cohort	Stroke	Stroke survivors	Health and social care	Data TreeAge	5 years, extended to 25 years (3 months)	Number of strokes, life years	Aspirin dominant
5 Ganesalingam (2015)	Model development	Decision tree State-transition, Cohort	Stroke	Adults suffering acute stroke	NHS	Software not stated	20 years (3 months)	QALY	ICER: £7,061 for mechanical thrombectomy compared to usual care
Abbreviations: TCF; Two-compound formulation.									

Table 3.3: Results of the replications compared to the published results, transformed to per patient to facilitate comparison.

		Results (Per patient)					
		Cost			Health Outcome		
Case Study	**Scenario Replicated**	**Original**	**Replication**	**Difference (%)**	**Original**	**Replication**	**Difference (%)**
1	Adults, no pimecrolimus (Base Case)	39.39	38.27	-1.12 (-2.84%)	0.968	0.968	0.000 (0.00%)
	Adults, pimecrolimus (2nd-line treatment)	70.58	79.69	9.11 (12.91%)	0.961	0.964	0.003 (0.31%)
	Adults, pimecrolimus (1st-line treatment)	135.44	157.78	22.34 (16.49%)	0.967	0.967	0.000 (0.00%)
	Children, no tacrolimus (Base Case)	956.47	1,989.44	1,032.97 (108.00%)	1.085	1.060	-0.248 (-2.29%)
2	Rabeprazole	1,414.00	1,431.00	17.00 (1.20%)	86.000	86.000	0.000 (0.00%)
	Lansoprazole	1,671.00	1,597.00	-74.00 (-4.43%)	68.000	68.000	0.000 (0.00%)
	Omeprazole	1,599.00	1,581.00	-18.00 (-1.13%)	81.000	81.000	0.000 (0.00%)
3	TCF gel as first-line therapy (Base Case)	241.86	230.89	-10.97 (-4.54%)	0.782	0.785	0.003 (0.37%)
	BMV as first-line therapy	255.12	255.29	0.17 (0.07%)	0.780	0.783	0.003 (0.40%)
4	No treatment (5 year time horizon)	15,093.00	14,955.00	-138.00 (-0.91%)	3.911	3.981	0.070 (1.79%)
	Aspirin (5 year time horizon)	14,817.00	14,717.00	-100.00 (-0.67%)	3.918	3.989	0.071 (1.81%)
	No treatment (25 year time horizon)	24,881.00	25,858.00	977.00 (3.93%)	7.607	7.585	-0.022 (-0.29%)
	Aspirin (25 year time horizon)	24,491.00	25,503.00	1,012.00 (4.13%)	7.664	7.643	-0.021 (-0.27%)

Table 3.3: Results of the replications compared to the published results, transformed to per patient to facilitate comparison. *(Continued)*

		Results (Per patient)					
		Cost			**Health Outcome**		
Case Study	**Scenario Replicated**	**Original**	**Replication**	**Difference (%)**	**Original**	**Replication**	**Difference (%)**
5	IV-tPA (Base Case)	52,495.00	53,545.00	1,050.00 (2.00%)	3.790	3.795	0.005 (0.13%)
	Thrombectomy	64,757.00	65,656.00	899.00 (1.39%)	4.842	4.800	-0.042 (-0.87%)
Abbreviations: TCS; Topical corticosteroid. TCF; Two-compound formulation. BMV; betamethasone valerate. IV-tPA; Intravenous tissue-type plasminogen activator. % difference calculated using the following formula: ((Replication – Original) / Original) x 100%							

3.3.1 Replication Case Studies

3.3.1.1 Case Study 1: Garside et al. (2005), Replicator: Emma McManus

The model for case study 1 was selected from a systematic review of decision models evaluating preventions or treatment for atopic eczema (McManus et al., 2017). This review identified 24 models, of which the model by Garside et al. (2005) was deemed to satisfy the most Philips criteria (satisfying 83% of applicable criteria).

The model was a state-transition model developed using Microsoft Excel and was described in a Health Technology Assessment monograph (Garside et al., 2005), and was funded by the NIHR. It evaluated the cost-effectiveness of two different topical calcineurin inhibitors (pimecrolimus and tacrolimus) for the treatment of atopic eczema, from the perspective of the United Kingdom (UK) NHS, across eight scenarios. These scenarios were divided into population: adults or children, location of eczema: facial or body, as well as severity: mild to moderate or moderate to severe. The models that evaluated the treatment of adults covered a time horizon of one year, with four weekly cycles, whereas the child models covered a 14 year time horizon, also with four weekly cycle lengths. The Health Technology Assessment monograph satisfied the majority of Philips checklist criteria, with 83% of applicable criteria being satisfied. This monograph was submitted as evidence to the NICE committee appraising the use of tacrolimus and pimecrolimus for the treatment of atopic eczema (National Institute for Health and Care Excellence, 2015). The NICE committee concluded that topical tacrolimus and pimecrolimus were not recommended for the treatment of mild atopic eczema or for first-line use for atopic eczema of any severity, whilst tacrolimus could be used for second line treatment of moderate to severe atopic eczema.

For this case study, two scenarios were chosen for replication, which were thought to encompass the widest range of options evaluated within the original study: the first evaluated pimecrolimus for the treatment of mild to moderate facial eczema within adults, whilst the second evaluated tacrolimus for moderate to severe body eczema amongst children. The models were constructed using treatment states comprised of different disease severity mixes, allowing for the fact that different disease severities could receive the same treatment (whilst at the same time having different health utility scores).

The results of the adult scenario replication varied by -2.84% to 16.49% for costs, 0.00% to 0.31% for outcomes, and the same overall conclusion regarding cost-effectiveness was found (topical corticosteroids dominated). In contrast, whilst replicating the childhood scenario numerous additional assumptions were required in order to make the model functional. Due to this and the extended time horizon of 14 years, any differences between the original and replicated model per cycle were amplified, which resulted in the replicated model returning values that were far removed from the original results; with costs varying by 108% of those originally reported. As such, no further attempts to replicate the other treatment pathways within this scenario were made. Whilst the costs were far removed, the outcomes replicated were relatively close to the original values (varying by -2.29%), which may suggest that rather than outright replicator error, the variation may have resulted from misinterpretation of costing assumptions.

The main barrier encountered when conducting this replication was the way in which the multiple scenarios were presented, which were all based on modifications of a general model. Although the transition probabilities were given, they were done so for all of the eight different scenarios together, with no clear labelling as to which transition probabilities related to which scenario. Additionally, some of the transition probabilities were presented instead as the likelihood of patients being offered different treatments, having previously failed a treatment. This was further complicated by conflicting information being presented within the text and the transitions presented in the table. For example, when recreating the adult scenario, it was stated that following a failed treatment of low-potency steroids, the probability that a patient would receive pimecrolimus was 0.85, mid-potency steroids, 0.1, and high-potency steroids, 0.05. However, this conflicted with information in the text, stating that high-potency steroids were not a treatment option within this scenario. Therefore, this left a 0.05 probability to be allocated to a treatment with no description about how this should be done. An author of the original publication was contacted to provide clarification (in 2015), however they were unable to help, citing the time that had passed since the publication, and current workload.

Facilitators:

- Detailed diagrams of the model structures, including possible transitions between states.

Barriers:

- Instances where the text and tables conflicted with one another.
- Transition probabilities were grouped together for multiple scenarios, instead of being presented individually.

- Lengthy time horizons meant that any differences between the replicated and original model were amplified over time (which occurred in the childhood scenario).
- Authors were unavailable to provide clarification.

3.3.1.2 Case Study 2: Dean et al. (2001), Replicator: Toochukwu Okoli

This case study replicated a decision tree, modelling the use of proton pump inhibitors for the maintenance therapy of erosive reflux oesophagitis over a one year time horizon, from the perspective of third party payer (Dean et al., 2001). The country within which this evaluation was set was not explicitly stated, however it may be inferred as the United States, due to the source of estimates used within the sensitivity analyses. This model was selected from the results of a systematic review (unpublished) which sought to identify published models evaluating different management approaches in Gastro-Oesophageal Reflux Disease (GORD). This systematic review identified 17 published models, of which the paper by Dean et al. satisfied the most Philips criteria (69% of applicable criteria were met). This was the lowest percentage of applicable criteria met across the five case studies. The journal article detailing the decision model was published in the American Journal of Health-System Pharmacy, a medical journal covering all aspects of drug therapy and pharmacy practice specific to secondary care. The work was funded by research grants from Janssen Pharmaceutica, Zynx Health and Ovation Research Group.

The replication was conducted in Microsoft Excel, instead of TreeAge, which was used in the original study as this proprietary software was not available to the replicator. The manuscript included a figure clearly showing the tree structure as well as a table of the probabilities used. Indeed, 69% of the applicable Philips checklist criteria were met, with the majority of criteria that were not satisfied mostly pertaining to the justification of modelling methods, as opposed to reporting clarity. Consequently, it was possible to replicate the decision tree closely, with the replicated outcomes matching exactly to that of the original, and replicated costs differing from the original by -4.43% to 1.20% and the same cost-effectiveness conclusions being reached (Rabeprazole dominant).

Despite the simplicity of the decision tree, there were still some barriers to exact replication. These included discrepancies between the text and branch structure presented in the model diagram (which was later assumed to be purely descriptive) along with a lack of clarity surrounding how to cost the treatments used as maintenance therapy.

Facilitators:

- Straightforward model structure.

Barriers:

- Instances where the text and model diagram conflicted with one another.
- Ambiguity regarding how costs were attributed during maintenance therapy.

3.3.1.3 Case Study 3: Affleck et al. (2011), Replicator: Haseeb Khawar

The model using in this replication study was selected from an unpublished systematic review, which aimed to identify decision models evaluating treatments for psoriasis. The review found 35 published decision models, of which the Affleck et al. (2011) study was deemed to satisfy the most of the Philips criteria (with 76% of applicable criteria met). This study was published in Current Medical Research and Opinion, a medical journal, and was funded by LEO Pharma, a pharmaceutical company.

Affleck et al. described a 15 state, state-transition model, built to evaluate treatment approaches for scalp psoriasis using four weekly cycles over a one year time horizon, from the perspective of the Scottish NHS (Affleck et al., 2011). Treatment pathways one and five were chosen for replication, on the basis that one of these evaluated the intervention and the other evaluated current treatment. The model was described comprehensively with tables of the transition probabilities, utilities, and descriptions of the health states being provided. Additionally, a detailed diagram of both the model and the different treatment pathways being evaluated, was given. This is reflected in the number of Philips criteria that were judged to be satisfied, 76% of those applicable. This enabled the replicated outcomes to vary from the original publication by only 0.37% and 0.40% across the two pathways, whereas the costs varied by -4.54% and 0.07%. The replication produced the same cost-effectiveness conclusions as in the original publication; that is, treatment pathway one dominated treatment pathway five. Only minor barriers to replication were encountered, involving the way some of the costs, assigned to each of the health states were described. For example, it was stated that a “weighted average of treatment modalities” was costed, although the weightings for this, were not given.

Facilitators:

- A table was provided which detailed the different health states of the model at baseline, any assumptions as well as the possible transitions from each state.

Barriers:

- Ambiguity surrounding the “weighted average” used when calculating the cost of treatments.

3.3.1.4 Case Study 4: M. Chambers et al. (1999), Replicator: David Turner

Case study 4 details the replication of a state-transition model developed by M. Chambers et al. (1999) evaluating the use of aspirin for stroke survivors, from a broad health and social care perspective, within the UK. This model was originally published in the journal of Pharmacoeconomics, a health economics journal. This model was also described in a subsequent publication by M. G. Chambers et al. (2002) in the journal Value in Health, an health economics journal. Both of these publications declared funding by Boehringer Ingelheim, a pharmaceutical company. The original publication satisfied 79% of applicable Philips criteria. Due to the multiple publications, some of the values used in the replication attempt were from the subsequent publication. In the base case of the model, the time horizon was five years, however in other iterations; this was varied between two and 25 years. Attempts were made to replicate the results from both the five and 25 year analyses, using Microsoft Excel.

In the base case, costs were replicated to within -0.91% and -0.67%, and outcomes were within 1.79% and 1.81%, in comparison to the reported results. Increased variation in costs was seen when the time horizon was extended to 25 years, with variation of 3.93% and 4.13%, although outcomes remained close to the variation found in the base case, varying by -0.29% and -0.27%.

Whilst conducting the replication, there was uncertainty relating to some parameter values. This was due to the table giving a range for each of the parameters, instead of listing individual values for each of the time points. Whilst this simplified reporting, it was unclear as to what value was used in particular cycles. Additionally, some values were reported with limited numbers of decimal places. In the model replication, total long-term costs over 25 years were overestimated by approximately 4%. Although total estimates of life years were very similar; there were small discrepancies as disabled life years were overestimated and disability free life years were slightly underestimated. Long-term care costs were the largest cost and estimates per cycle were much higher for disabled stroke survivors. Given the overestimate of disabled life years (from the original model) and the higher costs associated with this state, it is likely that this small discrepancy in overestimating disabled life years accounted for the additional estimates in cost. As a result, very small discrepancies in the number of individuals in disabled states had the potential for much larger discrepancies in expected costs.

Facilitators:

- Tables detailing how the main costs were derived, as well as a complete table of costs entered in the model, greatly facilitated replication.

Barriers:

- Ranges were given for the parameters, instead of individual values.

3.3.1.5 Case Study 5: Ganesalingam et al. (2015), Replicator: Ewan Gray

The final case study replicated an evaluation conducted by Ganesalingam et al. (2015) which was published in the medical journal, Stroke and funded by the NIHR. This model compared mechanical thrombectomy to standard care alone which was defined as: Intravenous tissue-type plasminogen activator (IV-Tpa), in cases of acute stroke, from the perspective of the UK NHS. Analyses were carried out using a combined short-term decision tree and state-transition cohort model, with a hypothetical cohort of 1,000 patients simulated. The time horizon was 20 years with discounting of costs and outcomes at 3.5%. The replication was conducted using Microsoft Excel, although the original modelling software was not stated.

The replicated costs varied by 2.00% and 1.39% in comparison to the original, whereas the outcomes varied by 0.13% and -0.87%. This case study was the only one where the interventions were not dominant or dominated, and so it enabled an ICER to be replicated. The original ICER was £11,651 per QALY, in comparison to £12,051 when using the replicated values, a total variation of 3.43%.

The original publication thoroughly reported the model, with a diagram being provided and all of the parameters required to recreate the main analyses being clearly listed in a table. The cost per cycle for two of the model states was also given, which further facilitated the replication. This was reflected in the number of Philips Criteria that were satisfied, 74% of applicable criteria.

Despite the parameters being comprehensively reported, several barriers to replication were still encountered which required additional assumptions during the replication.

These included uncertainties about the allocation of treatment costs following recurrent stroke, as well as how discounting was applied. It was unclear whether the first cycle was considered as time zero (given that 3 months was meant to have elapsed within the decision tree) and therefore whether the cycles within the first year were discounted or not.

The probabilistic sensitivity analyses were also replicated in this case study. During which, it became apparent that not all of the distribution parameters were included in the publication (for example, no shape parameters were provided for any of the gamma distributions), and although some of these were available in online supplementary materials, additional assumptions about the distributions used were required. As well, the shape parameters reported for the beta distributions to generate utilities were implausible, as they generated values that were far lower than the point estimates reported in the base case analysis and univariate sensitivity analysis.

To demonstrate this, the values reported in the manuscript for the beta distributions used in the probabilistic analysis, alongside the calculated mean and mode of the beta distribution using the reported shape parameters are shown in Table 3.4.

For all of the distributions reported, the calculated mean is well below the base-case value, and outside of the reported univariate sensitivity range. This suggests that at the very least, there has been a reporting error or potentially an error within the technical specification of the model when carrying out the probabilistic sensitivity analysis. Whilst the results of the probabilistic sensitivity analysis are not extensively discussed in the manuscript, a figure is presented of the Monte Carlo simulation for 1,000 patients using 10,000 simulations (Figure 3 in Ganesalingam et al. (2015)), which shows the simulation points broadly clustering around the base-case results, suggesting that the inappropriate shape parameters are more likely to be a reporting error.

Table 3.4: Beta distributions used in the probabilistic analysis of Case Study 5.

Reported in publication						**Calculated statistics**	
Utilities	Base-case value	Univariate sensitivity analysis	Distribution	Range	Alpha-Beta	Mean (2dp)	Mode
Independent mRS 0-1-2	0.74	0.70–0.77	Beta	0–1	684-3,021	0.18	0.18
Dependent mRS 3-4-5	0.38	0.29–0.47	Beta	0–1	60-590	0.09	0.09
Recurrent stroke	0.34	0.32–0.36	Beta	0–1	540-5,685	0.09	0.09

Mean calculated using Gupta and Nadarajah (2004):

$$\mu = \frac{\alpha}{\alpha + \beta}$$

Mode calculated using Gupta and Nadarajah (2004):

$$\hat{x} = \frac{\alpha - 1}{\alpha + \ \beta - 2}$$

Whilst carrying out this replication, an attempt was made to contact two of the original authors, to ask if they would be willing to share the original model code, however no response was received.

Facilitators:

- Model parameters clearly listed.
- Example of costs per cycle were given for two of the model states.

Barriers:

- Ambiguity about the assumptions made with treatment costs following recurrent stroke.
- Lack of clarity surrounding the three month decision tree and how this affected subsequent cycle discounting.
- Not all distribution parameters used in the probabilistic sensitivity analyses were given. As well, implausible shape parameters were given for the beta distributions.

- Authors were unavailable to provide clarification as they did not respond to requests for help.

3.3.2 Philips Criteria

Across the five publications replicated, all appeared to satisfy the majority of Philips criteria (as shown in Chapter 3 Appendix, Table A3.1). Whilst it may have appeared that the reporting of the studies was thorough, the replication case studies still highlighted areas where the reporting prevented replication. Table 3.5 shows a subset of the Philips checklist that were perceived to be most relevant in facilitating replication.

In example of this, the following criteria: “Have all data incorporated into the model been described and referenced in sufficient detail?” and “Is the process of data incorporation transparent?” were considered to be satisfied for all of the five case studies. However, as documented above, there were certainly issues with the reporting of model parameters (as seen in case studies 1 and 4) and ambiguity around how costs were incorporated, as with the “weighted average” of treatments used in case study 3 and as well in case studies 2 and 5. Moreover, when comparing the barriers identified in the replications and the Philips checklist responses, it does not appear that the checklist criteria were able to suggest the presence of such barriers, such as conflicting information being presented; or issues with the way that parameter information was presented.

Table 3.5: Shortened Philips checklist, with items directly related to reporting thoroughness.

	Case Study				
Checklist Item	1	2	3	4	5
Has the scope of the model been stated and justified?	y	y	y	y	y
Has the evidence regarding the model structure been described?	y	n	y	y	n
Are the sources of data used to develop the structure of the model specified?	y	n	y	n	n
Are the structural assumptions transparent and justified?	y	y	y	y	y
Is there a clear definition of the options under evaluation?	y	y	y	y	y
Is the time horizon of the model, and the duration of treatment and treatment effect described and justified?	y	y	y	y	y
Is the cycle length defined and justified in terms of the natural history of disease?	p	n/a	y	p	p
Is the choice of baseline data described and justified?	y	y	y	y	y
Are transition probabilities calculated appropriately?	y	y	y	y	y
Has a half cycle correction been applied to both cost and outcome?	n	n/a	n	n	n
Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified? Have alternative assumptions been explored through sensitivity analysis?	y	y	y	y	y
Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?	y	y	n	y	y

Table 3.5: Shortened Philips checklist, with items directly related to reporting thoroughness. *(Continued)*

	Case Study				
Checklist Item	1	2	3	4	5
Is the source for the utility weights referenced?	y	n/a	y	n/a	y
Have all data incorporated into the model been described and referenced in sufficient detail?	y	y	y	y	y
Is the process of data incorporation transparent?	y	y	y	y	y
If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	n/a	n/a	n/a	n/a	p
If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated and justified?	y	y	p	y	p
Abbreviations: Y: Yes; N: No; P: Partial; N/A: Not applicable.					

3.3.3 Evaluating the definitions of success

Table 3.6 contrasts the definitions identified in Chapter 2, to the results of the replication case studies conducted above. In doing so, it is possible to see whether the replications would be deemed as a success or failure, under each of the proposed definitions.

All of the replications satisfied the least specific definition (#1); that the same conclusions for intervention cost-effectiveness were reached. In contrast, none of the case studies met the strictest definition of replication success (#6); that the replication yielded the exact same results. Between these extremes, the other definitions resulted in greater variation. Definition #2, which looked at the percentage of ICER variation, was not applicable for the majority of case studies (four out of five) due to treatment dominance making it inappropriate to calculate ICER values. The only case study to replicate an ICER, case study 5, would have been considered a replication success if the percentage for variance was set at 5% (given total variation of 3.43%). Case study 3 on the other hand, was the only replication study that satisfied the third definition suggested, which considered a successful replication as the costs and outcomes being replicated for some treatment pathways and not others. Definition #4 was not applicable to any of the replications, given that none of them attempted to replicate the original figures. Finally, for definition #5, which required a variation limit to be specified, I chose the conventional value used in tests of statistical significance (5%) along with a stricter requirement of 1%. Using the 5% level, the majority of case studies (four out of five) were considered a successful replication, whereas none of the case studies satisfied the 1% threshold wholly.

In applying these definitions to the results of the case studies, several points were observed in terms of their workability. The first, was the importance of the original study's results and how these influenced whether or not the definition was useful as a measure of 'replication success'. This is a particularly important consideration for definitions #1, #2 and #4. Definition #1 relies on the same cost-effectiveness results being produced. However, this may be easier or more difficult to achieve depending on the original study results. For example, in the replication case studies above, the majority were either dominant/dominated or comfortably under conventional willingness to pay thresholds. This meant that even with large variations in the original results compared to the replication, the same conclusion would be reached. Therefore, this definition could be satisfied despite having considerable variability between original and replicated results whereas much smaller variability would be accepted if the original results were nearer to willingness to pay thresholds. This allows for potential unfair variability depending on the original study results. Similarly, depending on the original results, definitions #2 and #4 could be redundant: given it is inappropriate to calculate an ICER when a treatment is dominant. This implies that either the original study results need to be first considered before choosing an applicable definition; or that these definitions would not be functional in future replication studies (given their lack of generalisability). As was expected, definition #6 was too specific, and resulted in none of the case studies being deemed a success. This went against some of the subjective views of the replicators, which suggests that it may not be a workable definition. Of the remaining definitions, definition #3 was found to be overly specific, with only one case study meeting the requirements. In contrast, definition #5 allowed some variability, whilst ensuring that the same cost-effectiveness results were derived. It is therefore likely that this is the most workable definition, albeit with some further clarifications. Currently the definition does not specify whether it should apply to only some or all of the pathways replicated. As such, all of the case studies were labelled with “Partial” when looking at a threshold of 1%, due to some of the treatment pathways being replicated closely. It is likely that the most workable version of this definition is to allow for the majority of pathways to be replicated. Given this, I propose an update of definition #5:

“Results for the costs and outcomes vary by only XX% compared to the original for the majority of the treatment pathways replicated, AND are consistent with the original conclusions.”

Table 3.6: Deeming the replications a success or failure according to the definitions proposed in Chapter 2.

		Case Study					Comments
Proposed definition		**1**	**2**	**3**	**4**	**5**	
1. The same conclusions for intervention cost-effectiveness were reached		Yes	Yes	Yes	Yes	Yes	All of the replications found this.
2. The calculated ICER varies by only XX% compared to the original		N/A	N/A	N/A	N/A	Yes	For case study 5, there was a total variation of 3.43% between ICERs.
3. Costs and outcomes replicated for some treatment pathways/model scenarios and not others		No	No	Yes	No	No	Case study 2 and 5 replicated the outcomes in some scenarios, but not costs. Case study 3 satisfied this, conditional on rounding (0dp for costs and 2dp for outcomes).
4. Cost-effectiveness figures could be reproduced to a reasonable degree of success (for example, the cost-effectiveness acceptability curve)		N/A	N/A	N/A	N/A	N/A	
5. Results for the costs and outcomes vary by only XX% compared to the original, AND are consistent with the original conclusions	±1%	Partial	Partial	Partial	Partial	Partial	A distinction needs to be made for this definition about whether it is just for some of the pathways replicated or for all of the scenarios. The majority of case studies had some instances where the costs or outcomes
	±5%	Partial	Yes	Yes	Yes	Yes	

Table 3.6: Deeming the replications a success or failure according to the definitions proposed in Chapter 2. *(Continued)*

	Case Study					**Comments**
Proposed definition	**1**	**2**	**3**	**4**	**5**	
	🟧	🟩	🟩	🟩	🟩	were replicated within 1%. The choice of the degree to which results can vary is arbitrary, and for the purposes of this definition, 1% and 5% were chosen (in keeping with traditional levels of statistical significance).
6. Identical results are produced	No	No	No	No	No	None of the replication attempts met this definition.

3.3.4 Subsequent findings from Schwander et al. (2021)

There are several similarities between the barriers and facilitators identified by Schwander et al. (2021) and those detailed in the above case studies. In example of common barriers, in two out of the four case studies conducted by Schwander et al, there was insufficient information on the parameter values used, meaning that the probabilistic sensitivity analyses could not be replicated (as with case study 5 above). In another of their case studies, Schwander et al. found that parameter information was only presented in a figure. The authors also cited a similar replication facilitator, that the inclusion of clinical event frequencies greatly facilitated the replication as it made “it possible to check whether the event simulation and hence the clinical heart of the replicated model is working correctly” (Schwander et al., 2021).

More generally, the results of the replication case studies also showed similarities to those conducted above, in that the replicated costs deviated more than outcomes. The average variation in replicated costs compared to the original was 3.78% (ranging from -3.9 to 16.1%) and for QALYs, -0.11% (ranging from -3.7 to 2.1%), excluding the case study where variations of over 100% were found. This echoes the suggestion made above, that further detail regarding any costing assumptions used, are required.

Schwander et al. evaluated the reporting thoroughness of the original publications using the CHEERS checklist (Husereau et al., 2013). Whilst a different checklist was used, their findings were also similar. They cited that often it was not clear that crucial information was omitted from the publication until the replication was conducted; “the missing information on clinical event results (case studies 1 and 2) had no impact on the CHEERS rating on the quality of reporting results (CHEERS item #19)” (Schwander et al., 2021).

When applying the definitions of replication success, similar to the findings above, all of their case studies met the loosest definition (#1) and no case study met the strictest definition (#6). Where a threshold of variability needed to be specified in the definitions, Schwander et al. used a range of values: 5%, 10% and 20%, allowing for substantially larger levels of variation than the 5% and 1% used in the thesis study. In addition to calculating the variation in terms of costs and outcomes, the authors also calculated percentage variation in terms of the ICERs. For the latter, the variation rarely fell within the permitted ranges, with the authors concluding that: “for the assessment of incremental costs, QALYs, and ICERs, the calculation of relative variations may be misleading”. Instead, they proposed that the results should be visualised on a cost-effectiveness coordinate plane citing that this would allow the size of the differences to be put into perspective. The authors concluded that any definition needed to: meet the same cost-effectiveness conclusions, permit some variation in terms of replicated costs and outcomes (suggesting 5%), and that incremental results should be visualised to ensure that they are “fairly comparable”.

3.4 Discussion

In this chapter, five case studies were considered, attempting to replicate published decision models. In doing so, common barriers and facilitators to replication were found, which may inform how future modelling studies are presented, which is the focus of the next chapter, Chapter 4.

Common facilitators included providing a clear diagram of the model, detailing all potential transitions, and clearly reporting transition probabilities alongside the treatment pathways. As well as this, providing example cost calculations, documenting the cost per cycle for model states (as was seen in case study 5), and providing example calculations for transitions, were also considered great facilitators to model replication.

In contrast, common barriers included the use of conflicting information, particularly between model diagrams and the manuscript text or tables. In addition, despite all of the papers providing some description of model input parameters, these were commonly grouped for multiple treatment pathways or time horizons, instead of being specific. Consequently, it was difficult to appreciate which parameter referred to which model iteration or scenario, and thus was a common barrier to replication. As well, the case studies that had longer time horizons proved to be harder to replicate given that any, even minor, discrepancies in costs, outcomes or transition probabilities between the original and replicated model became amplified over time. Notably in two of the case studies an author was unsuccessfully contacted for clarification; whilst this cannot be counted as a barrier to replication per se, it should act as a reminder to modellers to archive and thoroughly annotate their work, to facilitate future enquiries, regardless of the time that may have elapsed since the original work was conducted and published.

The results of the replications also revealed another finding, which was that there was often greater variation in the replicated costs than outcomes. This can be seen when comparing the range of variation in costs: -4.54% to 108.00%, compared to outcomes: -3.81% to 0.40%. This is not unexpected, given the multiple components that may feed in to the total costs of an intervention (intervention costs, primary and secondary care resource use, productivity costs and so on), compared to health state utilities, which usually have a single value per state. With that being said, this may suggest that future modelling studies should consider including greater detail on how costs were derived for the model, perhaps by including example calculations, or making explicit all of the assumptions made.

These findings also build on the existing replication literature. Smolen et al. (2015) found replication variances ranging from 0.00% and 1.92% in their base case replication. The lead author of this replication was contacted (as described in Chapter 2), to further understand their experiences of carrying out the replication, particularly to find out about perceived facilitators or barriers. Whilst the published replication study stated that all parameters were clearly reported within the original study, Smolen clarified that to generate the transition probabilities used in the model special 'pixel counting' software was used, as the parameters were presented in graphical form. Therefore exposing a barrier to replication. Barriers identified in the Bermejo et al. (2017a) study which detailed the replication of five models, also reveal commonalities. In one of the replications, the authors stated that "no details were provided on… covariate or coefficients" for a generalised linear model that was used. In another replication, they found that "the model structure diagram did not exactly match the implemented model"; which echoes the findings in the above work, with case studies 1 and 2.

Reporting checklists, such as the Philips checklist are often advocated to encourage transparent research practices. The responses to the Philips checklist for the studies included in the case studies, suggest that the original publications were comprehensively reported, shown by the fact that the majority of applicable criteria were satisfied in each of the case studies, ranging from 69% to 83%. Indeed, three of the case studies were chosen because they satisfied the most criteria out of those identified in a review and the latter case studies were chosen by the replicator presumably with some thought towards how well the studies were described to facilitate replication. With this being said, there were still a number of barriers encountered to replication, which may suggest that the Philips checklist is not able to discern whether studies are reported thoroughly enough to facilitate replication. This may be due to the fact that the checklist focuses more so on model quality, for example ensuring that appropriate justifications for the methods used are given. It could also be that the checklist may be completed in different contexts which may be less scrupulous to the threshold for replication, such as peer reviewing or quality assessment for a peer review. These are similar to the findings reported by Mount Hood (as described in previous chapters), who undertook several replications of diabetes simulation models. The authors referred to the commonly used checklists of CHEERS (Husereau et al., 2013) and Philips (Philips et al., 2004) as being potentially "overly general to satisfy the needs in complicated multifactorial disease areas" to facilitate adequately transparent reporting and thus a models replication.

Given the barriers and facilitators identified, along with the potential inadequacies of reporting checklists, if future models are to be more easily replicated, instead of checklists, a more formalised process of how models are presented is required. For example, whilst a table of input parameters may be enough to satisfy a checklist item, it may not be apparent until a replication is conducted if any parameters have been omitted. Examples of how the presentation process may be more formalised include necessitating a table with the total costs associated with every model state. In doing so, any implicit assumptions that the modeller had failed to document in the manuscript could be deciphered. Other suggestions could include providing a summary of results for a shorter time horizon (for example a year), given that models with a longer time horizon proved harder to replicate. This would give the replicator values to check against, before running the model over many more cycles, and therefore inflating any discrepancies.

Alternatively, a more thorough presentation could be encouraged by changing workflow practices to give more consideration to replicability. Replication of the model programming by two individuals within the study team, both working from the same analysis plan, would encourage clear and unambiguous descriptions of the model structure. This type of redundancy in workflow is already commonly practised by statisticians analysing clinical trial data as well as in software development (detailed in 'Good Clinical Practice Guide', Chapter 9.7.9, Quality Control, Medicines and Healthcare products Regulatory Agency (2018)).

It should also be mentioned that whilst this chapter has focused on how authors could present their modelling studies to facilitate replication, there are other factors that could greatly influence replicability. These include: journal data sharing policies, word limits and the use of supplementary materials. As well, if model registries (Arnold & Ekins, 2010; Sampson, 2012; Sampson & Wrightson, 2017) or the publishing of open-source models (Dunlop et al., 2017) were more commonplace, replication may be more easily facilitated, given that a replicator could access and inspect model code (albeit that to be understood by a third party, this would still require detailed annotation).

The five replication case studies also presented an opportunity to test out the definitions of 'replication success' proposed in Chapter 2, and to evaluate the workability of the definitions proposed. In doing so, it was found that all of the case studies satisfied the loosest definition of success, whereas none met the stricter criteria of identical results being produced. In applying the definitions, it was also shown that some of the definitions were not applicable to certain studies, depending on the original results of the model being replicated. In example, replicating an ICER value would not be appropriate for studies that found a treatment to be dominant. This is an important finding from trying to apply the definitions proposed, and shows that a universal definition would need to be applicable regardless of the original study's results. From this work, an updated definition was proposed, with an additional clause, clarifying that a replication success should be consistent with the original study's conclusions, and that the majority of treatment pathways can be replicated to within a reasonable degree of accuracy.

3.5 Strengths and limitations

Whilst efforts were made to ensure the breadth of models replicated, the five case studies only covered Markov and decision-tree models. As such, the potential barriers and facilitators to model replications for more complex models, such as discrete event simulation models have not been explored. Moreover, whilst the suggestions made to enhance the likelihood of replication in terms of model presentation may appear feasible for more simplistic models, it is recognised that transparent reporting is likely to be far more challenging with complex models. The lack of diversity in the case studies is also reflected in the software used within the models, with the majority using Microsoft Excel, hence preventing the evaluation of how other software types or programming languages, such as R, may facilitate or impair replication. This may have occurred due to the replicators unconsciously opting for models that looked to be more feasibly replicated.

It is also important to acknowledge that the results of the replications are highly dependent on replicator skill. Therefore, it is possible that given a different replicator, the same case study may have had more or less success. Also, the inability to replicate does not necessarily infer errors within the model, but may just suggest a lack of information within the report.

Efforts were made to mitigate the effect of the replicator however, by using different modellers with varying levels of experience, to conduct each of the case study replications. It is also promising that common barriers and facilitators emerged across the five studies, indicating that they were more than a subjective experience. Moreover, whilst there was variation in the motivation for selecting the replicated models (Philips criteria fulfilled and pragmatism), it should be acknowledged that the models chosen pragmatically are also likely to have also been chosen on the basis that the replicator believed that the models were described in sufficient detail to facilitate replication. Thus, it is possible that the models selected were of a higher reporting standard than the ‘average’ published modelling study. If less thoroughly reported models were included in the replication attempts, it is possible that additional barriers may have been identified, perhaps for example the use of proprietary data.

Importantly, it should be reiterated that none of the case studies were chosen to single out any author, institution or journal.

3.6 Conclusion

This study has highlighted several barriers and facilitators that may influence how replicable a modelling study is. It is hoped that these can be used to inform the way future models are presented; this might particularly apply to how thoroughly costing assumptions are reported, given that costs were often replicated with greater variation to the original, than outcomes. This study has also shown that the Philips checklist, is not indicative of whether or not a model is likely to be replicable, given that all of the case studies appeared to be relatively well reported. The replications conducted also presented an opportunity to apply the definitions of ‘replication success’ derived in the Chapter 2. Applying these definitions highlighted that several of these definitions were reliant on a certain cost-effectiveness outcome (i.e. that no treatment dominated) which would prevent the generalisability of any definition. As such, an updated definition was proposed.

It is reassuring to see that the findings of these replication case studies and applying the definition of success were echoed in the publication by Schwander et al., which was published subsequent to this chapter’s work.

Chapter 4 Developing and reporting a decision model in a manner which seeks to aid future replication

4.1 Introduction

As evidenced throughout this thesis, there have been repeated calls for research transparency and reproducibility initiatives. Within health economics, these include: the establishment of modelling registries (Sampson & Wrightson, 2017) and publication of modelling code (Dunlop et al., 2017; Pouwels et al., 2022), and in other disciplines: conducting replication studies (Höffler, 2018). However, these calls rarely consider how easily they may be implemented by the individual researcher and thus may lack practicality. Most initiatives will require in the least additional researcher effort, either in the short-term whilst learning new programming languages or more generally when making aspects of the modelling process more transparent. This may include, for example, making modelling assumptions more explicit, or making code accessible and readable to an external party. The additional time required for this will incur a cost, either requiring more funding of researcher time or forcing researchers to do more within limited constraints. Hostler (2023) warns of the consequences of encouraging open research practices but not explicitly funding this additional work, in a concept they refer to as "workload creep". This is a phenomenon whereby researchers are expected to engage in additional tasks and commitments, but the time to do so is not specifically funded, potentially leading to higher rates of work-place burnout and stress (Hostler, 2023).

There has also been little discussion about the incentives or motivators for researchers in health economics to undertake the additional work to make their output replicable. Currently, there are few tangible incentives for such work. Motivation to improve replicability may come from an expectation that new standards will soon be adopted, either required by research funders or journals. Some research funders now request that findings are published in open-access journals (for example, UK Research and Innovation from April 2022), although publishing in an open-access format stops short of most reproducibility calls. A recent report by the House of Commons, Science, Innovation and Technology Committee exploring 'Reproducibility and Research Integrity' found that open-access calls "should go further in requiring the recipients of research grants to share data and code alongside the publications arising from the funded research" (House of Commons: Science Innovation and Technology Committee, 2023). Some journals now request replication packages to be published alongside the accepted article (as is now commonplace in American Economic Review (American Economic Association, 2020)), although this is not yet standard practice.

This chapter explores the practicalities of a researcher trying to engage in replicable practices, particularly, to develop a decision model with replication in mind. It uses this development process to explore how the replication barriers and facilitators identified in the previous chapter (Chapter 3) may be implemented and overcome. The learning points derived from the practicalities of incorporating these into a decision model are important to consider. If they cannot feasibly be achieved then the recommendations themselves will not improve standards. It aims to answer the following research question identified at the start of this thesis:

- What are the implications and challenges for modellers trying to incorporate replicability?

4.2 Methods

4.2.1 Reason for model development

The ‘Diabetes Prevention – Long Term Multimethod Assessment’ (DIPLOMA) study (National Institute for Health and Care Research, 2017) was commissioned by the NIHR to independently evaluate the NHS Diabetes Prevention Programme (NHS DPP), in terms of its effectiveness and cost-effectiveness from the perspective of the NHS. To evaluate the cost-effectiveness, the study pledged to undertake a short-term analysis using the observed data, and a longer-term analysis using a decision model. It is the latter model development which is described in this chapter. It is important to note that the model used in this PhD to evaluate the practicalities of incorporating replicable research practices was done so within the constraints of a major national funded research project, rather than a project solely for the purposes of the PhD, thus allowing it to closely mirror the experience of other researchers.

4.2.2 Replicable research practices

Whilst creating the model, care was taken to be mindful of the decisions being made at each stage, and to document the subsequent choices and rationale for each. In particular, there was a focus on ensuring the model code was transparent and replicable, programming the model in an open-source software and publishing the model code. There was also a focus on incorporating the barriers and facilitators identified in the previous chapter, along with those identified in the similar replication exercise subsequently published by Schwander et al. (2021). The barriers identified in these two studies are presented in Table 4.1, alongside proposed methods and planned deliverables for removing these barriers in subsequent models (if applicable). In both of these studies, key facilitators included providing a detailed description of the model structure, including all possible transitions, as well as a table of all of the input parameters. Other facilitators identified in the case studies detailed in Chapter 3 included giving example cost calculations for each model cycle or providing costs per model state, and including a table of health states and explicitly listing any assumptions made.

During the model development, effort was made to incorporate the planned deliverables into the model development whilst evaluating the feasibility of doing so.

4.2.3 Modelling Software

One of the most important aspects of replicable research practices is the software used to develop the model. Whilst it is important that it has a sufficient user base, it might also be important that the software is open-source, and therefore accessible to all. Software that can be used to develop decision models include, but are not limited to: Microsoft Excel, TreeAge, Simul8, and R. Both TreeAge and Simul8 are specialist software products for the development of models, whereas R is a general programming language often used for statistical analysis and Microsoft Excel is a spreadsheet software used for data visualisation and analysis. As part of their Health Technology Assessment (HTA) process, NICE currently accepts economic models built in Microsoft Excel, TreeAge, R or WinBUGs (with other software requiring special permission) (National Institute for Health and Care Excellence, 2022). Markedly, R is the only programming language and environment in this list, with the others being proprietary software, meaning that they are not open-source or free to use.

It is unclear how frequently different software is used when developing decision models (or indeed, how this may have changed over time). In an example of software use, a systematic review conducted by the author, identified 24 models evaluating treatments for atopic eczema. This found that the most popular software was Microsoft Excel (46%, 11/24), followed by TreeAge (21%, 5/24), although a large number did not report which software was used (33%, 8/24) (McManus et al., 2017).

More recently, there have been several calls for modellers to move away from developing models in Microsoft Excel (Incerti et al., 2019), given that these models can be difficult to navigate, and are prone to coding errors (due to formulae relating to several cells and the ability to easily overwrite text). R is often suggested as an alternative, with researchers identifying the following benefits: increased transparency, free to use, ability to conduct all analyses in one script - referred to as end-to-end functionality (Hart et al., 2020) and computational efficiency (Xin et al., 2021). With R, scripts are written linearly, which means that the code may be easier to follow than the multi-cell and worksheet approach used in Microsoft Excel.

There are also arguments against using a programming language like R. Some consultancy companies consider that their clients may be more familiar with Microsoft Excel and therefore less able to understand the model if developed in another software, and so Microsoft Excel may actually be perceived as more transparent (Hart et al., 2020). There may also be practical restrictions, such as some HTA agencies not accepting submissions in R. Up until recently, the HTA body in the Netherlands, Zorginstituut Nederland (ZIN) did not accept R submissions, although they have recently run a pilot scheme to evaluate the feasibility of doing so (Zorginstituut Nederland, 2022).

Given the perceived transparency benefits of using R rather than Microsoft Excel, it was decided to programme the model using R. R was chosen as it is open-source, which would enable anyone to inspect and run the model code. R code also means that the whole model could be coded in one script. Importantly, I had not used R to develop a decision model before, although had previously used it to carry out some statistical analyses.

Table 4.1: Barriers of model replication as identified in Chapter 3 and Schwander et al. (2021).

Barriers	**Planned Deliverables (if applicable)**
Identified in Chapter 3:	
Work flow	• Model to be programmed in an open-source software, such as R.
	• Model code to published, for example in a Github repository.
Authors unable to be contacted – requirement for authors to archive and annotate their work	• Publishing code on Github would mean that this was open to interrogation / and would not be dependent on the author responding to requests.
	• Development of an R Shiny App.
Lack of transparency around costs used	• Table of costs used for each model state.
	• Summary of costs for a shorter time horizon.
Lack of transparency around model parameters	• Table of parameters grouped for each of the scenarios modelled.
	• Diagram for each scenario, along with all possible transitions documented.
Identified in Schwander et al. (2021):	
Inability to recreate probabilistic sensitivity analysis due to lack of reporting of standard deviations or distribution parameters	• Ensure all shape parameters for probabilistic sensitivity analysis are reported.
Lack of reporting of clinical event results	• Present event and mortality results (for all simulated alternatives) over the whole time horizon of the model, for each model cycle.
Self-created regression analyses	• State all regression methods used, and provide details on how to apply/solve the regression equations correctly.
Lack of reporting of details on Life Tables	• Ensure that adequate detail is provided around the Life Tables, including the year used.

4.2.4 Model background

Type 2 diabetes is a chronic condition that affects insulin function and production, leading to high blood glucose levels. Symptoms include tiredness and excessive thirst, diabetes is also associated with an increased risk of developing other problems, such as cardiovascular disease, nerve and eye damage (Chatterjee et al., 2017). Worldwide, it is estimated that diabetes is the sixth leading cause of disability (Vos et al., 2016).

Whilst there are genetic risk factors, such as a family history of the disease and being from South-Asian, African-Caribbean or black African descent; the majority of risk factors are lifestyle based. These include, being overweight, smoking, poor diet and physical inactivity (Fletcher et al., 2002). People with type 2 diabetes are usually prescribed medications, such as Metformin, to regulate their blood sugar, although it can sometimes be managed with weight loss, diet and exercise alone (Krentz et al., 2008).

In the UK, prevalence of diabetes is increasing, with prevalence rates more than doubling between the years of 2000 and 2013 (2.39% to 5.32%) (Sharma et al., 2016), exacerbated by an increase in average weight, sedentary lifestyles, and an aging population. As well as the patient and carer burden, type 2 diabetes also places a considerable burden on healthcare systems, with a study estimating direct costs to the NHS each year of approximately £8.8bn (price year 2011-12) (Hex et al., 2012).

Before developing type 2 diabetes, many patients first develop non-diabetic hyperglycaemia or ‘pre-diabetes’, characterised by elevated blood glucose levels that are below the threshold of type 2 diabetes, but above normal ranges (Tabák et al., 2012). It is possible for individuals in this state, to progress to type 2 diabetes or return to normal blood glucose levels. It is estimated that 11% of individuals with obesity and non-diabetic hyperglycaemia will progress to type 2 diabetes annually (Tabák et al., 2012); with an estimated conversion rate of 7% within the first year of non-diabetic hyperglycaemia diagnosis (R Ravindrarajah et al., 2020).

Studies have shown that targeting this high-risk group of non-diabetic hyperglycaemia patients with interventions that combine diet and exercise advice can delay or prevent the progression to type 2 diabetes (Hemmingsen et al., 2017). A recent systematic review identified 29 studies evaluating a lifestyle intervention to prevent type 2 diabetes (M. Davies et al., 2017) and conducted a meta-analysis (using 25 studies). The meta-analysis found that lifestyle interventions result in a mean weight loss at 12 months of 2.31kg (95% CI: –2.87 to –1.76 kg), and a reduction of 0.11% (95% CI: –0.19% to –0.03%) in blood glucose levels, measured with HbA1c (using eight studies).

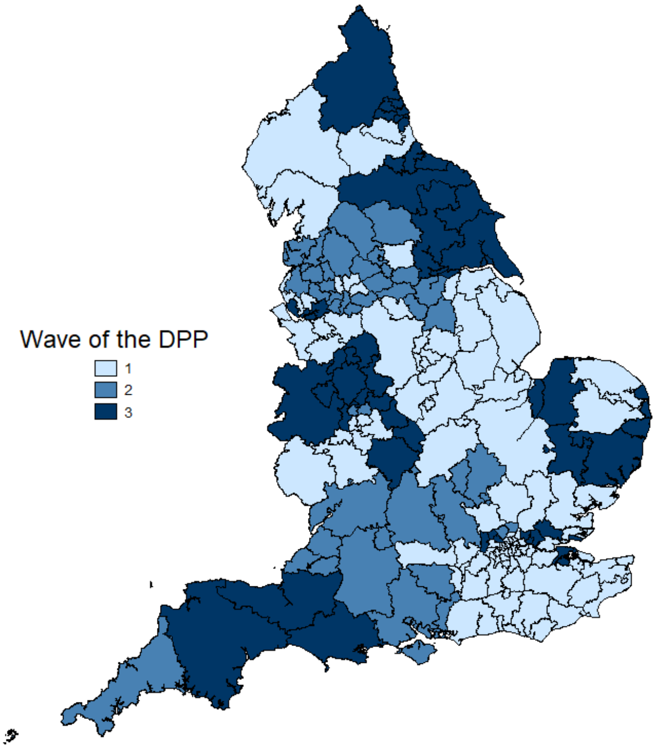
The NHS DPP was established as part of a major new emphasis in England on the need for targeted prevention strategies. The programme targets individuals with non-diabetic hyperglycaemia, with the aim of preventing these individuals from going on to develop type 2 diabetes.

Participants are primarily referred to the programme by a healthcare professional, during an NHS health check, or opportunistically during a consultation. To be referred, participants require a blood test to show that they are within the non-diabetic hyperglycaemia range (HbA1c 42–47 mmol/mol (6.0–6.4%)). HbA1c is a measure of average blood sugar over 8-12 weeks, with normal glucose tolerance considered to be a HbA1c concentration of below 42mmol/mol.

The programme comprises of at least 13 group-based behaviour change sessions, incorporating structured education on nutrition, physical activity and weight loss and typically lasts 9-12 months. It began in 2016, being rolled out across England in a series of waves (as shown in Figure 4.1), covering the whole of England by April 2018. To avoid placing additional burden on frontline NHS services, NHS England decided to procure the programme from external contractors. As part of the first framework, four commercial providers were selected in a national competition. Only these providers were able to bid to provide the programme at each site.

Figure 4.1: England Clinical Commissioning Groups according to the first wave in which they participated in the NHS Diabetes Prevention Programme (wave 1 from 2016, wave 2 from 2017 and wave 3 from 2018).

Figure created using data obtained NHS England.



NHS England conducted an impact assessment in 2016, which estimated the costs associated with implementing the NHS DPP in England from 2016-2021, along with the expected outcomes of the programme in the short- and long-term (NHS England, 2016) using estimates from the literature along with expert opinion. This report estimated that by 2021, a total of 18,000 cases of diabetes would have been prevented or delayed amongst a cohort of 390,000 participants. It also found that the programme was likely to be cost-effective, and by the year 2033/34 cost saving (when applying discounting), assuming a cost of £270 per participant.

The model described in this chapter sought to update these estimates of long-term cost-effectiveness, for the first time using observed data from the programme on participant retention and health outcomes, along with provider contract information.

4.2.5 Model description

The excerpts below are from the 'Methods' section of a publication currently under review describing the model that was developed:

A Markov cohort model was developed in R (Green et al., 2023; R Core Team, 2009) to evaluate the cost-effectiveness of the NHS DPP compared to usual care, from the perspective of the NHS. A model was developed as it is likely that many of the benefits of the NHS DPP occur in the long-term, beyond that which can be observed using routinely collected data. The model code in full is available via the author's Github repository (McManus, 2023) and an excerpt of the code is provided in Appendix Chapter 4, Code excerpt – Base case analysis.

4.2.5.1 Interventions analysed

Two strategies: 1) usual care and 2) referral to the diabetes prevention programme, in addition to usual care were considered. In this instance, usual care was defined as what existed prior to the introduction of the NHS DPP. Guidance from NICE, first published in 2012, recommends that individuals with non-diabetic hyperglycaemia be offered a blood test and assessment of their body mass index at least once a year (National Institute for Health and Care Excellence, 2012). The intervention evaluated is a referral to the NHS DPP. An individual referred to the programme is free to take up as much or as little of the programme as they wish. The usual programme pathway comprises of first an initial assessment and then a series of group-based sessions. Depending on the provider, there are either 13 or 18 total sessions. Further detail about the programme can be found in Barron et al. (2018).

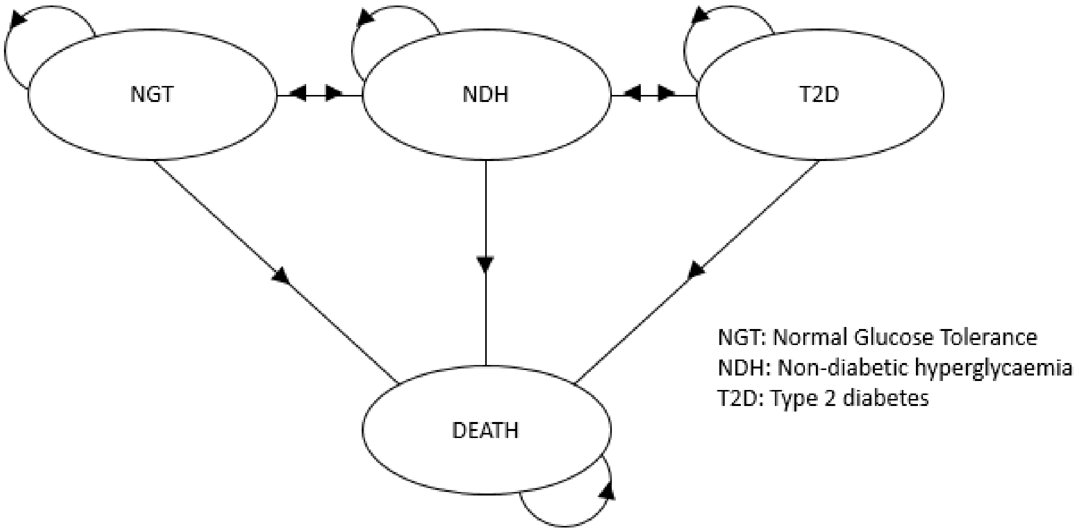
4.2.5.2 Model Structure

A Markov model was chosen due to the chronic nature of type 2 diabetes, which involves recurring events and continuous risk over time (Sonnenberg & Beck, 1993). The model comprises of the following states: Normal glucose tolerance (HbA1c below 42mmol/mol (6·0%)), Non-diabetic hyperglycaemia (defined as a HbA1c within the range of 42–47 mmol/mol (6·0–6·4%)), Type 2 diabetes (HbA1c above 47 mmol/mol (6·4%)) and Death. These states are mutually exclusive and exhaustive. A diagram of the model and the possible transitions between each of the states is shown in Figure 4.2.

In the model, it is possible for individuals with non-diabetic hyperglycaemia to transition to a normal glucose tolerance state, remain with non-diabetic hyperglycaemia, progress to type 2 diabetes or death. Individuals in the type 2 diabetes state can remain in the state, return to non-diabetic hyperglycaemia or die. However, individuals with type 2 diabetes cannot directly transition to normal glucose tolerance and likewise it is not possible for individuals with normal glucose tolerance to directly transition to type 2 diabetes.

The model structure was developed by reviewing existing model structures and in collaboration with a clinical expert to ensure it captured the clinical reality of the disease. This structure is also used in several studies (Frempong et al., 2021; A. J. Palmer et al., 2004; A. J. Palmer & Tucker, 2012; Roberts et al., 2018). Roberts et al. (2018) also state that this structure was developed following a review and in collaboration with a multi-disciplinary clinical team.

Figure 4.2: Markov model structure.


NGT
NDH
T2D
DEATH
NGT: Normal Glucose Tolerance
NDH: Non-diabetic hyperglycaemia
T2D: Type 2 diabetes

4.2.5.3 Model Parameters

The transition probabilities from and to each of the model states are described in Table 4.2, in terms of the point estimates and sampling distributions used. Where possible, age specific transition probabilities were obtained for the following age groups: <40, 40-49, 50-59, 60-69, 70-79 and 80+ years, although this was not possible for all parameters. Mortality rates according to five year age categories were used due to the availability of data.

The NHS DPP specific parameters were obtained from data used for the national evaluation of the programme. Other parameter values were obtained from peer-reviewed literature. The way these parameters were sourced and derived, as well as justifying where choices were made between sources are described fully in Appendix Chapter 4, Transition Probabilities.

Table 4.2: Transition probabilities and distributions.

	State to State transition	Age category (where available)	Point Estimate	Distribution	Source*
Normal glucose tolerance	Remaining in state	*Remainder from other transitions*		-	-
	to Non-diabetic hyperglycaemia	20<x≤39	0.0355	Beta (α=36.1, β=963.9)	Hadaegh et al. (2017) *Using estimates presented in Table 2, page 73*
		40<x≤59	0.0528	Beta (α=54.2, β=945.8)	
		x≥60	0.0742	Beta (α=77.1, β=922.9)	
	to Type 2 diabetes	*Transition not possible*		-	-
	to Death	30<x≤34	0.000674	Constant	National Life Tables (2018-2020) (Office for National Statistics, 2021)
		35<x≤39	0.00101		
		40<x≤44	0.00150		
		45<x≤49	0.00230		
		50<x≤54	0.00336		
		55<x≤59	0.00506		
		60<x≤64	0.00794		
		65<x≤69	0.0124		
		70<x≤74	0.0196		
		75<x≤79	0.0345		
		80<x≤84	0.0609		
		85<x≤89	0.111		
		90<x≤94	0.187		
		x≥95	0.296		

Table 4.2: Transition probabilities and distributions. *(Continued)*

	State to State transition	Age category (where available)	Point Estimate	Distribution	Source*
Non-diabetic hyperglycaemia	to Normal glucose tolerance		0.0795	Beta(α=335.0, β=3709.0)	Balk et al. (2015) *Using estimates presented in Figure 2, page 442*
	Remaining in state	*Remainder from other transitions*		-	-
	to Type 2 diabetes**	x<40	0.0268	Beta(α=27.1, β=972.9)	Linked National Diabetes Audit data (NHS Digital, 2023)
		40≤x<50	0.0339	Beta(α=34.5, β=965.5)	
		50≤x<60	0.0303	Beta(α=30.7, β=969.3)	
		60≤x<70	0.0251	Beta(α=25.5, β=974.5)	
		70≤x<80	0.0201	Beta(α=20.3, β=979.7)	
		x≥80	0.0141	Beta(α=14.2, β=985.8)	
	to Death	Age-specific mortality as described above (Normal glucose tolerance to Death)		Constant	National Life Tables (2018-2020) (Office for National Statistics, 2021)
Type 2 diabetes	to Normal glucose tolerance	*Transition not possible*		-	-
	to Non-diabetic hyperglycaemia		0.00280	Beta(α=2.8, β=997.2)	Karter et al. (2014) *Abstract, results.*
	Remaining in state	*Remainder from other transitions*		-	-
	to Death	Age-specific mortality as described above (Normal glucose tolerance to Death) multiplied by excess mortality		Constant	National Life Tables (2018-2020) (Office for National Statistics, 2021)

Table 4.2: Transition probabilities and distributions. *(Continued)*

	State to State transition	**Age category (where available)**	**Point Estimate**	**Distribution**	**Source***
	Excess Mortality		Hazard ratio: 1.60	LogNormal(μ=0.470, σ=0.064)	DECODE Study Group European Diabetes Epidemiology Group (2003) *Abstract, results.*
Dead	to any other state		0	-	-
	Remaining in state		1	-	-
Effectiveness of the NHS DPP* **			Hazard ratio: 0.80 (95% CI: 0.73 to 0.87)	LogNormal (μ=-0.223, σ=0.0448)	Rathi Ravindrarajah et al. (2023) *Abstract, findings.*

Notes:
*Further details on how parameters were derived from these sources are provided in Appendix Chapter 4, Transition Probabilities.
**Effect of referral to the NHS DPP applied here.
***In the base case, it is assumed that the effect of the NHS DPP lasts for 3 cycles and then stops (equivalent to an adjusted hazard ratio of 1.0).

4.2.5.4 Costs & Outcomes

Table 4.3 shows the costs used for each of the model states along with the sources for these values. Costs considered were from the perspective of the NHS and using UK pounds sterling for a price year of 2020, the mean and standard error of which are reported to 2 decimal places, as is convention with costs, and the shape parameters reported to 1 decimal place. Whilst there is not a state associated with type 2 diabetes-related complications in this model, the cost distribution used for the type 2 diabetes state was sourced from two studies which costed all of the primary and secondary healthcare resource use of a population of individuals with type 2 diabetes. As such it is expected that they will have incorporated a range of disease severities, including those who experienced diabetes-related complications; which will then be reflected in the cost distribution used in this model. Model outcomes were in terms of QALYs. To calculate these, each of the model states were assigned utility scores, which were then used to calculate QALYs depending on the time spent in that state. Table 4.4 shows the utility values used along with the source; the mean and standard error of these values are shown to 3 significant figures, so as to allow the reader (both here and in the submitted manuscript) to recreate the shape parameters for the distributions, and shape parameters are reported to 1 decimal place. A utility score of 0 was attributed to the dead state and assumed to be constant (hence no distribution). Further information on how these sources were selected is available in the Appendix Chapter 4, Costs and Utility Scores.

Table 4.3: Costs associated with each state of the model (2020 price year).

	Cost	Source	Distribution
Referral cost to the NHS DPP	£141.77 Standard error: 34.65	Analysis of provider and contract data provided by NHS England (average cost of referral and costs of implementation) (McManus et al., 2023)	Gamma(α=16.7, β=8.5)
State			
Normal glucose tolerance	£2,005.31 Standard error not reported*	Using resource use estimates from Nichols et al. (2008) and applying NHS specific resource use costs for 2020	Gamma(α=44.4, β=45.1)
Non-diabetic hyperglycaemia	£2,224.76 Standard error not reported*	Using resource use estimates from Nichols et al. (2008) and NHS specific applying resource use costs for 2020	Gamma(α=44.4, β=50.1)
Type 2 diabetes	£4,420.57 Standard error not reported*	Inflated estimate from Kanavos et al. (2012)	Gamma(α=44.4, β=99.5)
Death	0		

Notes:
*Further details on how state costs were derived from these sources are provided in Appendix Chapter 4, Costs.
**The source used to estimate the cost of diabetes, non-diabetic hyperglycaemia and normal glucose tolerance did not present the variation around the cost estimates. As such, a standard error of 15% of the mean is assumed.

Costings for model states include all NHS resource use, not just disease specific expenditure.

Table 4.4: Utility scores associated with each state of the model.

Age category	**Normal glucose tolerance**		**Non-diabetic hyperglycaemia**		**Type 2 diabetes**		**NHS DPP gain**	
	Mean (se)	Distribution	Mean (se)	Distribution	Mean (se)	Distribution	Mean (se)	1- DPP gain Distribution
x<40	0.890 (0.00408)	Beta(α= 5,243.2, β=647.9)	0.840 (0.00245)	Beta(α= 18,852.3, β=3,585.4)	0.776 (0.0684)	Beta(α= 27.9, β=8.1)	0.00431 (0.0000730)	Beta(α= 801,123.6, β=3,464.7)
40≤x<50	0.872 (0.00510)	Beta(α= 3,749.4, β=552.7)	0.829 (0.00171)	Beta(α= 40,397.0, β=8,321.1)	0.674 (0.0544)	Beta(α= 49.0, β=23.7)	0.00609 (0.0000603)	Beta(α= 1,653,302.7, β=10,122.6)
50≤x<60	0.855 (0.00496)	Beta(α= 4,317.4, β=733.5)	0.799 (0.00127)	Beta(α= 79,459.7, β=19,951.7)	0.724 (0.0301)	Beta(α= 159.0, β=60.7)	0.00891 (0.0000545)	Beta(α= 2,945,489.9, β=26,470.7)
60≤x<70	0.844 (0.00582)	Beta(α= 3,275.4, β=605.1)	0.810 (0.000954)	Beta(α= 137,085.2, β=32,219.1)	0.699 (0.0248)	Beta(α= 238.3, β=102.8)	0.0139 (0.0000589)	Beta(α= 3,884,458.6, β=54,584.9)
70≤x<80	0.825 (0.00653)	Beta(α= 2,793.5, β=592.8)	0.814 (0.000866)	Beta(α= 164,314.0, β=37,615.0)	0.745 (0.0219)	Beta(α= 294.0, β=100.8)	0.0153 (0.0000637)	Beta(α= 3,650,006.1, β=56,613.4)
x≥80	0.756 (0.0115)	Beta(α= 1,048.6, β=338.7)	0.781 (0.00158)	Beta(α= 53,445.9, β=14,946.6)	0.671 (0.0272)	Beta(α= 198.9, β=97.4)	0.0107 (0.0000914)	Beta(α= 1,250,791.6, β=13,497.3)

Notes:
The utility scores for Normal glucose tolerance and Type 2 diabetes were sourced from analysis of the 2018 Health Survey for England (NatCen Social Research, 2022).
Non-diabetic hyperglycaemia utility scores and the benefit from programme participation were sources from analysis of the NHS DPP provider data.

The cost of the NHS DPP was derived from data collected by providers of the programme up until April 2020 along with cost data provided by NHS England. These sources were combined to determine the cost of each referral made to the programme. This showed that the average cost per referral received was £118.98 (sd: 117.54), weighted by provider, with an additional £22.79 per referral estimated due to implementation costs of the programme (McManus et al., 2023).

As a probabilistic version of the model was implemented, point estimates were transformed into distributions using their respective means and standard errors. Beta distributions were used for transition probabilities, as they are continuous variables ranging between 0 and 1. Beta distributions were also used for the utility scores, which was appropriate as the health state utility values used were suitably far from zero and a constant utility score of 0 was applied to the dead state, with no distribution applied. In the instance where utility benefit from programme participation was parameterised, as this value was close to zero, a transformation was used of 1 minus the utility score. Gamma distributions were used for model costs given that costs are non-negative and finally a LogNormal distribution was used for the hazard ratios included. The formulas used to derive the shape parameters for each of these distributions are detailed in Appendix Chapter 4, Distributions.

4.2.5.5 NHS DPP Effectiveness

The effect of the NHS DPP was modelled in two ways. The first considered the utility gains obtained directly from individuals who participated in the programme, which are attributed to all individuals for the first cycle of the model. Previous analysis showed that for each session attended an individual gains 0.0042 in utility (95% CI: 0.0025 to 0.0059) (McManus et al., 2023). Provider data was used to determine the average number of sessions attended and then the associated utility benefit calculated. This estimate was then applied to the starting cohort of the NHS DPP in the model.

The second measure of NHS DPP effect considered the long-term effect of the programme in terms of delaying or preventing type 2 diabetes. For this, analysis conducted as part of the wider NHS DPP evaluation was used, which estimated the impact of being referred to the NHS DPP compared to not being referred, using a matched analysis. This analysis found an adjusted hazard ratio of 0.80 (95% CI: 0.73 to 0.87) for developing type 2 diabetes within 36 months (Rathi Ravindrarajah et al., 2023). When modelling the cohort exposed to the NHS DPP, this hazard ratio was applied to the transition probability for transitioning from ‘Non-diabetic hyperglycaemia’ to ‘Type 2 diabetes’. In the base case, this effect was assumed to be maintained for three model cycles, akin to the time horizon of effect within the observed analysis.

For both costs and outcomes, a discount rate of 3.5% was used as recommended in the NICE Reference Case (National Institute for Health and Care Excellence, 2013). A common price year of 2020 was employed for all costs and where applicable, costs were inflated using the annual Office for National Statistics Consumer Price Index (All items) (Office for National Statistics, 2022).

4.2.5.6 Cost-effectiveness Analysis

A cohort of 1,000 individuals was modelled, all of whom begin in the ‘Non-diabetic hyperglycaemia’ state, using a cycle length of one year and run over 35 cycles. To better reflect the reality of individuals referred to the NHS DPP, six age groups were considered (<40, 40-49, 50-59, 60-69, 70-79 and 80+) within this cohort, based on the observed age composition of individuals referred to the programme prior to April 2020. As such, 5.8% of the cohort began at an age of 34 years, 12.2% started at 45, 21.9% started at 55, 26.7% started at 65, 24.4% started at 75 and finally 9.0% started at 84.

The annual cycle length is consistent with the natural history of type 2 diabetes and is commonly used in other modelling studies in this disease area. A within-cycle correction was used (Barendregt, 2014; Gray et al., 2010) to account for the fact that transitions can occur at any point during the cycle and not at a discrete point in each cycle. This was applied by adding the state membership at time *t* to the state membership at time *t + 1* and then dividing by 2, using this value to then multiply by the relevant cost and outcome.

The primary analysis was conducted using probabilistic analysis with 10,000 Monte Carlo simulations. Running a probabilistic analysis as the base case of the model is in line with current recommendations by NICE (National Institute for Health and Care Excellence, 2013) and academic debate (Thom, 2022; Wilson, 2021).

From the results of the 10,000 Monte Carlo simulations, the expected costs and QALYs accrued for usual care and for the NHS DPP were calculated. From this the incremental costs and QALYs gained by the programme were calculated and averaged across the Monte Carlo simulations. The average incremental cost and QALYs associated with each individual for each of these simulations is plotted in a scatter plot to display the uncertainty around the simulation estimates.

The NHS DPP would be considered cost-effective if the incremental cost per additional QALY gained is below the currently accepted willingness to pay threshold. For this analysis, two thresholds were considered of £20,000 and £30,000, as these are the commonly used thresholds by NICE (National Institute for Health and Care Excellence, 2013). The NHS DPP would be considered to dominate usual care, if it incurred less costs and generated more QALYs over the course of the model (or dominated if the converse was true).

The incremental net-monetary benefit was also calculated, estimated as the average incremental benefit multiplied by the willingness to pay threshold, minus the incremental cost of the programme. Here, a positive net-monetary benefit amount would mean that the NHS DPP would be considered cost-effective, with higher values being more cost-effective than lower values. The probability of the NHS DPP being cost-effective compared to usual care was calculated at £20,000 and £30,000 per QALY by counting the proportion of simulations for which the incremental net benefit is positive. From this, cost-effectiveness acceptability curves (CEAC) were also plotted, which show the probability that the NHS DPP is cost-effective compared with usual care for a range of willingness to pay thresholds.

4.2.5.7 Impact of the NHS DPP

Using the results of the modelled cohort of 1,000 along with the number of actual referrals received by the NHS DPP (526,283 referrals received by 31st March 2020) the total incremental costs and benefits were calculated for the whole of the programme.

4.2.5.8 Sensitivity Analyses

Finally, the sensitivity of these results were tested according to several scenarios. Firstly, different levels of effectiveness were considered, in terms of the length of time for which the probability of transitioning to type 2 diabetes from non-diabetic hyperglycaemia was reduced. In the base-case analysis, a hazard ratio of 0.80 was applied to this transition probability for three cycles of the model. This effectiveness was based on what has been observed from analysis of routinely collected data. In sensitivity analyses, two different scenarios for the continued effectiveness of diabetes prevention were considered. The first scenario was based on evidence from the US Diabetes Prevention Programme (US DPP), a large clinical trial evaluating three different treatment groups: placebo, metformin and a lifestyle programme. This study found that there was still an observable effect in diabetes incidence at 10 years following the lifestyle programme (a hazard ratio of 0.66 at 10 years from 0.42 after three years) in comparison to placebo (Diabetes Prevention Program Research Group, 2009). Calculating the equivalent proportional change for the 0.80 hazard ratio observed in the NHS DPP base case, this was equivalent to an effect of 0.88, where it is assumed to be maintained for seven years to parallel what was observed in the US study, after which there was no lasting effect of the programme within the model.

The second sensitivity analysis used subsequent analysis of the US DPP which showed that after 15 years there was still an observable difference between lifestyle programmes and placebo, with a hazard ratio of 0.73 (Diabetes Prevention Program Research Group, 2015). As above, the equivalent proportional change was calculated and it was assumed that this effect was maintained for the subsequent five years, with the hazard ratio going from 0.80 (three cycles) to 0.88 (seven cycles) to 0.91 (five cycles) after which the risk of diabetes returns to normal, equivalent to a hazard ratio of 1.

There were also several sources for model parameters that could have been selected. As such, sensitivity analyses were conducted which used alternative sources for the costs associated with each of the model states, which were used in another published model (Roberts et al., 2018). In this modelling study, the authors used estimates from Hex et al. (2012) to determine the annual cost of type 2 diabetes, and did not include healthcare costs that were unrelated to diabetes or its complications in their base case analysis. They assumed an annual cost of £773.00 (standard error (se): 102.63) for normal glucose tolerance, a cost of £869.00 (se: 104.56) for non-diabetic hyperglycaemia and an annual cost of type 2 diabetes which increased linearly over 15 years from £1,179.00 to £2,939.00 (se: 270.00), using a price year of 2015. As the model used in this case was a Markov cohort model, it was not possible to determine how long an individual had remained in the type 2 diabetes state, instead an average of this range was taken, £2,059. These estimates were inflated to a 2020 price year using Shemilt et al. (2010) (using option ‘IMF’ as the data source for purchasing power parities) to obtain the following annual cost estimates: £853.48 (se: 113.32) for normal glucose tolerance, £959.47 (se: 115.45) for non-diabetic hyperglycaemia and £2,273.37 (se: 298.11) for type 2 diabetes.

The uncertainty surrounding the transition parameter from non-diabetic hyperglycaemia to type 2 diabetes was also examined in another sensitivity analysis. In the base case this was sourced from analysis of routinely collected primary care data and produced a probability of 0.0249. However, this was a lower probability than what had been reported in other studies (particularly clinical trials), and therefore what had been used in other modelling studies. The sensitivity analysis conducted looked at the impact of using two higher rates, the first reported by Herman et al. (2005) who used data from the US DPP to obtain a transition probability of 0.108, and the second used by Leal et al. (2020), who reported an event rate of 80.4 per 1,000 person years for their base case (equivalent to a transition probability of: 0.077). Neither of these estimates presented the different transition rates across age categories, and as such it is assumed that it is the same across all ages modelled.

In the final sensitivity analysis, an alternative source to estimate the utility score of individuals with type 2 diabetes is used. In the base case, the utility scores are estimated from data obtained from the Health Survey for England, using responses from 491 individuals. Whilst using data from the survey enabled a utility score to be estimated for different age categories, the sample size was relatively low and there was no information about the length of time an individual was diagnosed with diabetes or if they were suffering from any diabetes-related complications. Thus instead, estimates presented by Keng et al. (2022) were used, which detailed the health utility score of 11,683 individuals with established diabetes (93.9% with type 2 diabetes and 6.1% with type 1 diabetes). This population had a mean diabetes duration of 16.4 years and 14.7% were recorded as having at least one comorbid adverse event. The utility scores presented by Keng et al. (2022) were not broken down by diabetes type or age category, and thus the mean utility score presented (after imputation) of 0.771 (with standard deviation: 0.221), was used for all age categories in the model.

The different sensitivity analyses are described in Table 4.5.

Table 4.5: Different probabilistic sensitivity analyses conducted.

	Scenario	New distribution
	Changing effectiveness	
SA1	Effect maintained for 3 years (HR 0.800), and then reduced to (HR: 0.883) for the following 7 years, then nothing	First 3 years: LogNormal (μ=-0.223, σ=0.0448) Following 7 years: LogNormal (μ=-0.125, σ=0.0405)
SA2	Effect maintained for 3 years (HR 0.800), and then reduced to (HR: 0.883) for the following 7 years, then reduced to (HR: 0.907) for the following 5 years then nothing	First 3 years: LogNormal (μ=-0.223, σ=0.0448) Following 7 years: LogNormal (μ=-0.125, σ=0.0405) Following 5 years: LogNormal (μ=-0.0977, σ=0.0395)
	Costs from alternative source	

SA3	Using costs estimates used in Roberts et al. (2018) inflated to 2020 price year	Normal glucose tolerance:
		Gamma(α=56.7, β=15.0)
		Non-diabetic hyperglycaemia:
		Gamma(α=69.1, β=13.9)
		Type 2 diabetes:
		Gamma(α=58.2, β=39.1)

	Transition to type 2 diabetes from alternative source	
SA4	US DPP trial data used as parameter source in Herman et al. (2005)	Normal(mean=0.108, sd=0.00950) *Assume the same across all ages modelled
SA5	NAVIGATOR trial data (NAVIGATOR Study Group, 2010) (n=4,661 participants in placebo arm) used as parameter source in Leal et al. (2020)	Beta(α=80.4, β=919.6) *Assume the same across all ages modelled
	Utility scores from alternative source	
SA6	Utility score associated with type 2 diabetes, sourced from Keng et al. (2022)	Beta(α= 32,561.2, β=9,671.2) *Assume the same across all ages modelled

Notes:
HR: Hazard ratio; Sd: Standard deviation.

4.2.5.9 Validation

The mathematical programming of the model was validated by double programming a deterministic version of the model using the point estimates described, in both Microsoft Excel and R, to ensure the same results were obtained. The parameters used in the distributions were checked by plotting the distributions and ensuring that they clustered around the point estimate they were based upon. The model outputs were also compared against results from other modelling studies in the same disease area by comparing state memberships over time to ensure the same general trends were observed.

4.3 Results

4.3.1 Model Results

4.3.1.1 Base Case – Effectiveness observed for 3 years

Table 4.6 presents the costs and QALYs generated according to the two strategies: NHS DPP and usual care. Over the course of the 35 model cycles, the cohort referred to the NHS DPP incurred less costs on average than usual care, and generated more QALYs. This suggests that the NHS DPP dominates usual care. Across the 10,000 Monte Carlo simulations, on average the NHS DPP resulted in cost savings of £135,755 with QALY gains of 40.8, compared to usual care alone (for the cohort of 1,000).

The uncertainty surrounding the estimates of expected costs and effects, is shown in a scatterplot of the incremental cost and QALY pairs from the 10,000 Monte Carlo simulations, comparing the NHS DPP to usual care in Figure 4.3, alongside a willingness to pay threshold of £20,000 per QALY (the black line). The majority of these points fell within the south-east quadrant (86.1%) indicating that the NHS DPP was both cost-saving and generated more QALYs.

The CEAC from the base case analysis is shown in Figure 4.4. This plot shows the likelihood of the NHS DPP being cost-effective at different willingness-to-pay thresholds. The plot begins at 86.1% due to the number of simulations that both generated additional QALYs and incurred less costs. The probability of the NHS DPP being effective at a willingness to pay threshold of £20,000 per QALY generated was 98.1% which increased to 98.4% with a willingness to pay threshold of £30,000.

Scaling up these cost-savings to the number of referrals actually received by the NHS DPP by the end of March 2020 (526,283) equated to an additional 21,472 QALYs generated and cost savings of £71.4 million over the course of the 35 year time horizon.

An example of a cohort trace from one of the Monte Carlo simulations is shown in Appendix Chapter 4, Table A4.6.

4.3.1.2 Sensitivity Analyses

The results of the sensitivity analyses conducted are shown in Table 4.6 and the respective scatter plots of incremental cost and QALY pairs from the 10,000 Monte Carlo simulations shown in Appendix Chapter 4, Figures A4.1 to A4.6, alongside the number of simulations that fell within each of the scatterplot quadrants. Across all scenarios modelled the NHS DPP dominates usual care. The scenario which resulted in the highest on average cost-savings was when the transition probability from ‘Non-diabetic hyperglycaemia’ to ‘Type 2 diabetes’ was updated with the transition probabilities reported in the US DPP trial (Herman et al., 2005) (SA4), which had a higher probability than that used in the base case analysis. This resulted in an average cost saving of £321,383 and an average increase of 90.2 QALYs across the cohort of 1,000. Indeed, 99.7% of the 10,000 Monte Carlo simulations fell within the south-east quadrant of the scatter plot, corresponding to the NHS DPP being both cost saving and generating additional utilities when compared to usual care (Figure A4.3, Appendix Chapter 4). Across the six sensitivity analyses, the scenario with the highest uncertainty and therefore the lowest probability that the NHS DPP was cost-effective at a willingness to pay threshold of £20,000 occurred when using the lower state costs (SA3). In this instance, 97.9% of the 10,000 Monte Carlo simulations were cost-effective at a willingness to pay of £20,000, which increased to 98.4% when using a willingness to pay of £30,000. There were three sensitivity analyses in which all of the Monte Carlo simulations fell below the willingness to pay threshold of £20,000, these were: when the effect was maintained for 15 years (SA2), and in the two instances (SA4 and SA5) when the transition probability from ‘Non-diabetic hyperglycaemia’ to ‘Type 2 diabetes’ was updated using trial data.

Table 4.6**:** Model results for base case and sensitivity analyses (cohort of 1,000 with 10,000 Monte Carlo simulations).

Intervention	**Total Mean costs (£, 2020)**	**Total Mean QALYs**	**Cohort Incremental cost (£, 2020)**	**Cohort Incremental effect (QALY)**	**Mean Incremental Net Monetary Benefit per individual at £20,000 (£30,000)**	**Probability cost-effective at £20,000 (£30,000) WTP threshold**
Base Case						
No intervention	31,998,394	10,765.8				
NHS DPP	31,862,639	10,806.6	-135,755	40.8	951.94 (1360.04)	98.1% (98.4%)
Sensitivity Analyses						
Changing effectiveness						
SA1 – effect maintained for 10 years						
No intervention	31,998,394	10,765.8				
NHS DPP	31,711,537	10,826.4	-286,857	60.6	1,499.55 (2,105.90)	99.9% (99.9%)
SA2 – effect maintained for 15 years						
No intervention	31,998,394	10,765.8				
NHS DPP	31,677,544	10,831.1	-320,850	65.3	1,626.93 (2,279.97)	100.0% (100.0%)
Changing cost estimates						
SA3 – Using costs estimates used in Roberts et al.						
No intervention	14,508,167	10,765.8				
NHS DPP	14,481,464	10,806.6	-26,703	40.8	842.89 (1,250.98)	97.9% (98.4%)
Changing type 2 diabetes transition						
SA4 – US DPP trial						
No intervention	39,687,183	9,648.2				
NHS DPP	39,365,800	9,738.4	-321,383	90.2	2,124.97 (3026.76)	100.0% (100.0%)
SA5 – NAVIGATOR trial						
No intervention	39,101,371	9,823.2				
NHS DPP	38,928,500	9,906.5	-172,871	83.3	1,838.77 (2,671.72)	100.0% (100.0%)

Table 4.6: Model results for base case and sensitivity analyses (cohort of 1,000 with 10,000 Monte Carlo simulations). *(Continued)*

Intervention	**Total Mean costs (£, 2020)**	**Total Mean QALYs**	**Cohort Incremental cost (£, 2020)**	**Cohort Incremental effect (QALY)**	**Mean Incremental Net Monetary Benefit per individual at £20,000 (£30,000)**	**Probability cost-effective at £20,000 (£30,000) WTP threshold**
Utility scores from alternative source						
SA6 – Using type 2 diabetes utility scores from Keng et al. (2022)						
No intervention	31,998,394	10,950.0				
NHS DPP	31,862,639	10,977.1	-135,755	27.1	679.25 (951.00)	98.0% (98.5%)
Abbreviations: NHS DPP: NHS Diabetes Prevention Programme; WTP: Willingness to pay.						

Figure 4.3: Base case analysis.

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 1,295, North-West: 91, South-East: 8,614, South-West: 0. Percentage cost-effective at £20,000 willingness to pay (£30,000): 98.1% (98.4%).

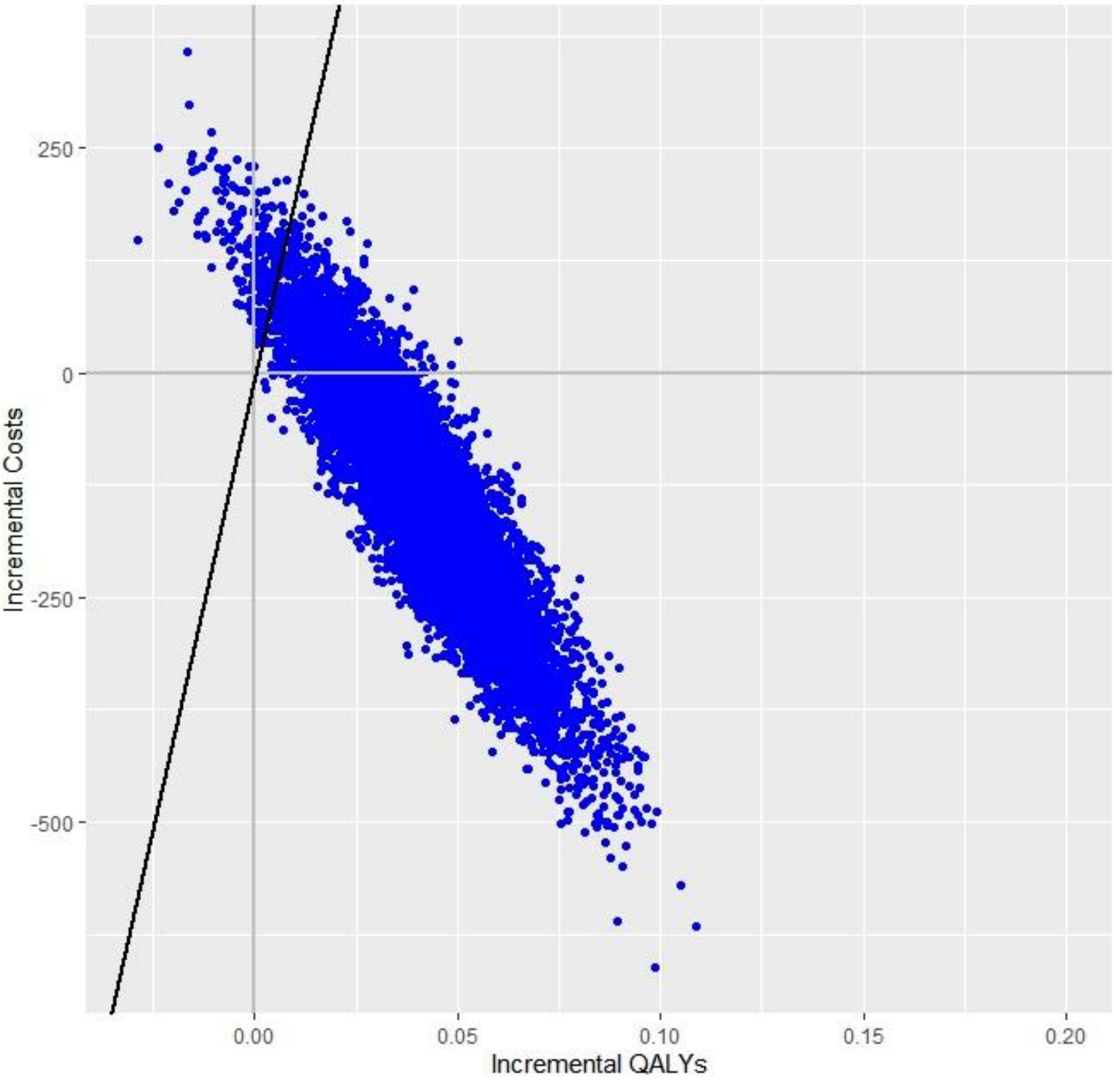
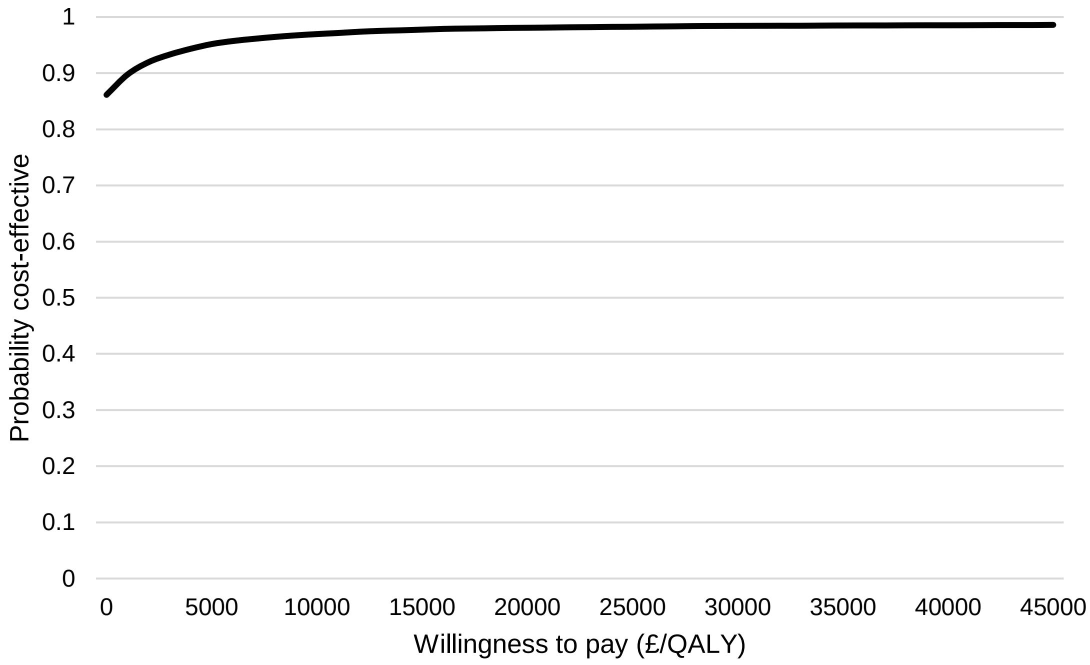


Figure 4.4: Cost-effectiveness acceptability curve for the strategy of NHS DPP compared to usual care (base case analysis)



4.3.2 Reflections on replicable practices

4.3.2.1 Incorporating facilitators

The identified facilitators to model replication were easily incorporated, these included a model diagram showing all transition pathways, along with a table of all of the input parameters which were reported alongside the different transitions.

4.3.2.2 Overcoming identified barriers

Table 4.7 presents what was developed in actuality alongside the proposed method of overcoming the identified barriers, and any positive or negatives from trying to address them. The majority of these were deemed to be successfully achieved, with only some not being incorporated: the development of an RShiny app to go alongside the model (as discussed further below), and providing example costs over short-term model cycles. These cost calculations were not directly included due to the model being probabilistic in the base case, however the average costs used to inform these distributions and cost calculations for how the estimates were derived, along with the costs per model state were reported. The provision of the model code was also thought to reduce the need for such calculations. One proposed method of addressing a barrier was not relevant in this study (providing regression equations). Key aspects of addressing these barriers are discussed in further detail below.

4.3.2.3 Learning R

The process of programming the model in R for an individual who had not used R to develop a model before, was time consuming. During the development process, example code published in a tutorial paper by Green et al. (2023) which detailed modelling a simple three state: sick/sicker/dead model, was adapted to develop the code for this model which had four health states. This tutorial was vital in developing my understanding of R. The learning curve associated with using R (even with the tutorial code as an example) was high and could be viewed as a barrier to this programming language being used by other modellers who do not have prior experience of using R, especially if there is a strict deadline and therefore no time for the modeller to spend learning a different programming language. Learning a new programming language also increased the potential for user error due to not fully understanding model code and how to implement it. This suggests that there is a fine balance between striving towards replicable models using open-source software and delivering a model free from coding errors in a more familiar software. It should also be noted that no one in my research team was familiar with R, which meant that there was nobody available to check the model code. Therefore, in this particular instance, a deterministic version of the model was double programmed in both R and in Microsoft Excel to check the coding of the model and to ensure that the same model results were being obtained. However this was not possible to do in the probabilistic version of the model, which was how the base case and sensitivity analyses of the final model were coded. The model was also intentionally coded without relying on R packages (such as Heemod (Filipović-Pierucci et al., 2017)), so as to ensure that the workings of the model and the code written were understood. Another consequence of being new to programming in R, was that the model code developed may not have been as concise or readable compared to the code of someone proficient in R and knowledgeable of the available R packages. This may then have an impact on the readability of the code by individuals who go on to access the model code, and therefore the transparency, despite efforts to annotate the code throughout. Another implication is that only those with experience of the R programming language will be able to inspect the code, which may reduce its usability.

4.3.2.4 Time constraints of the project

As this model was developed as part of funded work on a research project, there were time constraints relating to the end of the research grant. This meant that there was limited time to finesse the model code and ensure aspects of replicability once the model had been developed, given the project deadlines and the need to submit the manuscript in time for the final report submission (for example, in terms of ensuring the code was concise or enhancing readability to an external user). Other aspects that could have been developed included developing an RShiny (Smith & Schneider, 2020) version of the model embedded in a web browser. This would have enabled users who may not have any programming knowledge in R to change the parameters and rerun the model according to different scenario analyses. However in order to develop this it would have required substantially more time as well as learning another coding language (C++). Further transparency could also have been built into the model by using RMarkdown (Xie et al., 2018), both to structure the code of the R script and also to produce the report directly from the code. This would also have been another aspect of programming to learn and would have required additional time. As such, it was not possible to completely exploit the ‘end to end’ functionality of R, which is where the analysis is conducted (for say transition probabilities), the model is run and then the output is collated in a report, all in one script. This was partially due to the time constraints of the project, and also due to the way some of the data for the project was housed in a secure environment due to NHS Digital requirements. This meant that some analysis was performed in the secure environment and then the results exported.

4.3.2.5 Incentives and Motivating factors for reproducibility

Other than being motivated by interest in reproducibility and the knowledge that it was contributing to the thesis, there were actually very few incentives or motivators to conduct the work in a reproducible manner. The original journal the manuscript was submitted to (BMC Medicine) had no requirements that any reporting checklists were completed as part of the submission, and as such a checklist like CHEERS was not completed. This suggests that unless mandated by the journal, researchers under time constraints may not think to complete such reporting checklists. As part of the peer review process, the absence of a completed reporting checklist was not commented upon, nor did they comment on the efforts made to use an open-source programming language, the publishing of the model code or indicate that they had accessed the code to check that it was functional. The manuscript was unfortunately rejected from this first journal, and resubmitted to another (The European Journal of Health Economics). Similarly, this journal did not require any reporting checklists to be completed. Subsequent to these submissions, the CHEERS checklist was completed and can be seen in Appendix Chapter 4, Table A4.7.

Table 4.7: Barriers of model replication and how they were addressed, colour coded as to whether they were achieved or not: achieved (green), partially achieved (amber), not achieved (red), not applicable (grey).

Barriers	**Planned Deliverables**	**Actuality**	**Discussion/learning points, positive**	**Discussion/learning points, negative**
Work flow	• Model to be programmed in an open-source software, such as: R. • Model code to published, for example in a Github repository.	• Model developed in R. • Model code published on Github.	• R is open-source, and publishing code on Github means it can be openly interrogated. • R presents the model code linearly, in a step-wise format, which may be easier for the user to understand.	• Significant learning curve associated with R. • Lack of familiarity with R may have led to coding errors (if I had not have double coded the model in Excel). • R may be considered to lack transparency by some users (due to the learning curve). • Some HTAs do not accept submissions in R – so from the consultancy perspective, R models are not that useful. • Considerable time was spent programming the model in an open-source software. This might not always be possible given the time limits of the project.

Table 4.7: Barriers of model replication and how they were addressed, colour coded as to whether they were achieved or not: achieved (green), partially achieved (amber), not achieved (red), not applicable (grey). *(Continued)*

Barriers	**Planned Deliverables**	**Actuality**	**Discussion/learning points, positive**	**Discussion/learning points, negative**
Authors unable to be contacted – requirement for authors to archive and annotate their work	• Publishing code on Github would mean that this was open to interrogation / and would not be dependent on the author responding to requests.	• Model code published on Github.	• Github repository is linked in the final publication and code is annotated along with a 'Readme' file.	• There is a reliance on Github being maintained as a coding platform. • Github requires the user to create a (free) account to access model code, this might put some people off from accessing it.
	• Development of an R Shiny App.	• R Shiny not developed.	• Not applicable.	• There was insufficient time during the project to develop an RShiny version of the model. The learning curve represented by R was steep enough, let alone learning another programming language (C++).
Lack of transparency around costs used	• Table of costs used for each model state.	• Table of costs provided, along with shape parameters for distribution.	• This was easy to do.	• Commercially sensitive information surrounding costs by private providers, this meant the costs were

Table 4.7: Barriers of model replication and how they were addressed, colour coded as to whether they were achieved or not: achieved (green), partially achieved (amber), not achieved (red), not applicable (grey). *(Continued)*

Barriers	Planned Deliverables	Actuality	Discussion/learning points, positive	Discussion/learning points, negative
				a weighted average across providers. • If the model is transparent in all other respects, it may be that the model can easily be interrogated through a series of sensitivity analyses, meaning the costs can be calculated.
	• Summary of costs for a shorter time horizon.	• The final model did not explicitly describe the costs over a shorter time horizon.	• It was assumed the open model code would negate the need for this.	• This comment fits more with deterministic models. When running a model that is probabilistic in the base case (with 10,000 simulations), it would only be possible to provide a few examples of the cost traces over time.
Lack of transparency around model parameters	• Table of parameters grouped for each of the scenarios modelled.	• Table of model parameters provided, split across the different age.	• This was easy to do and incorporate in the manuscript using a table in the form of	• There is potential that whilst I may think that I have reported all of the model parameters,

Table 4.7: Barriers of model replication and how they were addressed, colour coded as to whether they were achieved or not: achieved (green), partially achieved (amber), not achieved (red), not applicable (grey). *(Continued)*

Barriers	Planned Deliverables	Actuality	Discussion/learning points, positive	Discussion/learning points, negative
	- Diagram for each scenario, along with all possible transitions documented.	categories modelled - Diagram of model included (although this did not change across scenarios). - Example state membership over time given for one of the PSA iterations included.	state transition matrix to make sure that all possible transitions were included.	some may be omitted.
Inability to recreate PSA due to lack of reporting of standard deviations or distribution parameters	- Ensure all shape parameters for PSA are reported.	- All parameters were detailed in a table in the manuscript.	- This was easily incorporated.	- None.
Lack of reporting of clinical event results	- Present event and mortality results (for all simulated alternatives) over the whole time horizon of the model, for each model cycle.	- An example cohort trace was provided for one of the Monte Carlo simulations.	- None.	- This point is less appropriate for models that are run as probabilistic in the base case.
Self-created regression analyses	- State all regression methods used, and provide details on how to apply/solve the	- This was not applicable to the model developed.	- Not applicable.	- Not applicable.

Table 4.7: Barriers of model replication and how they were addressed, colour coded as to whether they were achieved or not: achieved (green), partially achieved (amber), not achieved (red), not applicable (grey). *(Continued)*

Barriers	**Planned Deliverables**	**Actuality**	**Discussion/learning points, positive**	**Discussion/learning points, negative**
	regression equations correctly.			
Lack of reporting of details on Life Tables	• Ensure that adequate detail is provided around the Life Tables, including the year used.	• Where life tables were used, the version and web address for the resource were included in the manuscript.	• This was easily incorporated.	• None.

4.4 Discussion

This chapter sought to explore the practical implications of researchers developing replicable decision models, particularly in incorporating the facilitators and overcoming the barriers identified in the previous chapter. The learning points of trying to practically incorporate these into decision model development are vital to consider if replicable research practices are going to be recommended and pursued more generally going forward. This chapter produced several key findings: the steep learning curve in developing a model in a new, open-source, programming language, the additional time required to develop models with replicability and transparency in mind and the lack of motivators or incentives for researchers to do this. The facilitators and proposed solutions for the barriers were largely easily implemented, however some of them were no longer relevant given that the model code was openly published.

Whilst programming the model in an open-source programming language meant that it may have had greater accessibility and would allow individuals to see the step-wise development of the model, it was not without its challenges. Leaning a new programming language was particularly time-consuming and may have led to more mistakes in the code, given lack of user proficiency. This experience of using R to develop model code appeared to be similar to that of other researchers who looked to compare the use of R and Microsoft Excel when creating models. In the publication by Xin et al. (2021), a group of authors sought to replicate a published Microsoft Excel model in R. The authors reflected that learning the programming language of R was challenging and that the process of modelling in this new environment represented a time trade-off between the familiarity of working in Microsoft Excel and the learning curve required to program the model in R. Finally, these authors also remarked that the R code they used to develop the model had scope to be made more “elegant or efficient” but cited their inexperience and lack of time as a barrier in doing this. Similarly, Hart et al. remarked that developing an R model presented coding challenges that were not present when developing the Microsoft Excel model, and as such the development of the model in R took “significantly longer” to build (Hart et al., 2020).

The development of a replicable model was also found to be a lengthy process. It may not be practicable for researchers to prioritise replicable practices, given the often strict time constraints of funded research projects. Whilst some of this time would be reduced once researchers were more familiar with open coding practices and using new programming languages, there would still be time spent making the code transparent, well coded and accessible. This risk was also identified by Sampson et al. when they considered the benefits and risks of incorporating transparency into decision models (Sampson et al., 2019) , and also by Harris et al. who conducted a survey of public health analysts and found that 41.7% of those surveyed cited a lack of time as a barrier to reproducible practices (Harris et al., 2018). Given the increased time required, there is a financial cost associated with replication. Although this cost might be offset by the reduced researcher effort amongst those that then go on to use the open-source code or model. To ensure researchers have the time to dedicate to reproducible research methods, researchers must be given either explicit incentives to make their model code replicable, or alternatively have time built into the project in order to concentrate on the write up of model code and the automation of reports. One solution might be for universities to include replication efforts as a factor considered in promotion. Alternatively, funding bodies could incorporate funded time for researchers to finalise their code and ensure that it is transparent and readable to external users, however, this would require explicit commitment from the funders of research. This would be an important commitment beyond the current transparency requirements of some funders to publish in open-access journals. The potential of research funding was highlighted as an area of possible intervention in a scoping report looking at how to increase reproducibility of scientific results in the European Union, describing it as a "great potential … lever to increase reproducibility" (Lusoli, 2020), alongside other potential interventions such as the development of guidelines and increased researcher training. Another paper suggested this too, advising that grants could include and fund "data curation time, expertise in developing reproducible and transparent research workflows and infrastructure for data curation" (Stewart et al., 2021).

When considering the implementation of the facilitators to replication identified in Chapter 3 and by Schwander et al. (2021), the majority related to clear and specific aspects of reporting. Several checklists exist to improve reporting thoroughness, and many journals, although not all, require these (particularly CHEERS) to be submitted as part of the submission process. However, authors of publications often state that an item is reported when in actuality it is not (as shown in the replication case studies conducted in Chapter 3 and by the case studies conducted by Schwander et al. (2021)). This might be due to several reasons, one of which could be author fatigue, as by the time a manuscript has been prepared and is ready to be submitted, the checklist is often just a part of the submission process, at which point they are focused on achieving the submission and may rush its completion. Alternatively, it may be that authors are so familiar with their work that they are not able to see what they have and have not included due to their high level of involvement. A suggestion to rectify this, may be to require peer reviewers to complete a reporting checklist as part of their appraisal. Not only will the peer reviewers be adequately removed from the publication in that they are able to see what is and what is not reported, but they will also have less of a vested interest in saying that an aspect is reported. The journal could then include this feedback in the reviewer comments and ensure that any items that were inadequately reported were then incorporated in the manuscript revisions. A caveat of this approach is that peer reviewers are unpaid and may fill in the requested checklists to different degrees of meticulousness. Furthermore, adding another aspect to the often lengthy review process may make peer reviewers less likely to accept the review invitation. The completion of checklists could also be modernised, by using online tools where individuals click through the different criteria, with a link to the online tool being included in the peer review interface. An online tool would then allow the responses to be automatically included in any revision requests to the author. An example of this is the recently developed “Criteria for Health Economic Quality Evaluation” (CHEQUE) tool, which is an online tool which can be used to appraise reporting quality (Kim et al., 2023). However to my knowledge this tool has yet to be embedded into any journal peer review process.

Similarly, whilst model code is made available, it may not actually run. Whilst some journals now have data editors who have the role of checking that the data deposited can be opened and that associated code runs, this is not commonplace. Alternatively there are several initiatives which verify code can be run, for example, Code Check (CODECHECK, 2023), where individuals submit their code and receive a certificate if the code can be independently run.

4.5 Strengths and limitations

To my knowledge, this is the first study to consider from the modeller’s perspective the implications of replicable research practices and the practicalities of implementation. In developing the model alongside a funded research project, this study is likely to capture the practicalities that may have been missed if attempting only to develop a model purely with the goal of increased transparency and replicability. With that being said, there are also some limitations. The first being that these were only the experiences of the author. As such, if another researcher within a different research team or project attempted a similar task, it may be that they would have had an entirely different experience, finding it either to be more feasible or potentially encountering other barriers that have not been observed in this study. As well, the lack of familiarity with R and the barriers encountered, may be self-correcting amongst future health economists, as more universities may start to incorporate R into their curriculum and training courses emerge relating to using R in HTA (R for Health Technology Assessment, 2023). Alternatively, other software and programming languages may become more fashionable, leading to similar problems being encountered in the future.

4.6 Conclusion

Whilst this work has shown that it is possible for researchers to develop a decision model with replicable research practices in mind, it has highlighted that this is not without a substantial investment in researcher time and effort. In order for these research practices to become widespread, more motivators are needed. These may include the funding of researcher time explicitly in the grant to write up model code in a transparent way or in the form of additional researcher training to facilitate the development of models in open-source programming languages.

Chapter 5 Discussion

5.1 Introduction

This thesis aimed to examine the role and value of replication in health economic decision models. The concept of replication has been discussed widely in other scientific disciplines, such as biomedicine (Iqbal et al., 2016), computational science (Peng, 2011; Rougier et al., 2017), psychology (Makel et al., 2012) and epidemiology (Peng et al., 2006). However, how replication might apply and its value to decision modelling has been less well explored. This is of importance for several reasons. Decision models are used to inform health policy and the funding decisions surrounding treatments and interventions, yet they are also often seen as 'black boxes' that lack transparency. Decision models may also be vulnerable to manipulation for perceived personal or commercial gain (Z. M. Khan & Miller, 1999). Possible manipulation paired with the potential for perverse incentives to have models show a certain result, may mean they are reported with a lack of transparency. This lack of transparency may be particularly problematic regarding the motivation for sourcing model parameters, which may be influential to the model results obtained.

In addressing this overarching topic, this thesis sought to answer the following research questions:

- RQ1. Why is replication needed?
- RQ2. How do other scientific disciplines approach replication?
- RQ3. What is the role of replication in decision models within health economics research?
- RQ4. How could a successful replication be defined?
- RQ5. What are the barriers and facilitators to replicating decision models?
- RQ6. What are the implications of a model being replicable (or not)?
- RQ7. What are the implications and challenges for modellers trying to incorporate replicability?
- RQ8. Does the ability to replicate lead to greater transparency?

The thesis includes three original studies, all of which incorporate a discussion and a strengths and limitations section. This final chapter will summarise the key findings of each of the chapters of this thesis, as well as present the overarching discussion points, implications for future health economic analyses and the overall strengths and limitations of the thesis work. Finally, the chapter will suggest potential areas for further research.

5.2 Summary of thesis

5.2.1 Chapter 1 Summary

- RQ1. Why is replication needed?

Chapter 1 introduced the general concept of health economics, with a particular focus on decision modelling. It discussed the different types of decision models commonly used, along with the scenarios in which certain types of models may be best implemented. It also presented the concepts of research transparency and replicability and described the current initiatives ongoing within the discipline of health economics to facilitate transparent and replicable research. These include, but are not limited to, the call for modelling registries, the use of health economics analysis plans and exercises in comparative modelling, whereby modelling studies use the same input parameters to highlight differences in results obtained.

5.2.2 Chapter 2 Summary

- RQ2. How do other scientific disciplines approach replication?
- RQ3. What is the role of replication in decision models within health economics research?
- RQ4. How could a successful replication be defined?

Chapter 2 reviewed the existing literature to determine how replication and the concept of ‘replication success’ had been defined across scientific disciplines. This was the first published review of its kind in any discipline (McManus, Turner, & Sach, 2019). Whilst there was considerable literature discussing the concept of research replication, there was little that explicitly proposed how to define a successful replication and what this entailed. This may suggest a reluctance amongst researchers to label replications as a success or failure. This could be due to the potential reputational damage or alienation of colleagues if a replication is found to not reproduce the same results, or fear from the replicator since their skills may also come into question if different results were found. Replications may also not yet be reported for replications sake (and hence not require a definition of success) and instead be used as a way to develop other models. This review also highlighted a paucity of literature on replication within the field of health economics, although there was discussion around other transparency initiatives, such as modelling registries.

Using the definitions identified within other scientific disciplines, this chapter concluded by proposing six definitions of what might constitute a successful replication of a decision model. These ranged in specificity, from broad definitions to the narrowest requiring an exact replication of model results. Whilst several definitions were proposed, it was not possible to identify a single definition, given that the definition of a successful replication may depend on the motivation for the replication and importantly, the need for greater engagement from the wider health economics community.

5.2.3 Chapter 3 Summary

- RQ5. What are the barriers and facilitators to replicating decision models?
- RQ6. What are the implications of a model being replicable (or not)?

Chapter 3 described five case studies that sought to replicate published decision models. In doing so, it aimed to understand the current replicability of published decision models and to identify the common barriers and facilitators of replication. This chapter also evaluated the definitions of replication success proposed in Chapter 2 and sought to update these based on the replication findings. The replication studies identified several common barriers and facilitators to model replication. Common barriers to replication included: the presentation of conflicting information; that not all parameters were adequately described; that longer time horizons amplified any discrepancies between the replicated model and the original; authors not being contactable regarding their original work and that too much time had elapsed since the original publication. Facilitators included: the provision of clear model diagrams detailing all possible transitions and providing example cost calculations. The replication case studies also demonstrated that reporting checklists such as CHEERS (Husereau et al., 2022; Husereau et al., 2013) or Philips (Philips et al., 2004) may not be able to pick up on reporting nuances, given that the majority of the five case studies appeared to be well reported but were still missing key details needed for model replication. Therefore, checklists may not be sufficient to ascertain if a model is replicable or be an adequate tool for ensuring research transparency or the thoroughness of reporting required for model replication.

When implementing the proposed ‘replication success’ definitions, all five case studies satisfied the broadest of definitions, but none satisfied the narrowest which required an exact replication. It became evident that the model results influenced whether a definition could be used. One of the proposed definitions was found to be not applicable to four out of five of the case studies as it related to ICER variation and the model results found the intervention to either be dominant or be dominated by usual care. The most workable definition of those proposed in Chapter 2 was then adapted in Chapter 3 as follows:

> “Results for the costs and outcomes vary by only XX% compared to the original for the majority of the treatment pathways replicated, AND are consistent with the original conclusions.”

Here the percentage of variation (XX) is still to be determined, depending on wider community input.

5.2.4 Chapter 4 Summary

- RQ7. What are the implications and challenges for modellers trying to incorporate replicability?

In Chapter 4 a decision model was developed with a focus on trying to develop it in a manner that supported it being replicable. This contrasts with the previous thesis chapters, which evaluated the replicability of existing research models. This gave a different perspective, evaluating whether current replication initiatives along with the suggested facilitators and solutions to barriers identified in Chapter 3 were achievable from the perspective of the researcher developing the model.

The model construction was undertaken within a funded research project to mirror the experience of researchers more generally. In doing so, this chapter identified several issues, particularly relating to the use of open-source modelling software or programming language and the steep learning curve associated with this, along with the time required in making model code available and ensuring that it was readable to other users. Due to the significant time investment, this work identified that buy-in from research funders is required to fund researcher time to devote to such transparency initiatives. Without explicit funding and given the conflicting demands on researcher time to begin other projects, it is unlikely that researchers would engage in these additional steps to achieve transparency and replicability, which has been seen to date. Notably, the development of the model within this chapter highlighted the lack of incentives or motivators for researchers to engage in research transparency. For instance, there was no explicit push for developing the model using open-source software from the wider research team. There was also no mandated submission of a completed reporting checklist from several of the high-impact journals to which the research was submitted. Finally, it was found to be feasible to overcome the barriers and implement the facilitators identified in Chapter 3, however some were no longer relevant given that the model code was made publicly available.

5.3 Reflecting across the thesis

The work presented in this thesis represents an integrated body of work exploring the role and value of replication in decision models, with each chapter building upon the findings of the previous. Considering the combined results of these chapters, there are several common themes that have emerged. The first is that reporting checklists may be inadequate to discern if a model is sufficiently reported to allow replication. Secondly, the importance and need for a more developed research infrastructure to drive forward the replication initiative, as current research culture does not encourage or incentivise transparency and replicability. Finally, that replication is not a stand-alone concept, but one that must be employed alongside other transparency initiatives. Using these findings, it is possible to consider the overarching value of replication and what replication case studies are able to contribute to the transparency of health economic decision models. These themes are discussed in further detail below.

Reporting checklists are one of the more widespread approaches to increase the transparency of research, with checklists such as CHEERS (Husereau et al., 2022) being commonly recommended for the reporting of health economic research. Indeed, the CHEERS checklist was updated in 2022 and now includes a prompt to report if the model “is publicly available and where it can be accessed” in a further nod to the importance of transparency, although notably this is within a modelling item, rather than a standalone requirement. Whilst such checklists have a role in ensuring that key aspects of research methodology are reported, this thesis has shown that they are inadequate in assessing whether models are sufficiently reported to facilitate model replication and therefore are unable to fully assess if research transparency is achieved. This is because they do not specify the level of detail required and are unable to detect for example, if all parameters are sufficiently reported (as encountered in Chapter 3), which may only be discovered when a replication is attempted. These checklists are also not routinely mandated by every journal, as identified during the journal submission process of the model developed in Chapter 4. Whilst the importance of reporting transparency is accepted by the majority of researchers, there appears to be less consensus surrounding the role and value of replication, as shown by the relative lack of literature on the topic within health economics (Chapter 2) which may be why the role of checklists in facilitating replication has been less well explored.

Another key finding of this thesis is that in order to move towards replicable health economic research, more motivations or incentives for researchers are required. Currently, replicability appears to be a focus of a few select researchers with a passion for the topic, but it is not a widespread expectation or something that is routinely engaged with by the broader community with any sincerity or importance. This may be due to the high cost associated with conducting replications, for example in terms of researcher time, although the benefit of greater research transparency and the learning points from such studies may be worth such costs. In order to make this a more mainstream endeavour, changes in the current research culture and infrastructure are required. Presently, the focus of researcher time is on generating original outputs such as publications and impact. Due to this focus on metrics, there is potentially less time for researchers to focus on how the research is achieved and on research integrity. This is seen in the research promotions criteria, which focuses on the above metrics, rather than specifically on good research practice (McKiernan et al., 2019). In example, one study conducted a survey amongst researchers to identify factors they perceived to affect their research integrity (Vitae, 2020). This review found that journal impact factors, institutional workload models and how researchers are assessed for promotion were all perceived to have a strongly negative impact on their research integrity. Elsewhere, it has been suggested that incentives for research transparency should also be incorporated within the job-hiring phase (Gernsbacher, 2018), with candidates being asked to demonstrate how they have engaged in transparency initiatives within their research and being assessed against such criteria, which may be akin to the recent focus on demonstrating research impact.

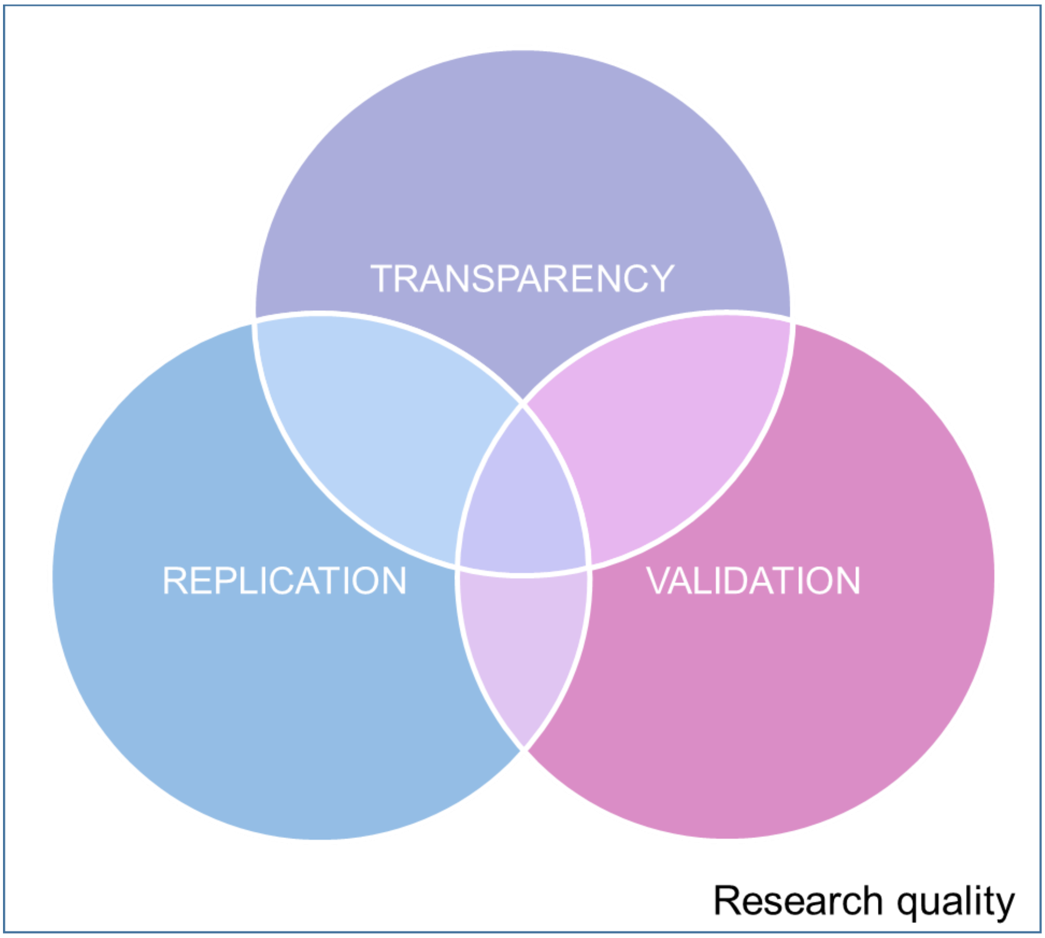
Not only did the work within this thesis suggest a need for a change in the overall research infrastructure and culture, but it also demonstrated the role of research funders and to lesser extent journals in facilitating replicable research. Such changes could come in the form of explicit funding for researchers to write up their analysis using open-source software, or time to improve the readability of their code before uploading to coding repositories. This would need to be explicit in the funding rather than incorporated in the general time allocated to develop the model, given the learning curve associated with developing such materials and the conflicting pulls on researcher time, which would mean it could easily be lost to other aspects of the research. This is especially true towards the end of research projects (which may be when researchers focus most on transparency and replication efforts) as focus may start to drift towards upcoming projects. Alternatively, research funders could mandate that all code be uploaded in an open-source framework, which would ensure that researchers engage in more replicable research practices. This may be especially relevant for research funders that use public money (such as the NIHR), to ensure that the methods as well as the results of the research can be used by others, hence helping to ensure the best value for money. Whilst the open sharing of code and results is currently mandated by some journals in wider economic disciplines, such as the American Economic Review (American Economic Association, 2020), there is yet to be such a policy amongst any of the main journals publishing health economic decision models. Furthermore, researchers have a choice not to submit to these more stringent journals, which may reduce the impact of these initiatives, as they could opt to submit to a journal without such policies. Outside of health economics, whilst some journals require the uploading of data and analysis code, there is variation in the quality and readability of such code, with one study finding that the uploaded data and code did not even run (Trisovic et al., 2022) and another study finding that authors refused to provide code despite it being mandated (Wood et al., 2018). Journals also have a role in allowing for the publication of supplementary materials, which were shown in Chapter 3 to help facilitate replications. Moreover, they may also be able to influence how replication is received, by publishing replication studies, which would increase awareness of the concept amongst their readership and also show researchers that their replication efforts can lead to valued research output.

This thesis sought to assess the role and value of replication within health economic decision models. The combined findings from the empirical work suggest that replication is of value and needed when assessing the current standard of reporting transparency for decision models, with replication highlighting areas that may need to be reported more consistently and thus helping to inform reporting guidelines and practices. Currently research transparency is most often assessed by reporting checklists, however this work has shown that such checklists fail to identify if a decision model is reported thoroughly enough to facilitate replication and therefore currently published works may be lacking in areas of transparency. The value of replication is both in highlighting general trends for the improvement of research reporting, and at a more individual level, potentially highlighting coding errors within the model. Replication may also be an informative endeavour to the replicator, allowing them to learn from the methods used by the original author and help to give an understanding of how aspects of the model need to be reported in order to facilitate replication, which may help to improve their own work. As well, conducting replications can help to speed up the research process, allowing the replicator to build on what others have developed, rather than developing a completely new model for each research question.

Whilst it has been shown that conducting replications may be a useful endeavour to the individual researcher and can suggest the general state of transparency, it is unclear how individual replication studies are used by other researchers (outside of the replicator and the original author). This may be due to the lack of replications currently conducted within health economics or being published under such a label, as it is still a new and emerging field. It is likely that some replications are carried out when researchers are in the process of developing a new model, but that they might not be published or branded as such, making them difficult to identify in the literature. However, a study within psychology looked at how replications were used alongside original studies and found that less than 3% of articles citing the original studies also then cited the replication attempt (von Hippel, 2022), regardless of whether the replication study was a success or failure. This might suggest that individual replication studies lack value as individual pieces of work and that further work may be needed in broadcasting the results of replications, encouraging their use and potentially the need to link the replication and the original work so that the replication is identified alongside the original study when readers access it. One way around this, may be to include a replication attempt or analysis check within the original paper’s supplementary material and have the replicator be included as an author in turn allowing them to get recognition for their work. This would make sure that the replication is highlighted alongside the original paper and also allow for the replication to identify any errors or areas where the reporting thoroughness is lacking, before the paper was published. However, this would require a replicator to be invited to replicate by the original authors (perhaps from the same institution), and would also rely on the original authors waiting for the replication to be conducted before publishing their work, which they may be reluctant to do. This method would also lack external validity, given that the replicator is likely to have links to the original authors and therefore may face pressure to report a successful replication.

It is also important to emphasise that replication is not a standalone or fix-all concept, but is something that needs to operate within a wider agenda of other research integrity and quality initiatives. Whilst a model may be perfectly replicable based on the information presented, it may be lacking in other areas of transparency, for example the description of why one source for a parameter was used over another, this would enhance the overall transparency but not directly impact on the replicability of the model. This is also the case with validation. Whilst replicating a model will show that the model is internally valid, it cannot inform other aspects of model validity, such as whether the model is clinically valid or speak to the external validity of the results. The converse is also true, whilst a model may not be replicable, it does not mean that it is not clinically valid or a good representation of the decision problem, merely that the methods have not been sufficiently reported. In this way, the three concepts may be viewed as an overlapping Venn diagram sitting within the wider concept of research quality, with some of each element covering the other, but neither concept fully encompassing another (as shown in Figure 5.1).

Figure 5.1 Example overlap of the concepts of Transparency, Replication and Validation in health economic modelling.



5.4 Strengths and limitations

A strength of this thesis is that it is one of the first to explore the concept of replication as a distinct topic in the field of health economics. It helps to address a considerable research gap. This thesis implemented an iterative research approach, where chapters built on the work of the previous and revised any recommendations made.

Another key strength of this thesis is that it evaluated the concept of replication from the perspective of those using research and those developing it. This ensured that any of the recommendations made in earlier chapters could be feasibly implemented.

There are also some limitations of the thesis. Whilst this thesis pertains to the replicability of decision models in general, it has inadvertently focused primarily on state-transition models, with the majority of case studies in Chapter 3 being Markov models, and the model developed in Chapter 4 also being a Markov model. Thus, this thesis may not have captured findings from bespoke elements relating to more complex models, such as discrete event simulation or individual patient level simulations. Although, it could be argued that if issues with replicability were found within more simple models, they are also likely to be relevant to more complex models.

Furthermore, this research has focused on modelling studies published within academic journals, but it has not explored the role of replication for decision models developed as part of the NICE HTA process. In this, models are submitted in executable form to NICE as part of a technology appraisal, which are then interrogated by an evidence review group or an independent Technology Assessment Review (TAR), who then provide a review of the evidence. As part of this review, the "reliability" of the model is tested, which may involve conducting other exploratory analyses using the model, although aspects of the model may be redacted depending on commercial sensitivities (National Institute for Health and Care Excellence, 2018). Transparency and the value of replication may be warranted even more so in models that are developed using public funds and that will ultimately inform how health budgets are spent, where one might reasonably expect that they should be open to scrutiny.

This thesis has also focused on replication in its purest form; that is to directly replicate reported results (which is known as 'narrow' replication). However replication can also be considered in terms of a 'broad' replication which looks to test the robustness of results and use other data sources or assumptions to test whether the empirical findings can be repeated and are generalisable (Bettis et al., 2016; Pesaran, 2003). This focus may be considered a limitation, as there may be additional value in broad replications that are not captured within this thesis. An example of the broader concept of replication being considered has been described earlier in the thesis with the Mount Hood input checklist, where the impact of using a set of input parameters across 11 decision models was evaluated to see the variability in results and whether they would give similar conclusions regarding the cost-effectiveness of three interventions (Altunkaya et al., 2023). The thesis has also not explored the extent to which replication already takes place under the guise of model development (by building on existing models), nor have the chapters above explored the value of replication with this solely in mind. The use of replication in this way would be harder to identify in the literature, but doing so would provide important additional insight into the role of replication which is not currently captured here.

Moreover, this thesis has focused solely on the value of replication within decision modelling, and makes no assessment on how replication relates to the wider discipline of health economics such as studies relating to economic evaluation within clinical trials or elicitation exercises. This was beyond the scope of the thesis.

5.5 Implications for policy and research

The findings of this thesis have several important implications for research. There is a need for greater involvement and engagement by research funders if they consider it important to encourage replicable research. This could be in the form of mandating replicable health economic research practices, for example ensuring that researchers who receive funding to develop decision models openly share the models. Although careful consideration would be needed about the implications of this regarding intellectual property, to ensure that the original researchers retain the credit for their work, as highlighted by Padula et al. (2017) and Sampson et al. (2019). Another important implication is that time should be built into research grant expenses to allow researchers to further develop their code so it is readable and executable by others.

An important implication of this research for health service funders and for policy makers, is that current models that are used to inform health policy and the funding of medications may not be replicable. Whilst this may be for innocuous reasons, such as insufficient reporting of a model or replicator inability, it could also be due to coding errors within the published model, which would mean that funding decisions are being made based on inaccurate results. As per the findings in Chapter 3, the replicated costs had greater variation than the health outcomes.

Some of the findings of this thesis have already contributed to academic debate regarding the value of replication, with the definitions proposed in Chapter 2 being used and adapted by other health economists (Schwander et al. (2021)), as well as being cited by other papers exploring transparency initiatives (Hamilton et al., 2023; Imam, 2022; Otten et al., 2023; Schwander et al., 2022; Zawadzki & Hay, 2020).

5.6 Future research

A major finding of this thesis is the need for researchers to be motivated to engage in transparency initiatives that may facilitate replication. The logical next steps would be to engage with research funders to argue for time for transparent practices to be built into grant proposals and for sufficient funding to be allocated to support health economic researchers with such endeavours. This would help to move beyond the current transparency initiatives such as journal open-access, which whilst important, have been shown in this thesis and in other studies, to fall short of ensuring complete transparency and replicability (McCullough et al., 2006; Wood et al., 2018). Another way of increasing researcher awareness of the importance of replication, would be for replication studies to become more commonplace. Initiatives to help with this could be akin to the Replication Games, which are hosted by the Institute for Replication (Institute for Replication, 2023). These games bring together teams of researchers and tasks them with replicating studies within economics. However, most of the studies chosen for replication come from journals where it is required to upload data and analysis code. Therefore there is a greater focus on the broader replicability of the research rather than direct ‘narrow’ replications. A similar initiative could be started to increase awareness of the importance of replication within health economics research.

Research is also needed to obtain a consensus from health economists on what threshold should be used when defining a successful replication. Currently the proposed definition fails to specify the percentage by which costs and outcomes can vary. In order to make this decision workable, and to conduct further replication studies using it, a decision on the variability is required. A Delphi consensus method could be used to reach a consensus. Similar methods have been used to determine the contents of HEAPs (Thorn et al., 2021).

This thesis has also suggested that there may be a need for health economist training in the use of open software or programming languages such as R, given the current reliance on Microsoft Excel. This could be incorporated in university modules for health economists in training. Alternatively, work is currently ongoing by researchers at the University of Bristol on the automated conversion of models developed in Microsoft Excel to R (Thom, 2023) which in the first instance involved writing popular Microsoft Excel commands into R. This may allow modellers to continue to develop models in Microsoft Excel and then convert them to an R script using the software, which could then be openly interrogated.

Research funders' willingness to pay for reproducibility must also be determined, given the likely increased costs of such an initiative. As part of this, it is important that research funders understand that more replicable research practices would likely lead to less research waste, if researchers build upon existing models rather than recreating a model within a specific condition. This could potentially also allow more time to be spent on validation exercises that would strengthen existing models. More replicable and transparent research could lead to better decisions on drug and technology funding by national bodies such as NICE, leading to more efficient utilisation of health care technologies by health systems.

The work conducted in this thesis could also be extended to other types of decision models. It could be used in other context settings such as the process of HTA, or extended to other areas of health economics such as in economic evaluations alongside trials.

The CHEERS checklist was updated in 2022 to include a reference to open-source modelling and to state whether model code was made available. A future research study could look to investigate whether the updated CHEERS checklist has impacted the availability of model code or if this aspect of the checklist is currently being overlooked, after allowing a sufficient amount of time for these new recommendations to establish.

There is also scope to consider the role of artificial intelligence or machine learning, given that this is now an emerging field in research, and may have applications to health economics and outcomes research (Padula et al., 2022). Artificial intelligence uses algorithms according to the following broad categories: natural language processing, data mining, and machine learning. The natural language processing component could be used in an initial screening process of manuscripts by journals in order to populate reporting checklists, or depending on the sophistication, developed in a way to ensure that model parameters are reported adequately. Whilst artificial intelligence may be a new avenue of research within health economics, how these practices are implemented will require careful reporting in order to prevent another aspect of reduced transparency and replication barriers.

5.7 Conclusion

This thesis has demonstrated the lack of current research exploring the replication of decision models within health economics. It has also shown the value of conducting replications, as both a tool to enhance the transparency of research, as a springboard for further model development and as a way to identify computation errors. For decision model transparency and replication initiatives to progress there is a requirement to move beyond the individual researcher motivated to develop replicable research and instead to build it into research infrastructure so that it is mandated. This would most likely need to come from the research funder and perhaps employing institutions, so as to introduce incentives for researchers to engage in such practices. Journals similarly have a role, but this would more likely be in raising the profile of replication studies, by publishing them and showing researchers that there is value in them being conducted.

References

Affleck, A. G., Bottomley, J. M., Auland, M., Jackson, P., & Ryttov, J. (2011). Cost effectiveness of the two-compound formulation calcipotriol and betamethasone dipropionate gel in the treatment of scalp psoriasis in Scotland. *Current medical research and opinion, 27*(1), 269-284.

Alouki, K., Delisle, H., Bermúdez-Tamayo, C., & Johri, M. (2016). Lifestyle interventions to prevent type 2 diabetes: a systematic review of economic evaluation studies. *Journal of diabetes research, 2016*, 2159890.

Altunkaya, J., Li, X. L., Feenstra, T., Keng, M. J., Lamotte, M., McEwan, P., . . . Clarke, P. (2023). *Quantifying cross-model variability in cost-effectiveness estimates: A mass replication exercise using eleven type 2 diabetes models*. University of Oxford, HESG Summer Conference, 21st-23rd June 2023.

Alva, M., Gray, A., Mihaylova, B., & Clarke, P. (2014). The effect of diabetes complications on health-related quality of life: the importance of longitudinal data to address patient heterogeneity. *Health economics, 23*(4), 487-500.

Alva, M., Gray, A., Mihaylova, B., Leal, J., & Holman, R. (2015). The impact of diabetes-related complications on healthcare costs: new results from the UKPDS (UKPDS 84). *Diabetic medicine, 32*(4), 459-466.

American Economic Association. (2020). Data and Code Availability Policy. Retrieved from https://www.aeaweb.org/journals/data/data-code-policy

Anokye, N. K., Trueman, P., Green, C., Pavey, T. G., Hillsdon, M., & Taylor, R. S. (2011). The cost-effectiveness of exercise referral schemes. *BMC Public Health, 11*(1), 1-11.

Arnold, R. J., & Ekins, S. (2010). Time for cooperation in health economics among the modelling community. *PharmacoEconomics, 28*(8), 609-613.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economics Review,, 53*(5), 941-973.

Bächle, C., Claessen, H., Andrich, S., Brüne, M., Dintsios, C., Slomiany, U., . . . Icks, A. (2016). Direct costs in impaired glucose regulation: results from the population-based Heinz Nixdorf Recall study. *BMJ Open Diabetes Research and Care, 4*(1), e000172.

Baker, M. (2016). Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help. *Nature, 533*(7604), 452-455.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical digital libraries, 3*(1), 1-8.

Balk, E. M., Earley, A., Raman, G., Avendano, E. A., Pittas, A. G., & Remington, P. L. (2015). Combined diet and physical activity promotion programs to prevent type 2 diabetes among persons at increased risk: a systematic review for the

Community Preventive Services Task Force. *Annals of internal medicine, 163*(6), 437-451.

Barendregt, J. J. (2014). The life table method of half cycle correction: getting it right. *Medical Decision Making, 34*(3), 283-285.

Barron, E., Clark, R., Hewings, R., Smith, J., & Valabhji, J. (2018). Progress of the Healthier You: NHS Diabetes Prevention Programme: referrals, uptake and participant characteristics. *Diabetic medicine, 35*(4), 513-518.

Barton, P. (2011). Development of the Birmingham Rheumatoid Arthritis Model: past, present and future plans. *Rheumatology, 50*(suppl_4), iv32-iv38.

Barton, P., Bryan, S., & Robinson, S. (2004). Modelling in the economic evaluation of health care: selecting the appropriate approach. *Journal of Health Services Research & Policy, 2*(9), 110-118.

Batty, A. J., Hansen, R. N., Bloudek, L. M., Varon, S. F., Hayward, E. J., Pennington, B. W., . . . Sullivan, S. D. (2013). The cost-effectiveness of onabotulinumtoxinA for the prophylaxis of headache in adults with chronic migraine in the UK. *Journal of medical economics, 16*(7), 877-887.

Beca, J., Husereau, D., Chan, K. K., Hawkins, N., & Hoch, J. S. (2017). Oncology Modeling for Fun and Profit! Key Steps for Busy Analysts in Health Technology Assessment. *PharmacoEconomics, 36*, 7-15.

Berkeley Initiative for Transparency in the Social Sciences. (2017). Retrieved from http://www.bitss.org/

Bermejo, I., Tappenden, P., & Youn, J.-H. (2017a). Replicating Health Economic Models: Firm Foundations or a House of Cards? *PharmacoEconomics, 35*(11), 1113-1121.

Bermejo, I., Tappenden, P., & Youn, J.-H. (2017b). Response to 'Comment on "Replicating health economic models: firm foundations or a house of cards?"'. *PharmacoEconomics, 35*(11), 1189-1190.

Bettis, R. A., Helfat, C. E., & Shaver, J. M. (2016). The necessity, logic, and forms of replication. *Strategic Management Journal, 37*(11), 2193-2203.

Bhopal, R. S., Douglas, A., Wallia, S., Forbes, J. F., Lean, M. E., Gill, J. M., . . . Wild, S. H. (2014). Effect of a lifestyle intervention on weight change in south Asian individuals in the UK at high risk of type 2 diabetes: a family-cluster randomised controlled trial. *The lancet Diabetes & endocrinology, 2*(3), 218-227.

Birkmeyer, J. D., & Liu, J. Y. (2003). Decision analysis models: opening the black box. *Surgery, 133*(1), 1-4.

Breeze, P., Thomas, C., Squires, H., Brennan, A., Greaves, C., Diggle, P. J., . . . Chilcott, J. (2017). The impact of Type 2 diabetes prevention programmes based on risk-identification and lifestyle intervention intensity strategies: a cost-effectiveness analysis. *Diabetic medicine, 34*(5), 632-640.

Brennan, A., Chick, S. E., & Davies, R. (2006). A taxonomy of model structures for economic evaluation of health technologies. *Health economics, 12*(15), 1295-1310.

Briggs, A., & Sculpher, M. (1998). An introduction to Markov modelling for economic evaluation. *PharmacoEconomics, 13*, 397-409.

Briggs, A., Sculpher, M., & Claxton. (2006). In *Decision modelling for health economic evaluation*: Oxford university press.

Brown, A. N., Cameron, D. B., & Wood, B. D. (2014). Quality evidence for policymaking: I'll believe it when I see the replication. *Journal of Development Effectiveness, 6*(3), 215-235.

Brunner, J., & Schimmack, U. (2016). *How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test statistics in original studies*. Retrieved from University of Toronto: https://www.utstat.toronto.edu/~brunner/papers/HowReplicable.pdf

Caro, J. J. (2005). Pharmacoeconomic analyses using discrete event simulation. *PharmacoEconomics, 4*(23), 323-332.

Chambers, M., Hutton, J., & Gladman, J. (1999). Cost-effectiveness analysis of antiplatelet therapy in the prevention of recurrent stroke in the UK: aspirin, dipyridamole and aspirin-dipyridamole. *PharmacoEconomics, 16*, 577-593.

Chambers, M. G., Koch, P., & Hutton, J. (2002). Development of a decision-analytic model of stroke care in the United States and Europe. *Value in Health, 5*(2), 82-97.

Chan, A.-W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., . . . Van Der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *The Lancet, 383*(9913), 257-266.

Chang, A. C. (2017a). Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say “Often Not”. *Critical Finance Review*.

Chang, A. C. (2017b). A replication recipe: list your ingredients before you start cooking. Economics Discussion Papers, No 2017-74. *Kiel Institute for the World Economy, 25*.

Chang, A. C. (2018). A replication recipe: List your ingredients before you start cooking. *Economics, 12*(1), 20180039.

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say'usually not'. *Finance and Economics Discussion Series 2015-083*. doi:http://dx.doi.org/10.17016/FEDS.2015.083

Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. *The Lancet, 389*(10085), 2239-2251.

Chilcott, J., Tappenden, P., Rawdin, A., Johnson, M., Kaltenthaler, E., Paisley, S., . . . Shippam, A. (2010). Avoiding and identifying errors in health technology assessment models: qualitative study and methodological. *Health Technology Assessment, 14*(25), 1-152.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., . . . Sculpher, M. (2015). Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technology Assessment (Winchester, England), 19*(14), 1-542.

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys, 31*(1), 326-342.

CODECHECK. (2023). Retrieved from https://codecheck.org.uk/

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., . . . Colombo, M. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology, 12*, 9-44.

Currie, C. J., Gale, E., & Poole, C. D. (2010). Estimation of primary care treatment costs and treatment efficacy for people with type 1 and type 2 diabetes in the United Kingdom from 1997 to 2007. *Diabetic medicine, 27*(8), 938-948.

Curtis, L. A., & Burns, A. (2020). *Unit Costs of Health & Social Care 2020*: PSSRU, University of Kent.

Dakin, H. A., Farmer, A., Gray, A. M., & Holman, R. R. (2020). Economic Evaluation of Factorial Trials: Cost-Utility Analysis of the Atorvastatin in Factorial With Omega EE90 Risk Reduction in Diabetes 2× 2× 2 Factorial Trial of Atorvastatin, Omega-3 Fish Oil, and Action Planning. *Value in Health, 23*(10), 1340-1348.

Davies, M., Gray, L., Ahrabian, D., Carey, M., Farooqi, A., Gray, A., . . . Leal, J. (2017). A community-based primary prevention programme for type 2 diabetes mellitus integrating identification and lifestyle intervention for prevention: a cluster randomised controlled trial. *NIHR Programme Grants for Applied Research, 5*(2), 1-324.

Davies, M. J., Heller, S., Skinner, T., Campbell, M., Carey, M., Cradock, S., . . . Eaton, S. (2008). Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomised controlled trial. *BMJ, 336*(7642), 491-495.

Dean, B. B., Siddique, R. M., Yamashita, B. D., Bhattacharjya, A. S., & Ofman, J. J. (2001). Cost-effectiveness of proton-pump inhibitors for maintenance therapy of erosive reflux esophagitis. *American journal of health-system pharmacy, 58*(14), 1338-1346.

DECODE Study Group European Diabetes Epidemiology Group. (2003). Is the current definition for diabetes relevant to mortality risk from all causes and cardiovascular and noncardiovascular diseases? *Diabetes care, 26*(3), 688-696.

Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine, 346*(6), 393-403.

Diabetes Prevention Program Research Group. (2009). 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *The Lancet, 374*(9702), 1677-1686.

Diabetes Prevention Program Research Group. (2015). Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. *The lancet Diabetes & endocrinology, 3*(11), 866-875.

Dritsaki, M., Gray, A., Petrou, S., Dutton, S., Lamb, S. E., & Thorn, J. C. (2018). Current UK practices on health economics analysis plans (HEAPs): are we using heaps of them? *PharmacoEconomics, 36*, 253-257.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*: Oxford university press.

Dunkley, A. J., Bodicoat, D. H., Greaves, C. J., Russell, C., Yates, T., Davies, M. J., & Khunti, K. (2014). Diabetes prevention in the real world: effectiveness of pragmatic lifestyle interventions for the prevention of type 2 diabetes and of the impact of adherence to guideline recommendations: a systematic review and meta-analysis. *Diabetes care, 37*(4), 922-933.

Dunlop, W. C., Mason, N., Kenworthy, J., & Akehurst, R. L. (2017). Benefits, Challenges and Potential Strategies of Open Source Health Economic Models. *PharmacoEconomics, 35*(1), 125-128.

Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics? *American Economic Review, 107*(5), 46-51.

Duvendack, M., Palmer-Jones, R. W., & Reed, W. R. (2015). Replications in Economics: A progress report. *Econ Journal Watch, 12*(2), 164-191.

Eddy, D. M., Hollingworth, W., Caro, J. J., Tsevat, J., McDonald, K. M., & Wong, J. B. (2012). Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Value in Health, 15*(6), 843-850.

Emerson, J., Bacon, R., Kent, A., Neumann, P. J., & Cohen, J. T. (2019). Publication of decision model source code: attitudes of health economics authors. *PharmacoEconomics, 37*, 1409-1410.

Equator Network. (2020). Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement. Retrieved from https://www.equator-network.org/reporting-guidelines/cheers/

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one, 4*(5), e5738.

Filipović-Pierucci, A., Zarca, K., & Durand-Zaleski, I. (2017). Markov models for health economic evaluations: the R package heemod. *arXiv preprint arXiv:1702.03252*.

Fletcher, B., Gulanick, M., & Lamendola, C. (2002). Risk factors for type 2 diabetes mellitus. *Journal of Cardiovascular Nursing, 16*(2), 17-23.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

Frempong, S. N., Shinkins, B., Howdon, D., Messenger, M., Neal, R. D., & Sagoo, G. S. (2021). Early economic evaluation of an intervention to improve uptake of the NHS England Diabetes Prevention Programme. *Expert Review of Pharmacoeconomics & Outcomes Research, 22*(3), 417-427.

Ganesalingam, J., Pizzo, E., Morris, S., Sunderland, T., Ames, D., & Lobotesis, K. (2015). Cost-utility analysis of mechanical thrombectomy using stent retrievers in acute ischemic stroke. *Stroke, 46*(9), 2591-2598.

García, F. M. (2014). Do Kenyan Teenagers Respond to HIV Risk Information? A Procedural Replication of Dupas (2011). Retrieved from https://www.semanticscholar.org/paper/Do-Kenyan-Teenagers-Respond-to-HIV-Risk-Information-Garc%C3%ADa/489d9f23b0fc1e4de7b52b43961da95985f641b1

Garside, R., Stein, K., Castelnuovo, E., Pitt, M., Ashcroft, D., Dimmock, P., & Payne, L. (2005). The effectiveness and cost-effectiveness of pimecrolimus and tacrolimus for atopic eczema: a systematic review and economic evaluation. *Health Technology Assessment, 9*(29), 1-230.

Gernsbacher, M. A. (2018). Rewarding research transparency. *Trends in Cognitive Sciences, 22*(11), 953-956.

Gillett, M., Brennan, A., Watson, P., Khunti, K., Davies, M., Mostafa, S., & Gray, L. J. (2015). The cost-effectiveness of testing strategies for type 2 diabetes: a modelling study. *Health Technology Assessment, 19*(33), 1-80.

Gillies, C. L., Lambert, P. C., Abrams, K. R., Sutton, A. J., Cooper, N. J., Hsu, R. T., . . . Khunti, K. (2008). Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis. *BMJ, 336*(7654), 1180-1185.

Goldacre, B. (2014). *Bad pharma: how drug companies mislead doctors and harm patients*: Macmillan.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine, 8*(341), 341ps312-341ps312.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., & Wordsworth, S. (2010). *Applied methods of cost-effectiveness analysis in healthcare* (Vol. 3): OUP Oxford.

Green, N., Lamrock, F., Naylor, N., Williams, J., & Briggs, A. (2023). Health Economic Evaluation Using Markov Models in R for Microsoft Excel Users: A Tutorial. *PharmacoEconomics, 41*(1), 5-19.

Gupta, A. K., & Nadarajah, S. (2004). *Handbook of beta distribution and its applications*: CRC press.

Hadaegh, F., Derakhshan, A., Zafari, N., Khalili, D., Mirbolouk, M., Saadat, N., & Azizi, F. (2017). Pre-diabetes tsunami: incidence rates and risk factors of pre-diabetes and its different phenotypes over 9 years of follow-up. *Diabetic medicine, 34*(1), 69-78.

Hamilton, M. P., Gao, C. X., Wiesner, G., Filia, K. M., Menssink, J. M., Plencnerova, P., . . . Karnon, J. (2023). A prototype software framework for transparent, reusable and updatable computational health economic models. *arXiv preprint arXiv:2310.14138*.

Hardwicke, T. E., Mathur, M. B., & Frank, M. C. (2017). Pre-Registration: Evaluation of the open data policy at Cognition. Retrieved from https://osf.io/q4qy3/

Harris, J. K., Johnson, K. J., Carothers, B. J., Combs, T. B., Luke, D. A., & Wang, X. (2018). Use of reproducible research practices in public health: a survey of public health analysts. *PloS one, 13*(9), e0202447.

Hart, R., Burns, D., Ramaekers, B., Ren, S., Gladwell, D., Sullivan, W., . . . Cain, T. (2020). R and Shiny for cost-effectiveness analyses: why and when? A hypothetical case study. *PharmacoEconomics, 38*, 765-776.

Hemmingsen, B., Gimenez-Perez, G., Mauricio, D., i Figuls, M. R., Metzendorf, M. I., & Richter, B. (2017). Diet, physical activity or both for prevention or delay of type 2 diabetes mellitus and its associated complications in people at increased risk of developing type 2 diabetes mellitus. *Cochrane database of systematic reviews*(12).

Herman, W. H., Hoerger, T. J., Brandle, M., Hicks, K., Sorensen, S., Zhang, P., . . . Ratner, R. E. (2005). The cost-effectiveness of lifestyle modification or metformin in preventing type 2 diabetes in adults with impaired glucose tolerance. *Annals of internal medicine, 142*(5), 323-332.

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics, 38*(2), 257-279.

Hex, N., Bartlett, C., Wright, D., Taylor, M., & Varley, D. (2012). Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabetic medicine, 29*(7), 855-862.

Höffler, J. H. (2017). Replication and economics journal policies. *American Economic Review, 107*(5), 52-55.

Höffler, J. H. (2018). ReplicationWiki - Improving Transparency in the Social Sciences. Retrieved from http://replication.uni-goettingen.de/wiki/index.php/Main_Page

Holman, N., Wild, S. H., Khunti, K., Knighton, P., O'Keefe, J., Bakhai, C., . . . Gregg, E. W. (2022). Incidence and characteristics of remission of type 2 diabetes in England: a cohort study using the National Diabetes Audit. *Diabetes care, 45*(5), 1151-1161.

Hostler, T. J. (2023). The Invisible Workload of Open Research. *Journal of Trial & Error*.

House of Commons: Science Innovation and Technology Committee. (2023). *Reproducibility and Research Integrity*. Retrieved from https://publications.parliament.uk/pa/cm5803/cmselect/cmsctech/101/report.html

Husereau, D., Drummond, M., Augustovski, F., de Bekker-Grob, E., Briggs, A., Carswell, C., . . . Loder, E. (2022). Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022 explanation and elaboration: a report of the ISPOR CHEERS II good practices task force. *Value in Health, 25*(1), 10-31.

Husereau, D., Drummond, M., Petrou, S., Carswell, C., Moher, D., Greenberg, D., . . . Loder, E. (2013). Consolidated health economic evaluation reporting standards (CHEERS) statement. *Cost Effectiveness and Resource Allocation, 11*(1), 6.

Imam, A. A. (2022). Remarkably reproducible psychological (memory) phenomena in the classroom: some evidence for generality from small-N research. *BMC psychology, 10*(1), 1-16.

Incerti, D., Thom, H., Baio, G., & Jansen, J. P. (2019). R You Still Using Excel? The Advantages of Modern Software Tools for Health Technology Assessment. *Value in Health, 22*(5), 575-579.

Institute for Replication. (2023). Past & Future Games. Retrieved from https://i4replication.org/games.html

International Initiative for Impact Evaluation. (2017). About 3ie. Retrieved from http://www.3ieimpact.org/en/about/

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS biology, 14*(1), e1002333.

Irvine, L., Barton, G. R., Gasper, A. V., Murray, N., Clark, A., Scarpello, T., & Sampson, M. (2011). Cost-effectiveness of a lifestyle intervention in preventing Type 2 diabetes. *International journal of technology assessment in health care, 27*(4), 275-282.

Jansen, D., & Warren, K. (2020). What is research methodology? Simple definition (With examples). *Grad Coach*.

Johansen, P., Håkan-Bloch, J., Liu, A. R., Bech, P. G., Persson, S., & Leiter, L. A. (2019). Cost effectiveness of once-weekly Semaglutide versus once-weekly Dulaglutide in the treatment of type 2 diabetes in Canada. *PharmacoEconomics-open, 3*(4), 537-550.

Jones, L. E., & Ziebarth, N. R. (2016). Successful scientific replication and extension of Levitt (2008): Child seats are still no safer than seat belts. *Journal of Applied Econometrics, 31*(5), 920-928.

Kanavos, P., van den Aardweg, S., & Schurer, W. (2012). Diabetes expenditure, burden of disease and management in 5 EU countries. *LSE health, London school of Economics, 113*, 1-123.

Karnon, J. (2003). Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. *Health economics, 10*(12), 837-848.

Karter, A. J., Nundy, S., Parker, M. M., Moffet, H. H., & Huang, E. S. (2014). Incidence of remission in adults with type 2 diabetes: the diabetes & aging study. *Diabetes care, 37*(12), 3188-3195.

Keng, M. J., Leal, J., Bowman, L., Armitage, J., Mihaylova, B., & Group, A. S. C. (2022). Decrements in health-related quality of life associated with adverse events in people with diabetes. *Diabetes, Obesity and Metabolism, 24*(3), 530-538.

Khan, T., Tsipas, S., & Wozniak, G. (2017). Medical care expenditures for individuals with prediabetes: the potential cost savings in reducing the risk of developing diabetes. *Population health management, 20*(5), 389-396.

Khan, Z. M., & Miller, D. W. (1999). Modeling economic evaluations of pharmaceuticals: manipulation or valuable tool? *Clinical therapeutics, 21*(5), 896-908.

Kim, D. D., Do, L. A., Synnott, P. G., Lavelle, T. A., Prosser, L. A., Wong, J. B., & Neumann, P. J. (2023). Developing criteria for health economic quality evaluation tool. *Value in Health, 26*(8), 1225-1234.

Kim, D. D., & Neumann, P. J. (2019). Comparative Modeling to Inform Health Policy Decisions: A Step Forward. *Annals of internal medicine, 171*(11), 851-852.

Knapp, M. R. (1984). *The economics of social care*: Macmillan International Higher Education.

Krentz, A. J., Patel, M. B., & Bailey, C. J. (2008). New drugs for type 2 diabetes mellitus: what is their place in therapy? *Drugs, 68*, 2131-2162.

Leal, J., Morrow, L. M., Khurshid, W., Pagano, E., & Feenstra, T. (2019). Decision models of prediabetes populations: a systematic review. *Diabetes, Obesity Metabolism, 21*(7), 1558-1569.

Leal, J., Reed, S. D., Patel, R., Rivero-Arias, O., Li, Y., Schulman, K. A., . . . Gray, A. M. (2020). Benchmarking the Cost-Effectiveness of Interventions Delaying Diabetes: A Simulation Study Based on NAVIGATOR Data. *Diabetes care, 43*(10), 2485-2492.

Li, R., Zhang, P., Barker, L. E., Chowdhury, F. M., & Zhang, X. (2010). Cost-effectiveness of interventions to prevent and control diabetes mellitus: a systematic review. *Diabetes care, 33*(8), 1872-1894.

Low, V., & Macaulay, R. (2022). Accounting for inflation within NICE cost-effectiveness thresholds. *Expert Review of Pharmacoeconomics & Outcomes Research, 22*(1), 131-137.

Lusoli, W. (2020). *Reproducibility of Scientific Results in the EU: Scoping report*: Publications Office of the European Union.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542.

Martin, G., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in psychology, 8*, 1-6.

Martin, S., Claxton, K., Lomas, J., & Longo, F. (2023). The impact of different types of NHS expenditure on health: Marginal cost per QALY estimates for England for 2016/17. *Health Policy, 132*, 104800.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist, 70*(6), 487.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB Archive. *Journal of Money, Credit and Banking, 38*(4), 1093-1107.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue canadienne d'économique, 41*(4), 1406-1420.

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *Elife, 8*, e47338.

McManus, E. (2023). Model code. Retrieved from https://github.com/e-mcmanus/NHS-DPP

McManus, E., Meacock, R., Parkinson, B., & Sutton, M. (2023). Working paper: Evaluating the short-term costs and benefits of a nationwide diabetes prevention programme: retrospective observational study

McManus, E., & Sach, T. (2017). Comment on “replicating health economic models: firm foundations or a house of cards?”. *PharmacoEconomics, 35*(11), 1187-1188.

McManus, E., Sach, T., & Levell, N. (2017). The Use of Decision–Analytic Models in Atopic Eczema: A Systematic Review and Critical Appraisal. *PharmacoEconomics, 36*, 51-66.

McManus, E., Sach, T., & Levell, N. (2019). An introduction to the methods of decision-analytic modelling used in economic evaluations for Dermatologists. *Journal of the European Academy of Dermatology and Venereology, 33*(10), 1829-1836.

McManus, E., Turner, D., Gray, E., Khawar, H., Okoli, T., & Sach, T. (2019). The Barriers and Facilitators to Model Replication Within Health Economics. *Value in Health, 22*(9), 1018-1025. doi:10.1016/j.jval.2019.04.1928

McManus, E., Turner, D., & Sach, T. (2019). Can you repeat that? Exploring the definition of a successful model replication in health economics. *PharmacoEconomics, 37*(11), 1371-1381.

Medicines and Healthcare products Regulatory Agency. (2018). *Good Clinical Practice Guide* (9th ed.). London: The Stationery Office.

Meigs, J. B., Muller, D. C., Nathan, D. M., Blake, D. R., & Andres, R. (2003). The natural history of progression from normal glucose tolerance to type 2 diabetes in the Baltimore Longitudinal Study of Aging. *Diabetes, 52*(6), 1475-1484.

Miners, A., Harris, J., Felix, L., Murray, E., Michie, S., & Edwards, P. (2012). An economic evaluation of adaptive e-learning devices to promote weight loss via dietary change for people with obesity. *BMC Health Services Research, 12*, 1-9.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of clinical epidemiology, 63*(7), 737-745.

Morgan, C. L., Jenkins-Jones, S., Currie, C., Berni, E., Holden, S., & Qiao, Q. (2016). *Health service utilisation and costs of treatment with either exenatide twice daily or basal insulin for patients with type 2 diabetes: a retrospective UK study.* Paper presented at the Diabetologia.

Morgan, C. L., Peters, J., Dixon, S., & Currie, C. J. (2010). Estimated costs of acute hospital care for people with diabetes in the United Kingdom: a routine record linkage study in a large region. *Diabetic medicine, 27*(9), 1066-1073.

Morris, S. (1997). A comparison of economic modelling and clinical trials in the economic evaluation of cholesterol-modifying pharmacotherapy. *Health economics, 6*(6), 589-601.

Morris, S., Devlin, N., & Parkin, D. (2007). *Economic analysis in health care*: John Wiley & Sons.

Mount Hood Diabetes Challenge. (2016). The Mount Hood 2016 Challenge, Switzerland. Retrieved from https://docs.wixstatic.com/ugd/4e5824_36cb1fd0aca94f1980d8a2228cf7e6e8.pdf

Moylan, E. C., & Kowalczuk, M. K. (2016). Why articles are retracted: a retrospective cross-sectional study of retraction notices at BioMed Central. *BMJ open, 6*(11), e012047.

Moynihan, R., Macdonald, H., Heneghan, C., Bero, L., & Godlee, F. (2019). Commercial interests, transparency, and independence: a call for submissions. In (Vol. 365): British Medical Journal Publishing Group.

Mwachofi, A., & Al-Assaf, A. F. (2011). Health care market deviations from the ideal market. *Sultan Qaboos university medical journal, 11*(3), 328.

NatCen Social Research, U. C. L., Department of Epidemiology and Public Health,. (2022). Health Survey for England, 2018. [data collection]. *UK Data Service. SN: 8649*. doi:http://doi.org/10.5255/UKDA-SN-8649-2

National Institute for Health and Care Excellence. (2012, 15 September 2017). Type 2 diabetes: prevention in people at high risk. Retrieved from https://www.nice.org.uk/guidance/ph38

National Institute for Health and Care Excellence. (2013). Guide to the methods of technology appraisal. Retrieved from https://www.nice.org.uk/process/pmg9/chapter/the-reference-case

National Institute for Health and Care Excellence. (2015). Technology appraisal guidance [TA82]. Appendix B: Source of evidence considered by the Committee Retrieved from https://www.nice.org.uk/guidance/ta82/chapter/Appendix-B-Sources-of-evidence-considered-by-the-Committee

National Institute for Health and Care Excellence. (2018). Guide to the processes of technology appraisal. Retrieved from https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/technology-appraisal-processes-guide-apr-2018.pdf

National Institute for Health and Care Excellence. (2022). NICE health technology evaluations: the manual. Retrieved from https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation

National Institute for Health and Care Research. (2017). Evaluating the NHS Diabetes Prevention Programme (NHS DPP): the DIPLOMA research programme (Diabetes Prevention Long term Multimethod Assessment). Retrieved from https://fundingawards.nihr.ac.uk/award/16/48/07

NAVIGATOR Study Group. (2010). Effect of nateglinide on the incidence of diabetes and cardiovascular events. *New England Journal of Medicine, 362*(16), 1463-1476.

Neumann, A., Schoffer, O., Norström, F., Norberg, M., Klug, S. J., & Lindholm, L. (2014). Health-related quality of life for pre-diabetic states and type 2 diabetes mellitus: a cross-sectional study in Västerbotten Sweden. *Health and quality of life outcomes, 12*(1), 1-10.

NHS Business Services Authority. (2021). Prescription Cost Analysis – England 2020/21. Retrieved from https://www.nhsbsa.nhs.uk/statistical-collections/prescription-cost-analysis-england/prescription-cost-analysis-england-202021

NHS Digital. (2021). National Diabetes Audit, 2019-20, Report 1: Care Processes and Treatment Targets. Retrieved from https://files.digital.nhs.uk/42/B253B1/National%20Diabetes%20Audit%202019-20%20Full%20Report%201.pdf

NHS Digital. (2023). National Diabetes Audit Programme. Retrieved from https://digital.nhs.uk/data-and-information/clinical-audits-and-registries/national-diabetes-audit

NHS England. (2016). NHS England Impact Analysis of implementing NHS Diabetes Prevention Programme, 2016 to 2021. Retrieved from https://www.england.nhs.uk/wp-content/uploads/2016/08/impact-assessment-ndpp.pdf

NHS England. (2022). 2020/21 National Cost Collection Data Publication. Retrieved from https://www.england.nhs.uk/publication/2020-21-national-cost-collection-data-publication/

Nichols, G. A., Arondekar, B., & Herman, W. H. (2008). Medical care costs one year after identification of hyperglycemia below the threshold for diabetes. *Medical care, 46*(3), 287-292.

Office for National Statistics. (2021). National life tables – life expectancy in the UK: 2018 to 2020. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2018to2020

Office for National Statistics. (2022). CPI Annual Rate 00: All items 2015=100. Retrieved from https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7g7/mm23

Olsen, J. (2022). What makes the market for healthcare different. In *Principles in Health Economics and Policy: Oxford University Press*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Otten, T. M., Grimm, S. E., Ramaekers, B., & Joore, M. A. (2023). Comprehensive Review of Methods to Assess Uncertainty in Health Economic Evaluations. *PharmacoEconomics, 41*(6), 619-632.

Padula, W. V., Kreif, N., Vanness, D. J., Adamson, B., Rueda, J.-D., Felizzi, F., . . . Crown, W. (2022). Machine learning methods in health economics and outcomes research—the PALISADE checklist: a good practices report of an ISPOR task force. *Value in health, 25*(7), 1063-1080.

Padula, W. V., McQueen, R. B., & Pronovost, P. J. (2017). Can Economic Model Transparency Improve Provider Interpretation of Cost-effectiveness Analysis? Evaluating Tradeoffs Presented by the Second Panel on Cost-effectiveness in Health and Medicine. *Medical Care, 55*(11), 909-911.

Palmer, A. J., Roze, S., Valentine, W. J., Spinas, G. A., Shaw, J. E., & Zimmet, P. Z. (2004). Intensive lifestyle changes or metformin in patients with impaired glucose tolerance: modeling the long-term health economic implications of the diabetes prevention program in Australia, France, Germany, Switzerland, and the United Kingdom. *Clinical therapeutics, 26*(2), 304-321.

Palmer, A. J., Si, L., Tew, M., Hua, X., Willis, M. S., Asseburg, C., . . . Foos, V. (2018). Computer modeling of diabetes and its transparency: a report on the eighth mount hood challenge. *Value in Health, 21*(6), 724-731.

Palmer, A. J., & Tucker, D. (2012). Cost and clinical implications of diabetes prevention in an Australian setting: a long-term modeling analysis. *Primary care diabetes, 6*(2), 109-121.

Palmer, S., & Torgerson, D. J. (1999). Definitions of efficiency. *BMJ, 318*(7191), 1136.

Pashler, H., Spellman, B., Kang, S., & Holcombe, A. PsychFileDrawer: archive of replication attempts in experimental Psychology. Retrieved from http://www.psychfiledrawer.org/faq.php

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*(4), 539-544.

Paul, J. E., & Trueman, P. (2001). 'Fourth hurdle reviews', NICE, and database applications. *Pharmacoepidemiology and drug safety, 5*(10), 429-438.

Peng, R. D. (2009). Reproducible research and biostatistics. *Biostatistics, 10*(3), 405-408.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226-1227.

Peng, R. D., Dominici, F., & Zeger, S. L. (2006). Reproducible epidemiologic research. *American journal of epidemiology, 163*(9), 783-789.

Penn, L., White, M., Oldroyd, J., Walker, M., Alberti, K. G. M., & Mathers, J. C. (2009). Prevention of type 2 diabetes in adults with impaired glucose tolerance: the European Diabetes Prevention RCT in Newcastle upon Tyne, UK. *BMC Public Health, 9*, 1-14.

Pesaran, H. (2003). Introducing a replication section. *Journal of Applied Econometrics, 18*(1), 111-111.

Philips, Z., Ginnelly, L., Sculpher, M., Claxton, K., Golder, S., Riemsma, R., . . . Glanville, J. (2004). Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technology Assessment, 8*(36), 1-179.

Pouwels, X. G., Sampson, C. J., Arnold, R. J., Janodia, M. D., Henderson, R., Lamotte, M., . . . Udupa, N. (2022). Opportunities and Barriers to the Development and Use of Open Source Health Economic Models: A Survey. *Value in Health, 25*(4), 473-479.

Prager, E. M., Chambers, K. E., Plotkin, J. L., McArthur, D. L., Bandrowski, A. E., Bansal, N., . . . Graf, C. (2019). Improving transparency and scientific rigor in academic publishing. In (Vol. 97, pp. 377-390): Wiley Online Library.

R Core Team. (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

R for Health Technology Assessment. (2023). Mission. Retrieved from https://r-hta.org/

Raftery, J., Williams, H., Clarke, A., Thornton, J., Norrie, J., Snooks, H., & Stein, K. (2020). ‘Not clinically effective but cost-effective’-paradoxical conclusions in randomised controlled trials with ‘doubly null’results: a cross-sectional study. *BMJ Open, 10*(1), e029596.

Ramachandran, A., Snehalatha, C., Yamuna, A., Mary, S., & Ping, Z. (2007). Cost-effectiveness of the interventions in the primary prevention of diabetes among Asian Indians: within-trial results of the Indian Diabetes Prevention Programme (IDPP). *Diabetes care, 30*(10), 2548-2552.

Ravindrarajah, R., Reeves, D., Howarth, E., Meacock, R., Soiland-Reyes, C., Cotterill, S., . . . Kontopantelis, E. (2020). The epidemiology and determinants of non-diabetic hyperglycaemia and its conversion to type 2 diabetes mellitus, 2000-2015: cohort population study using UK electronic health records. *BMJ Open, 10*(9), e040201.

Ravindrarajah, R., Sutton, M., Reeves, D., Cotterill, S., Mcmanus, E., Meacock, R., . . . Bower, P. (2023). Referral to the NHS Diabetes Prevention Programme and conversion from nondiabetic hyperglycaemia to type 2 diabetes mellitus in England: A matched cohort analysis. *PLoS medicine, 20*(2), e1004177.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's ‘Retroactive Facilitation of Recall’ Effect. *PloS one, 7*(3), e33423.

Roberts, S., Barry, E., Craig, D., Airoldi, M., Bevan, G., & Greenhalgh, T. (2017). Preventing type 2 diabetes: systematic review of studies of cost-effectiveness of lifestyle programmes and metformin, with and without screening, for pre-diabetes. *BMJ Open, 7*(11), e017184.

Roberts, S., Craig, D., Adler, A., McPherson, K., & Greenhalgh, T. (2018). Economic evaluation of type 2 diabetes prevention programmes: Markov model of low-and high-intensity lifestyle programmes and metformin in participants with different categories of intermediate hyperglycaemia. *Bmc Medicine, 16*(1), 1-12.

Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, F. C., . . . Davison, A. P. (2017). Sustainable computational science: the ReScience initiative. *PeerJ Computer Science, 3*, e142.

Ryan-Wenger, N. A. (2017). The benefits of replication research. *Journal for Specialists in Pediatric Nursing, 22*(4), e12198.

Sampson, C. J. (2012). Call for a model registry. Retrieved from https://aheblog.com/2012/10/19/call-for-a-model-registry/

Sampson, C. J., Arnold, R., Bryan, S., Clarke, P., Ekins, S., Hatswell, A., . . . Sadatsafavi, M. (2019). Transparency in decision modelling: what, why, who and how? *PharmacoEconomics, 37*(11), 1355-1369.

Sampson, C. J., & Wrightson, T. (2017). Model Registration: A Call to Action. *PharmacoEconomics - Open, 1*(2), 73-77. doi:10.1007/s41669-017-0019-2

Santana, I. R., Mason, A., Gutacker, N., Kasteridis, P., Santos, R., & Rice, N. (2023). Need, demand, supply in health care: working definitions, and their implications for defining access. *Health Economics, Policy and Law, 18*(1), 1-13.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90.

Schwander, B., Kaier, K., Hiligsmann, M., Evers, S., & Nuijten, M. (2022). Does the Structure Matter? An External Validation and Health Economic Results Comparison of Event Simulation Approaches in Severe Obesity. *PharmacoEconomics, 40*(9), 901-915.

Schwander, B., Nuijten, M., Evers, S., & Hiligsmann, M. (2021). Replication of Published Health Economic Obesity Models: Assessment of Facilitators, Hurdles and Reproduction Success. *Pharmacoeconomics*, 1-14.

Science Innovation and Technology Committee. (2023). *Reproducibility and Research Integrity*. Retrieved from UK Parliament: https://committees.parliament.uk/publications/39343/documents/194466/default/

Seuring, T., Archangelidi, O., & Suhrcke, M. (2015). The economic costs of type 2 diabetes: a global systematic review. *PharmacoEconomics, 33*, 811-831.

Shao, H., Lin, J., Zhuo, X., Rolka, D. B., Gregg, E. W., & Zhang, P. (2019). Influence of diabetes complications on HbA1c treatment goals among older US adults: a cost-effectiveness analysis. *Diabetes care, 42*(11), 2136-2142.

Sharma, M., Nazareth, I., & Petersen, I. (2016). Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. *BMJ Open, 6*(1), e010210.

Shemilt, I., Thomas, J., & Morciano, M. (2010). A web-based tool for adjusting costs to a specific target currency and price year. *Evidence & Policy: A Journal of Research, Debate and Practice, 6*(1), 51-59.

Smith, R., & Schneider, P. (2020). Making health economic models Shiny: A tutorial. *Wellcome open research, 5*(69), 1-26.

Smolen, L. J., Klein, T. M., & Kelton, K. (2015). Replication Of A Published Markov Chronic Migraine Cost-Effectiveness Analysis Model For Purposes Of Early Phase Adaptation And Expansion. *Value in Health, 18*(3), A19.

Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: a practical guide. *Medical Decision Making, 13*(4), 322-338.

Stewart, A. J., Farran, E. K., Grange, J. A., Macleod, M., Munafò, M., Newton, P., & Shanks, D. R. (2021). Improving research quality: the view from the UK Reproducibility Network institutional leads for research improvement. *BMC research notes, 14*(1), 1-4.

Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). Prediabetes: a high-risk state for diabetes development. *The Lancet, 379*(9833), 2279-2290.

The Replication Network. (2017). Furthering the Practice of Replication in Economics. Retrieved from https://replicationnetwork.com/

The REWARD Alliance. (2016). The REWARD statement. Retrieved from http://researchwaste.net/reward-statement/

Thom, H. (2022). Deterministic and Probabilistic Analysis of a Simple Markov Model: How Different Could They Be? *Applied health economics and health policy, 20*(3), 447-449.

Thom, H. (2023). Value of information analysis for health care trial design, and its computational challenges. Retrieved from https://www.imsi.institute/videos/value-of-information-analysis-for-health-care-trial-design-and-its-computational-challenges/

Thorn, J. C., Davies, C. F., Brookes, S. T., Noble, S. M., Dritsaki, M., Gray, E., . . . Ridyard, C. (2021). Content of Health Economics Analysis Plans (HEAPs) for trial-based economic evaluations: expert Delphi consensus survey. *Value in Health, 24*(4), 539-547.

TreeAge Pro. (2021). TreeAge Software, Williamstown, MA; software available at http://www.treeage.com.

Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data, 9*(60), 1-16.

Vahdat, V., Griffin, J. A., Stahl, J. E., & Yang, F. C. (2018). Analysis of the effects of EHR implementation on timeliness of care in a dermatology clinic: a simulation study. *Journal of the American Medical Informatics Association, 25*(7), 827-832.

Van Dalen, H. P., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology, 63*(7), 1282-1293.

Van Hout, B., Janssen, M., Feng, Y.-S., Kohlmann, T., Busschbach, J., Golicki, D., . . . Pickard, A. S. (2012). Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health, 15*(5), 708-715.

Vitae. (2020). *Research integrity: a landscape study*. Retrieved from https://www.vitae.ac.uk/vitae-publications/research-integrity-a-landscape-study

von Hippel, P. T. (2022). Is psychological science self-correcting? Citations before and after successful and failed replications. *Perspectives on Psychological Science, 17*(6), 1556-1565.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., . . . Chen, A. Z. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388*(10053), 1545-1602.

Watts, R. D., & Li, I. W. (2019). Use of checklists in reviews of health economic evaluations, 2010 to 2018. *Value in Health, 22*(3), 377-382.

Weinstein, M. C., O'brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C., & Luce, B. R. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value in Health, 6*(1), 9-17.

Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: the QALY and utilities. *British medical bulletin, 96*(1), 5-21.

Wilson, E. C. (2021). Methodological note: reporting deterministic versus probabilistic results of markov, partitioned survival and other non-linear models. *Applied health economics and health policy, 19*, 789-795.

Wohlin, C. (2014). *Guidelines for snowballing in systematic literature studies and a replication in software engineering.* Paper presented at the Proceedings of the 18th international conference on evaluation and assessment in software engineering.

Wood, B. D., Müller, R., & Brown, A. N. (2018). Push button replication: Is impact evaluation evidence for international development verifiable? *PloS one, 13*(12), e0209416.

World Health Organization. (1948). Constitution of the World Health Organization, as adopted by the International Health Conference, New York, 19–22 June 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948. *WHO, Geneva, Switzerland*.

World Health Organization. (2023). Health Equity. Retrieved from https://www.who.int/health-topics/health-equity#tab=tab_1

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*: CRC Press.

Xin, Y., Gray, E., Robles-Zurita, J. A., Haghpanahan, H., Heggie, R., Kohli-Lynch, C., . . . Lewsey, J. (2021). From spreadsheets to script: experiences from converting a Scottish cardiovascular disease policy model into R. *Applied health economics and health policy, 20*, 149-158.

Zawadzki, N. K., & Hay, J. W. (2020). Characterizing the Validity and Real-World Utility of Health Technology Assessments in Healthcare: Future Directions: Comment on" Problems and Promises of Health Technologies: The Role of Early Health Economic Modelling". *International Journal of Health Policy and Management, 9*(8), 352-355.

Zhou, X., Siegel, K. R., Ng, B. P., Jawanda, S., Proia, K. K., Zhang, X., . . . Zhang, P. (2020). Cost-effectiveness of Diabetes Prevention Interventions Targeting High-risk Individuals and Whole Populations: A Systematic Review. *Diabetes care, 43*(7), 1593-1616.

Zorginstituut Nederland. (2022). Richtlijn voor kosteneffectiviteitsmodellen in R. Retrieved from https://www.zorginstituutnederland.nl/publicaties/publicatie/2022/12/15/richtlijn-kosteneffectiviteitsmodellen-in-r

Chapter 3 Appendix

Table A3.1: Responses to the Philips checklist criteria for each of the models replicated.

	Case Study				
Checklist Item	**1**	**2**	**3**	**4**	**5**
Is there a clear statement of the decision problem?	y	y	y	y	y
Is the objective of the evaluation and model specified and consistent with the stated decision problem?	y	y	y	y	y
Is the primary decision maker specified?	y	y	y	y	y
Is the perspective of the model stated clearly?	y	y	y	y	y
Are the model inputs consistent with the stated perspective?	y	y	y	y	y
***Has the scope of the model been stated and justified?**	**y**	**y**	**y**	**y**	**y**
Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	y	y	y	y	y
***Has the evidence regarding the model structure been described?**	**y**	**n**	**y**	**y**	**n**
Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	y	y	y	y	y
Have any competing theories regarding model structure been considered?	y	n	n	n	n
***Are the sources of data used to develop the structure of the model specified?**	**y**	**n**	**y**	**n**	**n**
Are the causal relationships described by the model structure justified appropriately?	y	y	y	y	y

Table A3.1: Responses to the Philips checklist criteria for each of the models replicated. *(Continued)*

	Case Study				
Checklist Item	**1**	**2**	**3**	**4**	**5**
***Are the structural assumptions transparent and justified?**	**y**	**y**	**y**	**y**	**y**
Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	y	y	y	y	y
***Is there a clear definition of the options under evaluation?**	**y**	**y**	**y**	**y**	**y**
Have all feasible and practical options been evaluated?	y	y	y	y	y
Is there justification for the exclusion of feasible options?	y	n/a	n/a	n/a	n/a
Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	y	n	y	y	y
Is the time horizon of the model sufficient to reflect all important differences between options?	y	n	y	y	y
***Is the time horizon of the model, and the duration of treatment and treatment effect described and justified?**	**y**	**y**	**y**	**y**	**y**
Has a lifetime horizon been used? If not, has a shorter time horizon been justified?	n	n	n	y	y
Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	y	y	y	y	y
***Is the cycle length defined and justified in terms of the natural history of disease?**	**p**	**n/a**	**y**	**p**	**p**
Are the data identification methods transparent and appropriate given the objectives of the model?	y	y	y	y	y

Table A3.1: Responses to the Philips checklist criteria for each of the models replicated. *(Continued)*

	Case Study				
Checklist Item	**1**	**2**	**3**	**4**	**5**
Where choices have been made between data sources, are these justified appropriately?	y	n/a	y	n/a	n/a
Has particular attention been paid to identifying data for the important parameters in the model?	y	y	y	y	y
Has the process of selecting key parameters been justified and systematic methods used to identify the most appropriate data?	y	y	y	y	y
Has the quality of the data been assessed appropriately?	y	y	y	y	n
Where expert opinion has been used, are the methods described and justified?	p	p	y	y	n/a
Are the pre-model data analysis methodology based on justifiable statistical and epidemiological techniques?	y	y	y	y	y
***Is the choice of baseline data described and justified?**	**y**	**y**	**y**	**y**	**y**
***Are transition probabilities calculated appropriately?**	**y**	**y**	**y**	**y**	**y**
Has a half cycle correction been applied to both cost and outcome?	n	n/a	n	n	n
If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	y	y	y	y	y
***Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified? Have alternative assumptions been explored through sensitivity analysis?**	**y**	**y**	**y**	**y**	**y**
***Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have**	**y**	**y**	**n**	**y**	**y**

Table A3.1: Responses to the Philips checklist criteria for each of the models replicated. *(Continued)*

	Case Study				
Checklist Item	**1**	**2**	**3**	**4**	**5**
alternative assumptions been explored through sensitivity analysis?	🟩	🟩	🟥	🟩	🟩
Are the utilities incorporated into the model appropriate?	y	n/a	y	n/a	y
***Is the source for the utility weights referenced?**	**y**	**n/a**	**y**	**n/a**	**y**
Are the methods of derivation for the utility weights justified?	y	n/a	y	n/a	y
***Have all data incorporated into the model been described and referenced in sufficient detail?**	**y**	**y**	**y**	**y**	**y**
Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	n/a	n/a	n/a	n/a	n/a
***Is the process of data incorporation transparent?**	**y**	**y**	**y**	**y**	**y**
***If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?**	**n/a**	**n/a**	**n/a**	**n/a**	**p**
Have the four principal types of uncertainty been addressed?	p	p	p	p	p
If not, has the omission of particular forms of uncertainty been justified?	n	n	n	n	n
Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	n	n	n	y	n
Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	n	y	n	n	y

Table A3.1: Responses to the Philips checklist criteria for each of the models replicated. *(Continued)*

	Case Study				
Checklist Item	**1**	**2**	**3**	**4**	**5**
Has heterogeneity been dealt with by running the model separately for different sub-groups?	y	n	n	n	n
Are the methods of assessment of parameter uncertainty appropriate?	y	y	y	y	y
Has probabilistic sensitivity analysis been done, if not has this been justified?	y	n	n	n	y
***If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated and justified?**	**y**	**y**	**p**	**y**	**p**
Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	n	n	n	n	n
Are the conclusions valid given the data presented?	y	y	y	y	y
Are any counterintuitive results from the model explained and justified?	n/a	n/a	n/a	n/a	n/a
If the model has been calibrated against independent data, have any differences been explained and justified?	y	n/a	n/a	n/a	n/a
Have the results of the model been compared with those of previous models and any differences in results explained?	y	n	y	y	y
TOTALS:					
Yes	83%	69%	76%	79%	74%
No	11%	27%	20%	17%	18%
Partial	6%	4%	4%	4%	8%

Table A3.1: Responses to the Philips checklist criteria for each of the models replicated. *(Continued)*

Abbreviations: Y: Yes; N: No; P: Partial; N/A: Not applicable. *The asterisked criteria were those that were thought to have the greatest potential to influence the replicability of a modelling study, due to the fact that they directly related to the reporting of items needed for replication.

Chapter 4 Appendix

Sourcing Model Parameters

Transition Probabilities

To source transition probabilities the existing peer reviewed literature was searched, along with consulting the sources for parameters of existing modelling studies. Below, it is described how and why each of the transition probabilities were selected, along with other potential sources that could have been used.

Where applicable, transition probabilities were calculated from rates using the following formula:

$$Annual\,probability = 1 - \, e^{-\left(\frac{events}{time\ period}\right)*1}$$

Normal glucose tolerance to Non-diabetic Hyperglycaemia

Two studies were identified that could be used to determine the rate of progression from normal glucose tolerance and non-diabetic hyperglycaemia: Meigs et al. (2003) and Hadaegh et al. (2017). The study by Hadaegh et al. was more recent and had a larger sample size (n=5,879), with a 9-year follow-up, however was based in Iran and as such the estimates may not be generalisable to the English population. This paper also presented the rates in terms of different age categories: 20-39, 40-59 and 60 years and over. The study by Meigs et al. was less recent with a smaller sample size (n=488) but had a similar follow-up period of 10 years and had a study population from the United States (largely from the Baltimore, Maryland, and Washington D.C. areas), reporting rates in terms of individuals above and below 65 years of age. This paper found that progression from normal glucose tolerance to non-diabetic hyperglycaemia (measured by: 2hPG>=7.8 mmol/l) was 12.1 per 100 person years amongst individuals aged 65 and over, which was equivalent to a transition probability of 0.114. Hadaegh et al. found a higher event rate of 77.1 events per 1,000 person years (amongst individuals that were aged 60 years and over), equivalent to a transition probability of: 0.0742.

The estimates presented by Hadaegh et al. were chosen to be used, due to this study being the more recent, having a much larger sample size and the rates being presented by more age categories.

Non-diabetic Hyperglycaemia to Normal glucose tolerance

A systematic review was identified that looked at studies reporting on regression to normal glucose tolerance from non-diabetic hyperglycaemia (Balk et al., 2015). This review identified six studies, of which four reported rates for 3 years of follow-up. The data presented in this systematic review was used to determine the transition probability to be used in the model, shown in Table A4.1.

Table A4.1: Study rates of individuals reverting to normal glucose tolerance from non-diabetic hyperglycaemia within a three year time horizon. Sourced from Balk et al. (2015)

	Usual Care		
Study	**Events**	**People**	**Person Life Years**
Bhopal et al. (2014)	32	82	246
Penn et al. (2009)	11	51	153
Ramachandran et al. (2007)	32	133	399
Diabetes Prevention Program Research Group (2002)	260	1,082	3,246
Total:	335		4,044
Rate:	0.0828 per 1 person year		
Transition probability	0.0795		
Distribution	Beta(α=335, β=3709)		

Non-diabetic Hyperglycaemia to Type 2 diabetes

Linked data from the National Diabetes Audits for individuals with non-diabetic hyperglycaemia and type 2 diabetes was used, along with the NHS DPP provider dataset, which included all referrals made to the programme. The National Diabetes Audit covered 99.3% of practices in England in 2019/20 (NHS Digital, 2021). As this dataset is nationally representative, this data was used to derive the transition probabilities from non-diabetic hyperglycaemia to type 2 diabetes, rather than trying to find a published estimate from the literature. Three extracts of the National Diabetes Audit were used, covering the periods: 1st April 2017 – 31st March 2018, 1st January 2018 – 31st March 2019 and 1st January 2019 – 31st March 2020. Individuals that were offered or referred to the programme were excluded, to determine the baseline probability of developing type 2 diabetes from a non-diabetic hyperglycaemic state.

There were a total of 1,069,790 subjects diagnosed with a Read Code of non-diabetic hyperglycaemia in their electronic health care record from 1st January 2016, this amounted to 2,071,280 person years from the time of the non-diabetic hyperglycaemia diagnosis to the date of type 2 diabetes diagnosis or 31st March 2020 (whichever came sooner). A total of 52,330 cases of type 2 diabetes were observed during this period, which was equivalent to: 25.27 cases per 1,000 person years. These calculations were repeated for the different age groups modelled.

Here numbers are rounded according to the output rules of the National Diabetes Audit.

Type 2 diabetes to Non-diabetic hyperglycaemia

Two studies were identified that looked at regression from type 2 diabetes to non-diabetic hyperglycaemia. The first, a study by Holman et al. (2022) used data from the National Diabetes Audit and found that amongst 2,297,700 people with type 2 diabetes, the overall incidence of remission per 1,000 person-years was 9.7 (95% CI 9.6–9.8). Whilst this study is representative of the English population being modelling, it unfortunately did not distinguish between remission state (for example normal glucose tolerance or prediabetes).

Another, US based study did differentiate between types of remission (Karter et al., 2014). They studied 122,781 adults with type 2 diabetes and found an incidence of remission to a “sub-diabetic hyperglycaemia” state of 2.8 (95% CI: 2.6–2.9) per 1,000 person years. This is equivalent to a transition probability of: 0.00280. As this study differentiated between normal glucose tolerance and non-diabetic hyperglycaemia, it is this estimate that was used.

Death

The probability of death was obtained from Office for National Statistics population life tables, 2018-2020 (Office for National Statistics, 2021). Here, “qx” was used, which represents the probability that a person aged x will die before reaching age (x +1). The estimates for male and females were combined by taking an average of the probabilities. These estimates were used for individuals in the normal glucose tolerance and the non-diabetic hyperglycaemic states. The Office for National Statistics is the best source for this data, as it is routinely collected and representative of the English population.

For individuals in the type 2 diabetes state, an excess mortality risk was applied. To do this, the population life table estimates of death were multiplied by a hazard ratio. Estimates from the DECODE study were used, which used data from 22 cohorts across Europe, amounting to 29,714 subjects to determine the relationship between plasma blood glucose levels and mortality (DECODE Study Group European Diabetes Epidemiology Group, 2003). This study reported a hazard ratio of 1.6 (95% CI: 1.4–1.8) for all-cause mortality amongst individuals with diabetes compared to normal glucose tolerance. No other source of excess mortality for type 2 diabetes could be found in the literature that had a similar number of subjects and were more relevant to the population being considered.

Costs

In the absence of primary data collected on participants of the NHS DPP, or routinely collected data demonstrating primary and secondary healthcare resource use, published literature was searched to determine the costs of being in each of the model states. These costs include: the annual cost of having type 2 diabetes, the annual cost of having non-diabetic hyperglycaemia and the annual cost of having normal glucose tolerance.

There have been several literature reviews of decision-analytic models in the prevention of type 2 diabetes (Alouki et al., 2016; Leal et al., 2019; Li et al., 2010; Roberts et al., 2017; Zhou et al., 2020). These reviews identified the following studies with a UK focus: A. J. Palmer et al. (2004), Gillies et al. (2008), Anokye et al. (2011)*, Miners et al. (2012)*[1], Gillett et al. (2015) and Breeze et al. (2017). There have also been three modelling studies focusing on the UK setting published since these reviews were conducted: Roberts et al. (2018), Leal et al. (2020) and Frempong et al. (2021). The five most recently published of these models were consulted to determine where they

[1] *Miners and Anokye focused more on obesity related interventions.

sourced their costings and what cost components they included. The most recent of these studies, Frempong et al. (2021), was consulted first, as these would perceivably be using the most up to date sources, assuming that they obtained their cost estimates from literature searches.

A table of the costs used in five of the most recent UK-based models (published from 2015 to 2021), along with the sources for the costings (where available) is shown below (Table A4.2).

In addition to this, a simple literature search was conducted to see if there were any recent publications related to the cost of treating non-diabetic hyperglycaemia or type 2 diabetes beyond what were used in these published modelling studies. To do this, the following search terms were used “cost” “non-diabetic hyperglycaemia/type 2 diabetes” and “UK” in the PubMed database.

Literature Search – Cost of Type 2 diabetes

This search identified no newly published studies on the cost of treating type 2 diabetes, beyond those use in the studies published by Roberts et al. (2018) and Leal et al. (2020). This lack of costing data was confirmed by a systematic review titled “The Economic Costs of Type 2 diabetes: A Global Systematic Review” (Seuring et al., 2015), which only found one UK specific study. This study looked at the cost associated with people with type 2 diabetes and their employment, rather than the cost of their treatment from the perspective of the NHS. A published report by Kanavos et al. (2012) titled “Diabetes expenditure, burden of disease and management in 5 EU countries” also identified the lack of data for the UK according to diabetes type, and as a result, they estimated the cost of type 2 diabetes using two studies: C Ll Morgan et al. (2010) and Currie et al. (2010), estimating an annual cost of £3,717 for type 2 diabetes. The way in which these costs were combined is not fully described, beyond stating “a simple combination” was used. These two studies are discussed below:

- C Ll Morgan et al. (2010) used data from an area of Wales for all residents (approximately 439,000) in 2004 to estimate the secondary health care costs of individuals with type 2 diabetes, it also reported on the healthcare resource use of individuals without type 2 diabetes.
- Currie et al. (2010) conducted a UK specific analysis using data from The Health Improvement Network (THIN) to estimate primary care costs for 114,752 individuals with type 2 diabetes from 1997 to 2007. Importantly, this study also

obtained healthcare resource use for a cohort of non-diabetic controls, matched by age, sex and their general practice. As such, this study could be used to estimate the average healthcare resource use of individuals with normal glucose tolerance. The authors reported the resource use units for each aspect of primary care, making it possible to cost these using updated costings.

The literature review also identified a poster presentation which costed healthcare resource use for individuals with type 2 diabetes prescribed Exenatide or Basal Insulin therapy within the UK (C LI Morgan et al., 2016).

Literature Search – Cost of Non-diabetic Hyperglycaemia

There were no newly published estimates of the cost of non-diabetic hyperglycaemia in the UK beyond what have been referenced in the modelling studies below (Table A4.2).

Literature Search – Cost of Normal glucose tolerance

No published estimates subsequent to those used in the modelling studies were identified.

Cost used for Type 2 diabetes

From the estimates identified from the literature and other modelling studies (Alva et al., 2015; M. J. Davies et al., 2008; Hex et al., 2012; Kanavos et al., 2012), the estimate published by Kanavos et al. (2012) was selected for use in the model, which combined data from C LI Morgan et al. (2010) and Currie et al. (2010). Both of these estimates used a large sample size and are relevant to the English setting of the model. As such they reflected the best source to use for the annual cost of the type 2 diabetes modelling state.

Cost used for Non-diabetic Hyperglycaemia & Normal glucose tolerance

Of the literature estimates, the study by Nichols et al. (2008) was used for both normal glucose tolerance and non-diabetic hyperglycaemia. Whilst this study was not directly representative of the English population modelled (being from the United States), they did present the mean number of each resource use type used, which enabled UK cost estimates to be applied. In this paper, an assumption is made that individuals with impaired fasting glucose, and impaired glucose tolerance have the same healthcare resource use as individuals with non-diabetic glycaemia identified with elevated HbA1c. The unit costs applied are shown in Table A4.3.

From these calculations, the annual cost of having normal glucose tolerance is estimated to be: £2,005.31, and the annual cost of having non-diabetic hyperglycaemic to be: £2,224.76.

Table A4.2: Cost and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives.

		Glycaemic State		
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**
Frempong et al. (2021) (2018 price year)	Cost used	£1,179 This paper did not assume a linear increase overtime.	£869	£773
	Source	Same cost sources as Roberts 2018 (described below)	Same cost sources as Roberts 2018 (described below)	Same cost sources as Roberts 2018 (described below)
	Notes on source	-	-	-
Leal et al. (2020) (2017 price year)	Cost used	£827 to £2,792 Costs varied by sex and age group (up to 50, 51-60, 61-70, 71-80, 81+) The above costs excluded complications.	£636 to £2,149 Costs varied by sex and age group (up to 50, 51-60, 61-70, 71-80, 81+) "IGT costs estimated by applying ratio of IGT and diabetes costs per individual from Khan 2007 (23) for the US setting (0.74) and DPP (22) for the UK setting (0.77)"	Does not have a normal glucose tolerance state.
	Source	Alva et al. (2015)	(22) T. Khan et al. (2017) (23) Herman et al. (2005)	

Table A4.2: Cost and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State		
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**
			It is unclear whether the references here are switched – in the publication they do not correspond to the author mentioned. As such, it was assumed that the UK estimates were sourced from Herman et al., rather than the (22) cited in the text of the manuscript (which corresponds to Khan).	
	Notes on source	This study used patient-level data from the UKPDS, Hospital Episode Statistics to estimate secondary care costs and resource use questionnaires were administered during clinic visits of the trial, to estimate primary care visits. It is difficult to disentangle how Leal et al. produced the cost estimates cited. The paper focused on diabetes complications. In the paper they stated that the average cost for non-inpatient	This study described a Markov simulation model using data from the US DPP and published reports. The Herman study estimated base case costs of prediabetes as $1,296, and sourced these from US DPP data.	

Table A4.2: Cost and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State		
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**
		resource use was: £676.21 (all), £550.75 (no complications). (Obtained from Table 5 in the supplementary material). The average cost for inpatient visits was £1,351.52 (all) and £767 (no complications) (obtained from Table 4 in the supplementary material).		
Roberts et al. (2018) (2015 price year)	Cost used	£1,179 to £2,939 The annual cost increased linearly from Year 1-15. For probabilistic sensitivity analysis, they used a Gamma distribution with a standard error of 270.00.	£869.00 In the model Roberts et al. evaluates the different categorisations of elevated blood glucose: IFG, IGT and HbA1c. “IFG costs are 73% of T2DM costs” where the type 2 diabetes (T2DM) costs excluded the cost of complications.	£773.00 Costs associated with normal glucose tolerance were assumed to be 66% of type 2 diabetes costs without complication costs.

Table A4.2: Cost and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State		
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**
			The authors make the assumption that individuals with non-diabetic hyperglycaemia (identified via raised HbA1c) are equal to those of IFG.	
	Source	Hex et al. (2012)	Bächle et al. (2016) Nichols et al. (2008)	Bächle et al. (2016) Nichols et al. (2008)
	Notes on source	Hex et al. estimated the prevalence of type 2 diabetes for 2010/11 as 3,419,727. They identified costs according to the following categories and estimated the annual cost associated with them. This paper looked at type 1 and type 2 diabetes, but presented the costs separately for each of these conditions: Screening £10,588,726 Treatment £1,75,615,980 Complications £7,000,037,553	Bächle et al. described a population based study from Germany. This study reported the average resource use, making it possible to assign UK costs to these estimates. Resource use was derived from the statutory health insurances' data. The Nichols et al. study is based in the US, looking at individuals identified with prediabetes between 1998 and 2004. They present estimates in terms of IFG and IFT.	Bächle et al. describes a population based study from Germany. This study reported the average resource use, making it possible to assign UK costs to these estimates. Resource use was derived from the statutory health insurances' data. The Nichols et al. study is based in the US, looking at individuals identified with prediabetes between 1998 and 2004.

Table A4.2: Cost and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State		
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**
		The treatment costs included costs for "education programmes" which would need to be excluded if used in the current model, as the DPP costs are assigned separately. This leads to an annual average cost of treating type 2 diabetes of £2,101.41 (2010/11 price year). Inflating this to 2020 price year, this is equivalent to an annual cost of £2,499.17 (inflated using: Shemilt et al. (2010)).	This study also presented the average resource use, making it possible to assign UK costs to these estimates.	This study also presented the average resource use, making it possible to assign UK costs to these estimates.
Breeze et al. (2017)		This paper described an individual patient simulation model, and as such did not describe costs by blood glucose category (Type 2 diabetes/ Non-diabetic hyperglycaemia /Normal control), and instead individually presented costs, prescriptions and healthcare resource use dependant on individual characteristics and comorbidities.		
Gillett et al. (2015)	Cost used	Not directly reported (used Sheffield type 2 diabetes model)	Not directly reported (used Sheffield type 2 diabetes model)	Not directly reported (used Sheffield type 2 diabetes model)
	Source	Reports using data directly from the DESMOND trial (M. J. Davies et al., 2008) (resource use over		

Table A4.2: Cost and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State		
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**
		12 months, recruited 2004-2006) for individuals newly diagnosed with Type 2 diabetes (n=824, 55% men, mean age 59.5 years). Only information on primary care health care resource use was collected.		
	Notes on source	Trial data were used to identify use of drugs and use of general practitioners' and other primary care professionals' time over 12 months. They did not collect secondary care use and instead used literature estimates. The authors present the mean use of primary care resources over 12 months (for both control and DESMOND trial members). Looking at the control arm, the average cost of primary care (excluding prescriptions) was: £264.47.		

Table A4.3: Calculating cost of normal glucose tolerance and non-diabetic hyperglycaemia health states from published resource use by Nichols et al. (2008).

	States of glycaemic control					
	Normal	**IFG**	**IFT**	**IFG/IFT**	**Unit Cost**	**Source**
Cost Element	(n=15,629)	(n=5,713)	(n=2,552)	(n=2,217)		
Inpatient						
Mean number reported	0.14	0.15	0.26	0.22	£2,941	Using weighted average of Elective Inpatients, Non Elective Inpatients and Non-elective short stay (NHS England, 2022)
Total cost for sample[1]	£6,436,146	£2,520,706	£1,951,734	£1,434,680		
Outpatient visits						
Mean number reported	3	3	3.6	3.5	£39	Assumed cost of primary care visit equates to a GP consultation (Curtis & Burns, 2020)
Total cost for sample[1]	£1,828,593	£668,421	£358,301	£302,621		
Urgent visits						
Mean number reported	0.8	0.7	1.1	1	£297	Assumed cost of A&E visit. Sourced from National Schedule of NHS Costs 2020 (NHS England, 2022)
Total cost for sample[1]	£3,713,450	£1,187,733	£833,738	£658,449		
Speciality visits						
Mean number reported	5.7	5.5	6	5.8	£182	Using weighted average of outpatient unit costs (NHS England, 2022)
Total cost for sample[1]	£16,172,270	£5,704,162	£2,779,693	£2,334,310		
Pharmaceutical						
Mean number reported	23.6	25.7	31.7	32.1	£8.65	Average cost per item (NHS Business Services Authority, 2021)

Table A4.3: Calculating cost of normal glucose tolerance and non-diabetic hyperglycaemia health states from published resource use by Nichols et al. (2008). *(Continued)*

	States of glycaemic control					
	Normal	**IFG**	**IFT**	**IFG/IFT**	**Unit Cost**	**Source**
Cost Element	(n=15,629)	(n=5,713)	(n=2,552)	(n=2,217)		
Total cost for sample[1]	£3,190,504	£1,270,028	£699,771	£615,583		
Overall total cost[2]	£31,340,963	£11,351,050	£6,623,238	5,345,643		
Average per person[3]	£2,005.31	£2,224.76				

Notes:

[1]Total cost for sample was calculated as the unit cost multiplied by the number in the sample multiplied by the mean number reported.

[2]Overall total cost is the sum of all of the cost element 'total cost for sample' costs.

[3]Weighted by number in each impaired glucose category. All resource use estimates sourced from Nichols et al. (2008).

Utility Scores

A table showing the sources of utility data for the five most recent UK modelling studies is shown in Table A4.4.

Data collected as part of the NHS DPP was used to inform the utility values used for individuals with non-diabetic hyperglycaemia. The responses to the EQ-5D-5L questionnaire, recorded at initial assessment of the NHS DPP, were valued using the Crosswalk value set (Van Hout et al., 2012). There were 116,004 referrals that had a baseline utility score recorded (out of referrals received to the programme prior to 1st April 2019). The average utility score for these referrals was: 0.808 (sd: 0.216) with a visual analogue score (VAS) of 74 (sd: 19). This estimate is higher than existing studies have previously estimated (ranging from 0.759 (Neumann et al., 2014) to 0.7302 (Herman et al., 2005)). This may be because clinical trials select individuals that are at a higher-risk and therefore, have a lower utility score. Reassuringly, there are other studies have found similar results to the estimates derived from NHS DPP data. In the feasibility of the Norfolk Diabetes Prevention Study, it was reported that the baseline characteristics of their non-diabetic hyperglycaemic population, had a utility score of 0.85 (n=177) (Irvine et al., 2011), which is higher than the average baseline of referrals to the NHS DPP.

Due to the higher baseline score, individuals in a pre-diabetic state would have a higher utility score than those in a normal state, which appears counterintuitive.

As such, data from the 2018 Health Survey for England (NatCen Social Research, 2022) was analysed to determine the utility scores of being in the following model states: normal glucose tolerance and type 2 diabetes. This is an annual survey which looks at the health and lifestyle of English people, surveying approximately 8,000 adults and 2,000 children. As part of this survey, the EQ-5D-5L is collected. The number of respondents used to estimate each utility score for the model states by age category is shown in Table A4.5.

To determine the utility score to be used in the type 2 diabetes state, the data was restricted to individuals who reported having type 2 diabetes in the survey and weighted according to the age/sex profile of individuals referred to the NHS DPP.

To determine the utility score to be used in the normal glucose tolerance state, the data was restricted to individuals not reporting diabetes, and excluded individuals with ‘bad’ self-reported health, as there was no question about prediabetes in the survey.

Table A4.4: Utility values and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives.

		Glycaemic State			
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**	**Incremental utility gain from DPP**
Frempong et al. (2021)	Value	0.738	0.759	0.768	0.0189
	Source	Neumann et al. (2014)	Neumann et al. (2014)	(Neumann et al., 2014)	Roberts et al. (2018)
	Notes on source	Measured using Short Form-36 questionnaire (SF-36). Study based in Sweden, between 2003 and 2013. 2,995 individuals with type 2 diabetes. Responses stratified by age, sex, education and BMI.	Measured using Short Form-36 questionnaire (SF-36). Study based in Sweden, between 2003 and 2013. 5,629 (IFG), 2,440 (IGT), 1,232 (IFG and IGT). Utility was 0.759 for IFG, 0.746 for IGT and 0.745 for those with both.	Measured using Short Form-36 questionnaire (SF-36). Study based in Sweden, between 2003 and 2013. 43,586 'healthy' individuals	Modelling study discussed below.
Leal et al. (2020)	Value	Initial utility 0.807 (although it is unclear if this corresponds to non-diabetic hyperglycaemia). Information was not presented according to different glycaemic control states. The supplementary material instead only stated an initial utility value, and gave decrements associated with complications. The model looked at delaying the onset of diabetes in individuals with cardiovascular disease or cardiovascular risk factors and IGT, and therefore it was assumed that this utility value referred to individuals with prediabetes.			

Table A4.4: Utility values and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State			
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**	**Incremental utility gain from DPP**
	Source	Alva et al. (2014)			
	Notes on source	EQ-5D questionnaires administered between 1997 and 2007 in the UK Prospective Diabetes Study (UKPDS). 0.807 is from a regression model looking at the presence of complications. Raw data from the waves of UKPDS show an average utility of 0.702 (amongst 3,380 individuals without complications who have type 2 diabetes).			
Roberts et al. (2018)	Value	0.738	0.759	0.768	0.0189
	Source	Neumann et al. (2014)	Neumann et al. (2014)	Neumann et al. (2014)	Herman et al. (2005)
	Notes on source	As described above.	As described above.	As described above.	States source as “DPP data”. The paper does not describe how this is determined or the length of time over which there is expected to be a benefit.

Table A4.4: Utility values and sources used in the five most-recently published, UK-specific, modelling studies evaluating diabetes prevention initiatives. *(Continued)*

		Glycaemic State			
Study		**Type 2 diabetes**	**Non-diabetic hyperglycaemia**	**Normal glucose tolerance**	**Incremental utility gain from DPP**
Breeze et al. (2017)	Value	Not directly stated.	Not directly stated.	Not directly stated.	0.0003 to 0.0009
	Source	Baseline utilities for all individuals in the cohort were extracted from the Health Survey for England, 2011. This data was not presented.			Dunkley et al. (2014)
	Notes on source	Authors of the modelling study state: “we assumed that a diagnosis of diabetes was not associated with a reduction in EQ-5D independent of the utility decrements associated with complications, comorbidities or depression”. “Utility was assumed to decline due to ageing independent of health status. In the simulation, utility declines by an absolute decrement of 0.004 per year” (sourced from supplementary material 1).			Meta-analysis looking at the effectiveness of diabetes prevention programmes. It included 22 studies with outcome data for weight loss at 12 months. The paper did not present results in terms of changes in quality adjusted life years or utility values.
Gillett et al. (2015)	Value	Not directly reported (used Sheffield type 2 diabetes model)			

Table A4.5: Sample size used for utility calculations.

	Sample size		
Age category	**Normal glucose tolerance**	**Non-diabetic hyperglycaemia***	**Type 2 diabetes**
<40	1,038	7,469	19
40-49	1,066	17,170	48
50-59	945	35,349	91
60-69	698	51,838	135
70-79	267	50,982	130
80+	304	15,642	68
Total sample size	4,318	178,450	491
*Utility scores for non-diabetic hyperglycaemia were estimated from referrals to the NHS DPP, recorded at initial assessment.			

Distributions

A distribution for each of the parameters included in the model was estimated (for example: transition probabilities, costs and utilities). As recommended by Briggs et al. (2006), Beta distributions were used for transition probabilities and utility estimates (given they were sufficiently removed from zero), and Gamma distributions for costs. The parameters for these distributions were obtained using the following formulas:

The alpha and beta parameters of the Gamma(α,β) distribution were calculated using:

$$\alpha = {}^{\mu^2}\!/_{s^2}$$

$$\beta = {}^{s^2}\!/_{\mu}$$

Where μ is sample mean, s^2 is variance.

The alpha and beta parameters of the Beta(α,β) distribution were calculated using:

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

$$\beta = \alpha * \frac{(1-\mu)}{\mu}$$

Where μ is sample mean, s^2 is variance.

Model Results

Example cohort trace

As the model was run using probabilistic analysis with 10,000 Monte Carlo simulations, it is not possible to provide all of the cohort traces. However, an example of one of the traces from these simulations is shown in Table A4.6.

Table A4.6: Example cohort trace over time (from 1 Monte Carlo simulation).

	Usual Care				**NHS DPP**			
Cycle	Normal glucose tolerance	Non-diabetic hyperglycaemia	Type 2 diabetes	Dead	Normal glucose tolerance	Non-diabetic hyperglycaemia	Type 2 diabetes	Dead
0	0	1000	0	0	0	1000	0	0
1	82.569	868.571	25.736	23.124	83.837	872.523	20.516	23.124
2	145.759	761.943	47.171	45.127	147.741	770.169	37.082	45.008
3	193.670	674.982	65.369	65.978	195.285	685.317	53.579	65.819
4	228.089	607.224	78.980	85.706	233.848	611.609	68.940	85.602
5	254.256	543.520	89.225	112.999	258.240	550.436	78.492	112.832
6	271.776	488.351	96.957	142.916	276.058	494.250	87.400	142.292
7	283.338	443.587	103.048	170.027	287.379	449.099	94.158	169.363
8	288.880	407.113	109.083	194.924	294.657	409.968	101.264	194.112
9	292.076	376.006	114.424	217.495	297.894	378.118	107.148	216.839
10	287.884	344.102	116.651	251.363	291.069	347.207	110.938	250.786
11	281.927	314.835	118.975	284.263	284.404	319.390	112.593	283.612
12	275.642	289.838	121.634	312.886	277.196	295.457	114.837	312.511
13	265.893	272.569	122.633	338.905	268.918	275.955	117.015	338.112
14	259.011	255.150	123.567	362.272	260.402	258.670	119.179	361.750

Table A4.6: Example cohort trace over time (from 1 Monte Carlo simulation). *(Continued)*

	Usual Care				**NHS DPP**			
Cycle	Normal glucose tolerance	Non-diabetic hyperglycaemia	Type 2 diabetes	Dead	Normal glucose tolerance	Non-diabetic hyperglycaemia	Type 2 diabetes	Dead
15	245.328	236.243	121.122	397.307	247.134	237.796	117.803	397.267
16	234.718	218.152	118.602	428.529	234.300	222.221	114.559	428.920
17	223.654	203.838	117.104	455.405	220.512	209.834	113.608	456.046
18	215.751	188.302	116.320	479.627	210.971	195.967	113.044	480.018
19	207.791	176.078	114.763	501.368	202.450	184.020	111.938	501.592
20	190.710	164.861	109.850	534.580	189.608	168.328	107.611	534.453
21	178.407	153.257	105.038	563.298	178.396	154.329	104.015	563.261
22	166.714	143.453	101.354	588.479	166.815	144.375	101.127	587.683
23	157.043	133.940	98.730	610.287	156.774	135.523	98.079	609.624
24	149.124	124.858	96.499	629.519	147.449	128.368	95.293	628.890
25	136.951	114.814	90.488	657.747	135.951	117.236	89.315	657.498
26	126.734	105.278	85.954	682.034	126.793	107.215	84.563	681.430
27	117.369	98.370	81.275	702.986	118.021	99.303	79.924	702.752
28	109.071	91.957	78.179	720.793	109.754	92.743	76.349	721.154
29	101.390	86.588	75.090	736.932	102.176	87.723	73.109	736.992
30	91.674	78.757	68.809	760.760	91.684	79.888	67.777	760.650
31	83.720	71.683	63.580	781.017	83.892	72.589	62.751	780.768
32	76.711	65.689	59.746	797.855	77.204	66.359	58.872	797.565
33	71.075	59.943	56.690	812.292	71.373	60.961	55.426	812.239
34	65.944	55.407	53.870	824.779	67.084	55.860	52.190	824.867
35	59.148	49.828	48.773	842.251	59.906	50.423	47.294	842.378

Scatter Plots from sensitivity analyses

Figure A4.1: Sensitivity analysis 1

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 89, North-West: 3, South-East: 9,907, South-West: 1. Percentage cost-effective at £20,000 willingness to pay (£30,000): 99.9% (99.9%)

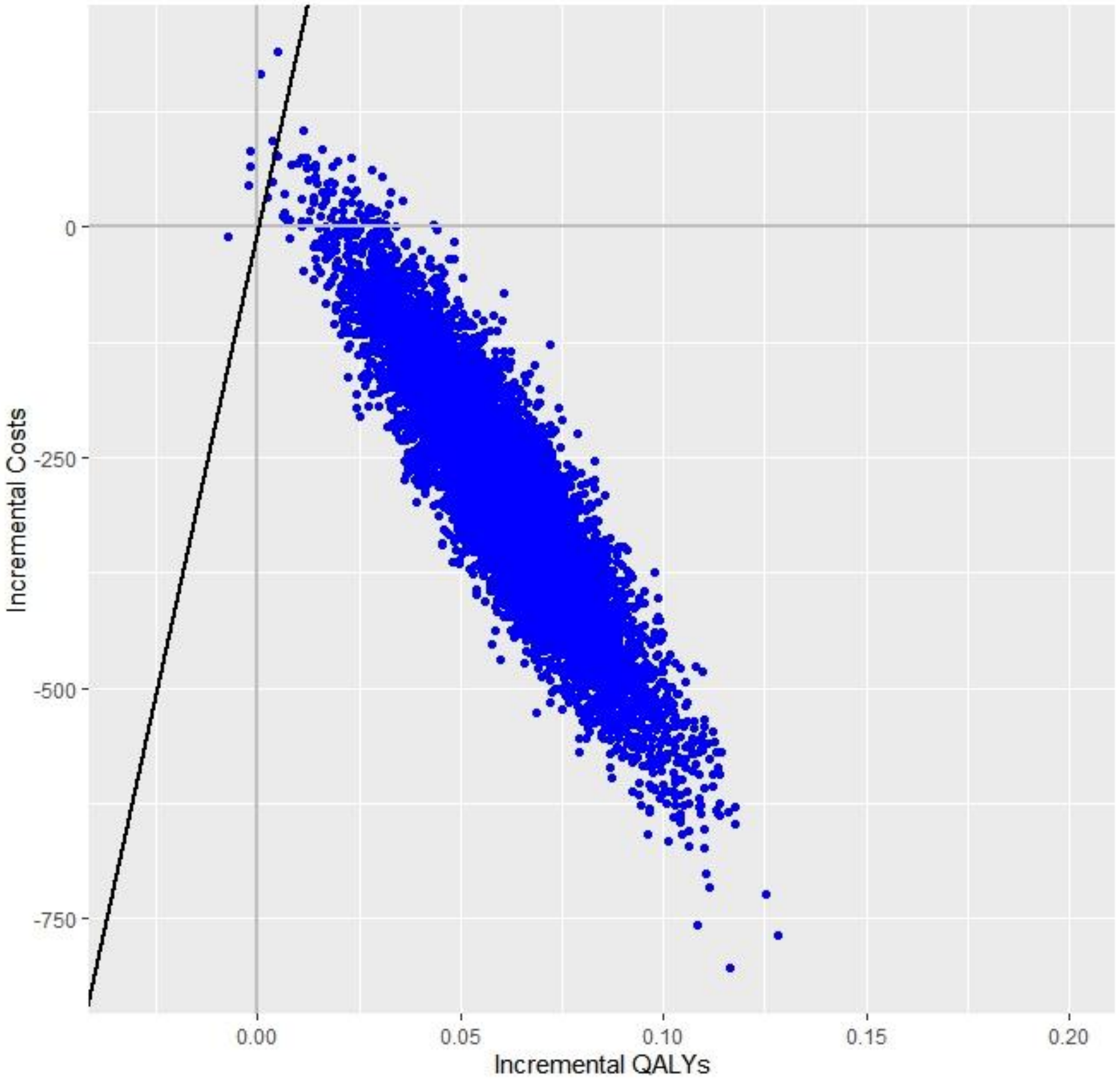


Figure A4.2: Sensitivity analysis 2

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 39, North-West: 0, South-East: 9,960, South-West: 1. Percentage cost-effective at £20,000 willingness to pay (£30,000): 100.0% (100.0%)

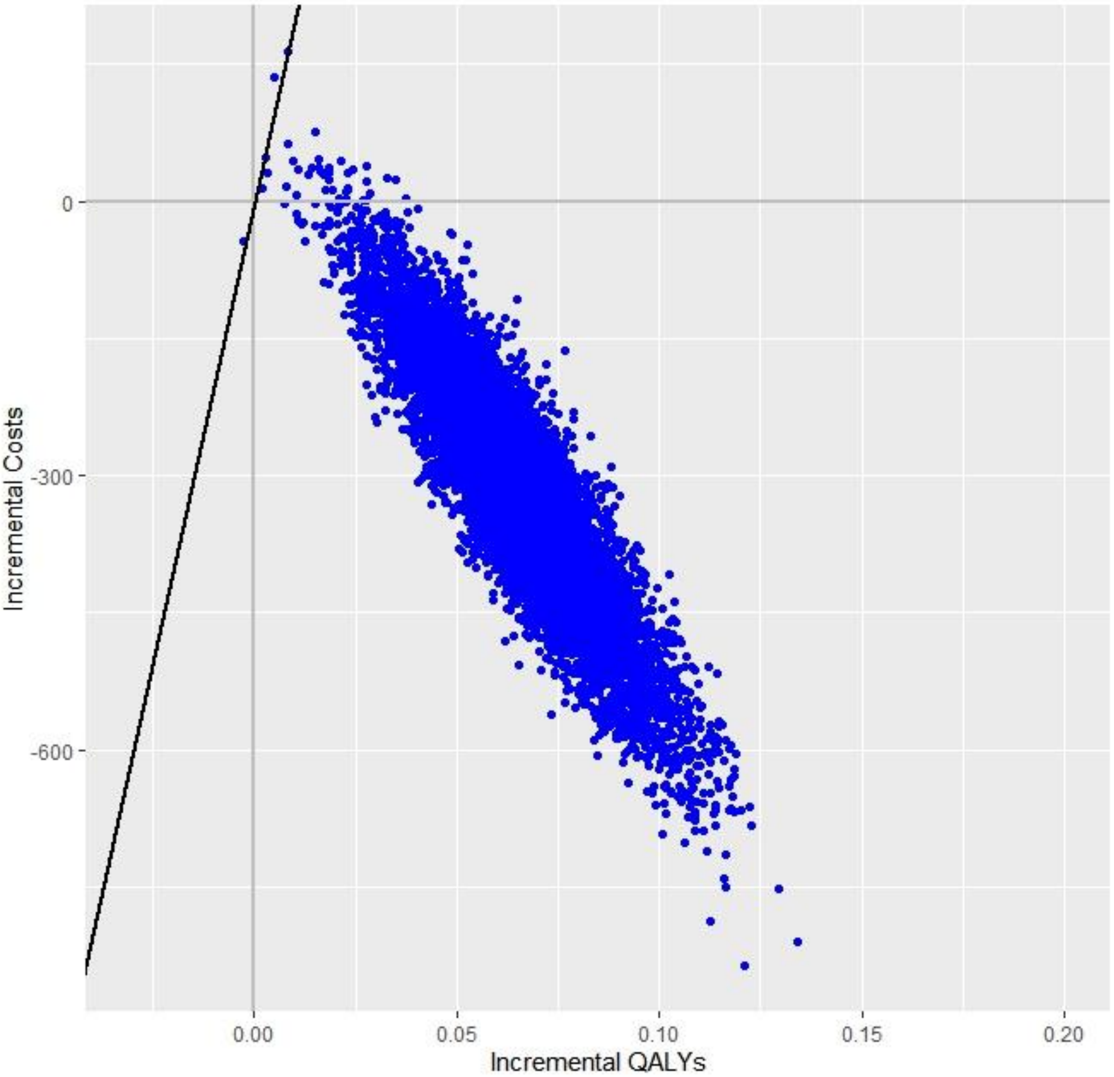


Figure A4.3: Sensitivity analysis 3

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 3,533, North-West: 91, South-East: 6,376, South-West: 0. Percentage cost-effective at £20,000 willingness to pay (£30,000): 97.9% (98.4%)

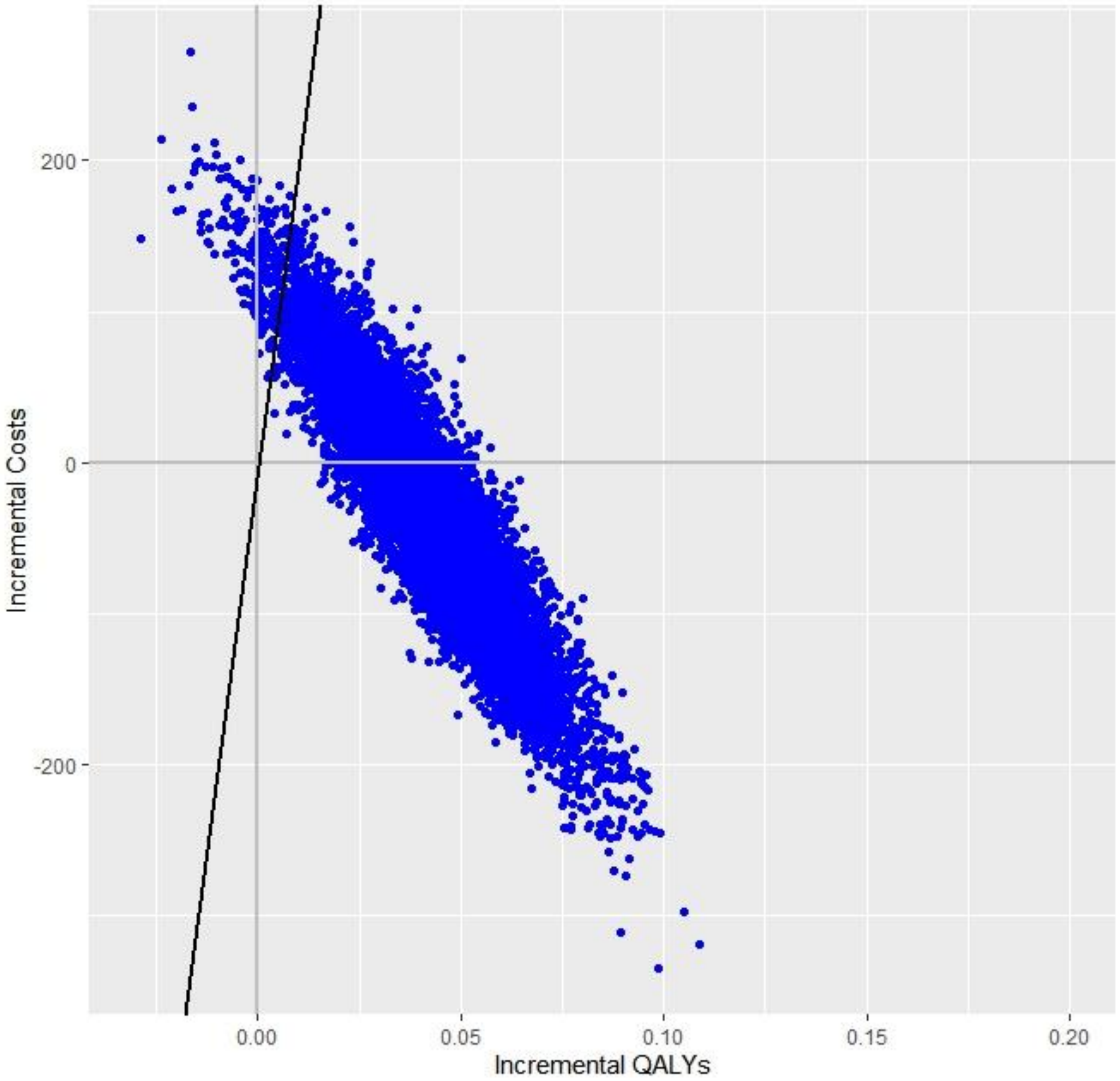


Figure A4.4: Sensitivity analysis 4

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 26, North-West: 0, South-East: 9,974, South-West: 0. Percentage cost-effective at £20,000 willingness to pay (£30,000): 100.0% (100.0%)

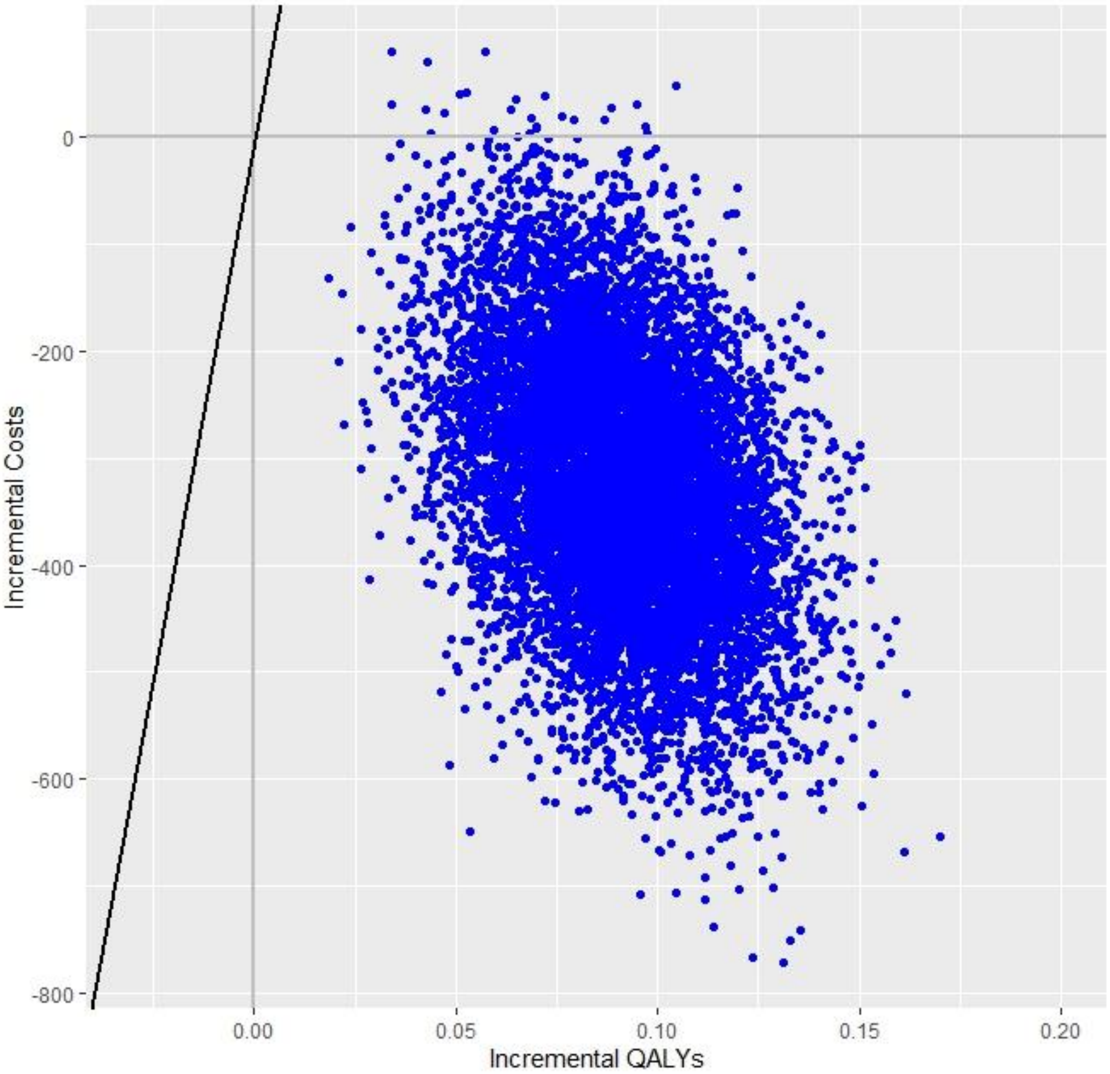


Figure A4.5: Sensitivity analysis 5

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 405, North-West: 0, South-East: 9,595, South-West: 0. Percentage cost-effective at £20,000 willingness to pay (£30,000): 100.0% (100.0%)

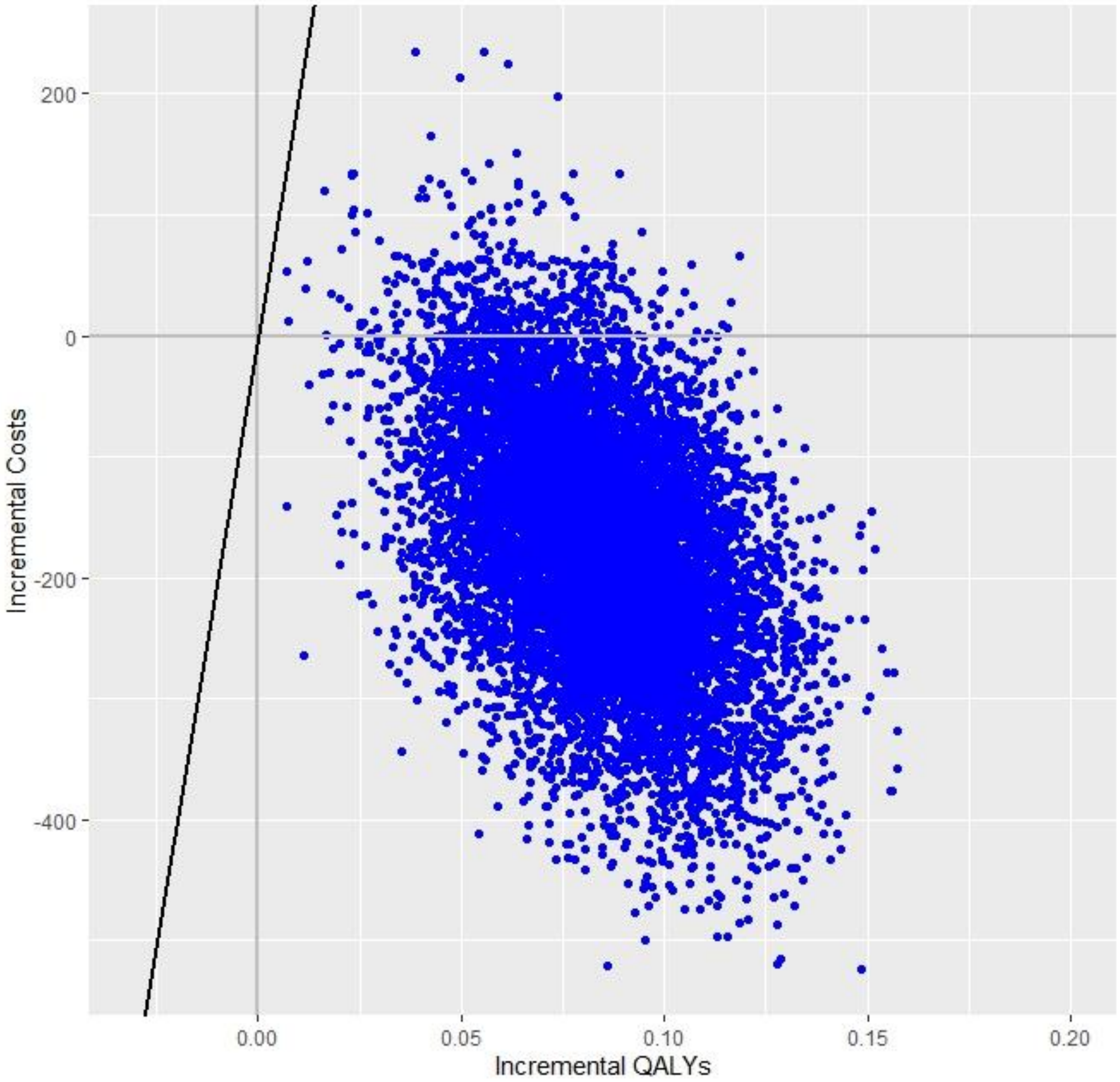
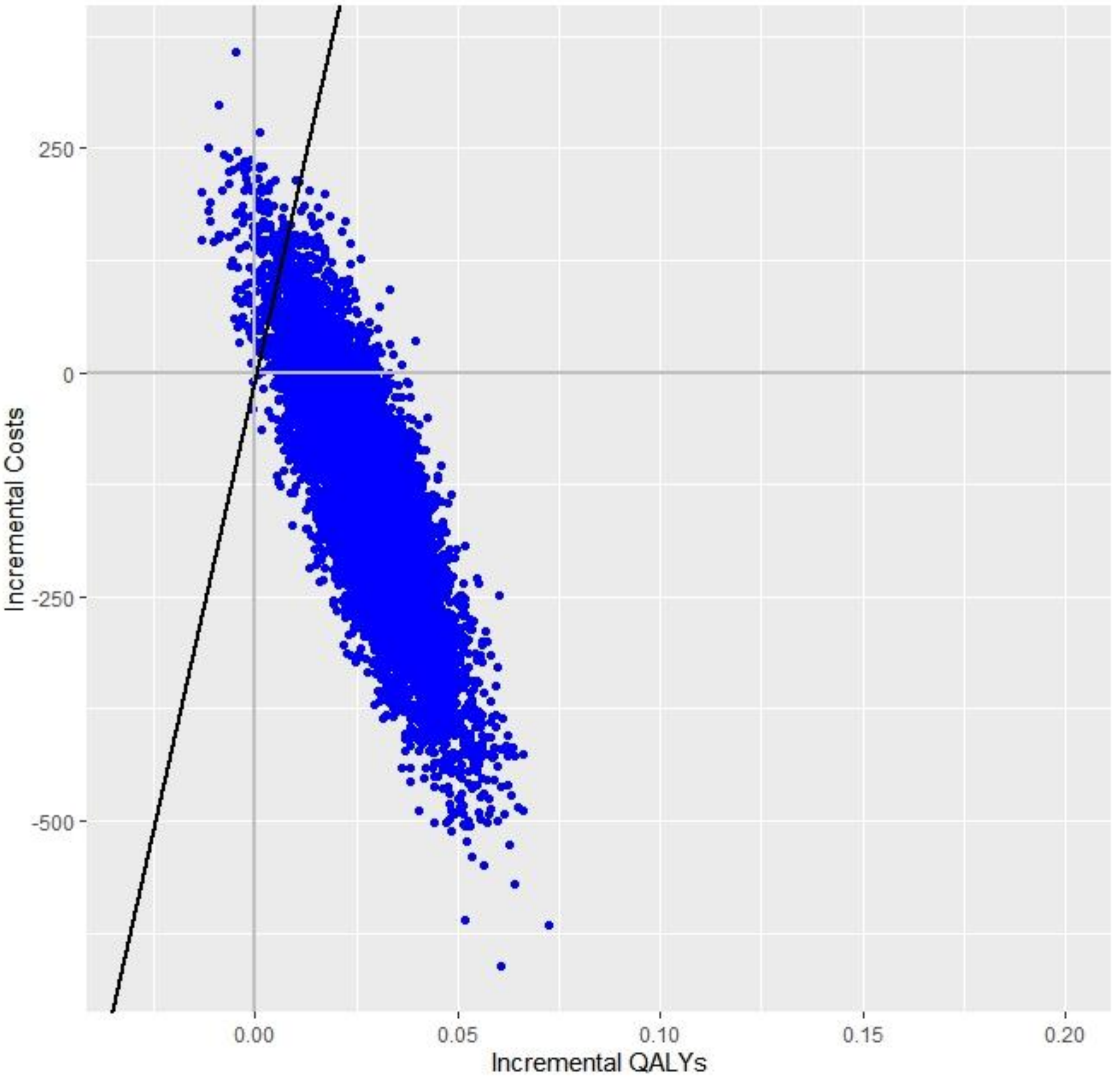


Figure A4.6: Sensitivity analysis 6

Scatter plot of incremental cost and QALY pairs from 10,000 Monte Carlo simulations, on average for an individual in the modelled cohort. The black line represents a willingness to pay threshold of £20,000 per QALY gained. The number of points in each quadrant are: North-East: 1,311, North-West: 75, South-East: 8,612, South-West: 2. Percentage cost-effective at £20,000 willingness to pay (£30,000): 98.0% (98.5%)



Code Excerpt – Base case analysis

Below is an excerpt of the model code used in the base case analysis of the model, which is available in full from the author’s Github repository (McManus, 2023).

It includes the code used for cohort of age category x<40, which is then repeated for the other age categories used (but for brevity omitted here). The model parameters are first defined as point estimates and then updated in the code using distributions. From line 372 is the code that is used to combine the results of the different age categories modelled and run analysis of the results.

```
1   ########################################
2   # Markov model: Evaluating the NHS DPP #
3   # 4 modelling states:
4   # Normal glycaemic control
5   # NDH
6   # Type 2 Diabetes
7   # Death
8   ########################################
9   getwd()
10  ## SET WORKING DIRECTORY ##
11  setwd("…")
12  library('ggplot2')
13  library('ggthemes')
14  library("dplyr")
15  #Cohort is made up of 6 age categories / model runs
16  ############
17  # AGE CAT 1
18  ############
19  set.seed(9009)
```

```
## model set-up ----
t_names <- c("without DPP", "with DPP")
n_treatments <- length(t_names)
s_names  <- c("Normal", "NDH", "T2D", "Dead")
n_states <- length(s_names)
n_cohort <- 58
n_cycles <- 36
Initial_age <- 35
#NUMBER OF PSA ITERATIONS
n_trials<-10000
costs <- matrix(NA,nrow=n_trials,ncol=n_treatments,
dimnames=list(NULL,t_names))
qalys <- matrix(NA,nrow=n_trials,ncol=n_treatments,
dimnames=list(NULL,t_names))
#Cost of states
cNorm <- rgamma(1, 44.44444, scale=45.119475)
cNDH <- rgamma(1,  44.44444, scale=50.0571)
cT2D <- rgamma(1,  44.44444, scale=99.4628)
cDeath <- 0
cDPP <- 0
#Utility of states
uNorm <- rbeta(1,27.53076303,4.883160879)
uNDH <- rbeta(1,26.33007731, 6.334144666)
uT2D <- rbeta(1, 26.52841, 10.65937)
uDeath <-0
uDPP<-0
#Discount rates
oDr <- 0.035
```

```
cDr <- 0.035
#utility gains from DPP
utility <- c(1-rbeta(n_cohort,801123.6265,3464.71833))
utility_sum <-sum(utility)
#costs of the DPP
dppcost <- c(rgamma(n_cohort,16.73964,scale=8.469118))
dppcost_sum <- sum(dppcost)
#Transition probabilities (later replaced with PSA values)
#normal to NDH
tpProgNDH <- 0.074202731
#dying from NGT or NDH
tpDeath <- 0.05
#Progression to T2D from NDH
tpProgT2D <- 0.0249
#From NDH to NGT
tpNorm <- 0.0795
#From T2D to NDH
tpRegress <- 0.002796084
#Additional risk of death for T2D
tpExcessDeath <- 1.6

#change in transition probability for DPP (later replaced with PSA values)
effect <- 0.8
# Cost of staying in state
state_c_matrix <-
  matrix(c(cNorm, cNDH, cT2D, 0,
           cNorm, cNDH + cDPP, cT2D, 0),
         byrow = TRUE,
         nrow = n_treatments,
         dimnames = list(t_names,
                         s_names))
# qaly when staying in state
state_q_matrix <-
  matrix(c(uNorm, uNDH, uT2D, 0,
           uNorm, uNDH + uDPP, uT2D, 0),
         byrow = TRUE,
         nrow = n_treatments,
         dimnames = list(t_names,
                         s_names))
# cost of moving to a state
# same for both treatments - no cost of death applied
trans_c_matrix <-
  matrix(c(0, 0, 0, cDeath,
           0, 0, 0, cDeath,
           0, 0, 0, cDeath,
           0, 0, 0, 0),
         byrow = TRUE,
         nrow = n_states,
         dimnames = list(from = s_names,
                         to = s_names))
# Transition probabilities ----
# time-homogeneous
#not by row - by column
p_matrix <- array(data = c(1-tpProgNDH - tpDeath, tpNorm, 0, 0,
                           tpProgNDH, 1-tpProgT2D-tpDeath-tpNorm, tpRegress,
0,
                           0, tpProgT2D,  1-tpRegress -
(tpDeath*tpExcessDeath), 0,
                           tpDeath, tpDeath, (tpDeath*tpExcessDeath), 1,
                           #below would be adjusted for effect of DPP
                           1-tpProgNDH - tpDeath, tpNorm, 0, 0,
                           tpProgNDH, 1-(tpProgT2D*effect)-tpDeath-tpNorm,
tpRegress, 0,
                           0, (tpProgT2D*effect),  1-tpRegress -
(tpDeath*tpExcessDeath), 0,
                           tpDeath, tpDeath, (tpDeath*tpExcessDeath), 1),
                  dim = c(n_states, n_states, n_treatments),
                  dimnames = list(from = s_names,
                                  to = s_names,
                                  t_names))
# Store population output for each cycle
# state populations
pop <- array(data = NA,
             dim = c(n_states, n_cycles, n_treatments),
             dimnames = list(state = s_names,
                             cycle = NULL,
                             treatment = t_names))
pop["Normal", cycle = 1, ] <-0
pop["NDH", cycle = 1, ] <- n_cohort
pop["T2D", cycle = 1, ] <- 0
pop["Dead", cycle = 1, ] <- 0

halfpop<-array(data=NA, dim=c(n_states,n_cycles,n_treatments),
               dimnames = list(state= s_names, cycle=NULL, treatment=
t_names))


halfpop["Normal", cycle=1, treatment="without DPP"]<- 0

halfpop["NDH", cycle=1, treatment="without DPP"]<- 0

halfpop["T2D", cycle=1, treatment="without DPP"]<- 0

halfpop["Dead", cycle=1, treatment="without DPP"]<- 0

halfpop["Normal", cycle=1, treatment="with DPP"]<- 0

halfpop["NDH", cycle=1, treatment="with DPP"]<- 0

halfpop["T2D", cycle=1, treatment="with DPP"]<- 0

halfpop["Dead", cycle=1, treatment="with DPP"]<- 0


# _arrived_ state populations

trans <- array(data = NA,

               dim = c(n_states, n_cycles, n_treatments),

               dimnames = list(state = s_names,

                               cycle = NULL,

                               treatment = t_names))

trans[, cycle = 1, ] <- 0

# Sum costs and QALYs for each cycle at a time for each treatment

cycle_empty_array <-

  array(NA,

        dim = c(n_treatments, n_cycles),

        dimnames = list(treatment = t_names,

                        cycle = NULL))


cycle_state_costs <- cycle_trans_costs <- cycle_qalys <- cycle_empty_array

cycle_costs <- cycle_QALYs <- cycle_empty_array
```

```
LE <- LYs <- cycle_empty_array    # life expectancy; life-years
cycle_QALE <- cycle_empty_array   # quality-adjusted life expectancy

total_costs <- setNames(c(NA, NA), t_names)
total_QALYs <- setNames(c(NA, NA), t_names)

# Transition probabilities (later updated with PSA)
#normal to NDH
tpProgNDH <- 0.074202731
# dying from NGT or NDH
tpDeath <- 0.05
# Progression from NDH to T2D
tpProgT2D <- 0.0249
# From NDH to NGT
tpNorm <- 0.0795
# From T2D to NDH
tpRegress <- 0.002796084
# Additional risk of death for T2D
tpExcessDeath <- 1.6

# change in transition probability for DPP
effect <- 0.8

tpNorm <- rbeta(1,335,3709)
tpProgT2D <- rbeta(1,25.27,974.73)
tpRegress <- rbeta(1,2.8,997.2)
tpProgNDH <- rbeta(1,77.1,922.9)
```

```
tpExcessDeath <- rlnorm(1, meanlog = 0.470004, sdlog = 0.064111)

# Time-dependent probability matrix ----
# Time dependent transition - risk of death
p_matrix_cycle <- function(p_matrix, age, cycle,
                           tpProgNDH = rbeta(1,77.1,922.9),
                           #tpProgT2D = rbeta(1,25.27,974.73),
                           tpNorm = rbeta(1,335,3709),
                           tpRegress = rbeta(1,2.8,997.2),
                           tpExcessDeath = rlnorm(1, meanlog = 0.470004,
sdlog = 0.064111)
) {
  tpDeath_lookup <-
    c("[30,34]" = 0.000674,
      "[35,39]" = 0.001007,
      "[40,44]" = 0.001499,
      "[45,49]" = 0.002298,
      "[50,54]" = 0.003359,
      "[55,59]" = 0.005056,
      "[60,64]" = 0.007940,
      "[65,69]" = 0.012389,
      "[70,74]" = 0.019640,
      "[75,79]" = 0.034483,
      "[80,84]" = 0.060938,
      "[85,89]" = 0.110627,
      "[90,94]" = 0.187364,
      "[95,120]" = 0.295984)

  age_grp <- cut(age, breaks = c(30,34,39,44,49,54,59,64,69,74,79, 84, 89,
94, 120))
  tpDeath <- tpDeath_lookup[age_grp]

  tpProgT2D_lookup <-
    c("[30,39]" = rbeta(1,27.12014,972.8799),
      "[40,49]" = rbeta(1,34.49617,965.5038),
      "[50,59]" = rbeta(1,30.72741,969.2726),
      "[60,69]" = rbeta(1,25.45268,974.5473),
      "[70,79]" = rbeta(1,20.27103,979.729),
      "[80,120]" = rbeta(1,14.16019,985.84731))

  age_grpT2D <- cut(age, breaks= c(30,39,49,59,69,79,120))
  tpProgT2D <-tpProgT2D_lookup[age_grpT2D]

  # [ includes, ( doesn't include
  effect_lookup <-
    c("(0,3]" = rlnorm(1, meanlog = -0.22314, sdlog = 0.044757),
      "(3,10]" = rnorm(1, mean = 1.000, sd = 0.00000001),
      "(10,15]"= rnorm(1, mean = 1.000, sd = 0.00000001),
      "(15,50]" = rnorm(1, mean = 1.000, sd = 0.00000001))

  cycle2 <- cut(cycle, breaks = c(0,3,10,15,50))
  effect <- effect_lookup[cycle2]
  #########
  p_matrix["Normal", "Normal", "without DPP"] <- 1 - tpProgNDH - tpDeath
  p_matrix["Normal", "NDH", "without DPP"] <- tpProgNDH
  p_matrix["Normal", "Dead", "without DPP"] <- tpDeath
  p_matrix["NDH", "Normal", "without DPP"] <- tpNorm
  p_matrix["NDH", "NDH", "without DPP"] <-1-tpProgT2D-tpDeath-tpNorm
  p_matrix["NDH", "T2D", "without DPP"] <- tpProgT2D
  p_matrix["NDH", "Dead", "without DPP"] <- tpDeath
  p_matrix["T2D", "NDH", "without DPP"] <- tpRegress
  p_matrix["T2D", "T2D", "without DPP"] <- 1-tpRegress -
(tpDeath*tpExcessDeath)
  p_matrix["T2D", "Dead", "without DPP"] <- (tpDeath)*tpExcessDeath
  p_matrix["Dead", "Dead", "without DPP"] <- 1
  p_matrix["Normal", "Normal", "with DPP"] <- 1 - tpProgNDH - tpDeath
  p_matrix["Normal", "NDH", "with DPP"] <- tpProgNDH
  p_matrix["Normal", "Dead", "with DPP"] <- tpDeath
  p_matrix["NDH", "Normal", "with DPP"] <- tpNorm
  p_matrix["NDH", "NDH", "with DPP"] <- 1 - (tpProgT2D*effect) - tpDeath -
tpNorm
  p_matrix["NDH", "T2D", "with DPP"] <- (tpProgT2D*effect)
  p_matrix["NDH", "Dead", "with DPP"] <- tpDeath
  p_matrix["T2D", "NDH", "with DPP"] <- tpRegress
  p_matrix["T2D", "T2D", "with DPP"] <- 1 - tpRegress -
(tpDeath*tpExcessDeath)
  p_matrix["T2D", "Dead", "with DPP"] <- tpDeath*tpExcessDeath
  p_matrix["Dead", "Dead", "with DPP"] <- 1
 return(p_matrix)
}
###########
#state_q_matrix
state_q_matrix_cycle <- function(state_q_matrix, age)
{
  #uNorm
  uNorm_lookup <-
    c("[30,39]" = rbeta(1,5243.210127, 647.8882745),
      "[40,49]" = rbeta(1,3749.390288, 552.6603545),
      "[50,59]" = rbeta(1,4317.392142, 733.4582261),
      "[60,69]" = rbeta(1,3275.399881, 605.1330164),
      "[70,79]" = rbeta(1,2793.5369, 592.8093771),
      "[80,120]" = rbeta(1,1048.602386, 338.6611273))

  age_grpnorm <- cut(age, breaks= c(30,39,49,59,69,79,120))
  uNorm <-uNorm_lookup[age_grpnorm]

#uNDH
  uNDH_lookup <-
    c("[30,39]" = rbeta(1,18852.2607, 3585.449629),
      "[40,49]" = rbeta(1,40397.03381, 8321.148331),
      "[50,59]" = rbeta(1,79459.6746, 19951.65504),
      "[60,69]" = rbeta(1,137085.1892, 32219.07565),
      "[70,79]" = rbeta(1,164314.0338, 37615.0213),
      "[80,120]" = rbeta(1,53445.85673, 14946.63982))
  age_grpNDH <- cut(age, breaks= c(30,39,49,59,69,79,120))
  uNDH <-uNDH_lookup[age_grpNDH]

#uT2D
  uT2D_lookup <-
    c("[30,39]" = rbeta(1,27.8648312, 8.05807557),
      "[40,49]" = rbeta(1,48.96668398, 23.6572448),
      "[50,59]" = rbeta(1,158.9763077, 60.70760667),
      "[60,69]" = rbeta(1,238.2666204, 102.8036175),
      "[70,79]" = rbeta(1,294.0078638, 100.8414828),
      "[80,120]" = rbeta(1,198.9254422, 97.37988722))
  age_grpT2D <- cut(age, breaks= c(30,39,49,59,69,79,120))
  uT2D <-uT2D_lookup[age_grpT2D]

  state_q_matrix[1,1]<- uNorm
  state_q_matrix["without DPP", "NDH"] <- uNDH
  state_q_matrix["without DPP", "T2D"] <- uT2D
  state_q_matrix["without DPP", "Dead"] <-0
  state_q_matrix["with DPP", "Normal"] <- uNorm
  state_q_matrix["with DPP", "NDH"] <- uNDH
  state_q_matrix["with DPP", "T2D"] <-uT2D
  state_q_matrix["with DPP", "Dead"]<-0
  return(state_q_matrix)
}
## Run model ----
for(n in 1:n_trials) {
    for (i in 1:n_treatments) {
    age <- Initial_age
    for (j in 2:n_cycles) {
      p_matrix <- p_matrix_cycle(p_matrix, age, j - 1)
      pop[, cycle = j, treatment = i] <-
      pop[, cycle = j - 1, treatment = i] %*% p_matrix[, , treatment = i]
      trans[, cycle = j, treatment = i] <-
      pop[, cycle = j - 1, treatment = i] %*% (trans_c_matrix * p_matrix[, ,
treatment = i])

```

```
      halfpop["Normal", cycle=j, treatment=i]<- 0.5*(pop["Normal", cycle=(j-
1), treatment=i]+pop["Normal", cycle=j, treatment=i])
      halfpop["NDH", cycle=j, treatment=i]<- 0.5*(pop["NDH", cycle=(j-1),
treatment=i]+pop["NDH", cycle=j, treatment=i])
      halfpop["T2D", cycle=j, treatment=i]<- 0.5*(pop["T2D", cycle=(j-1),
treatment=i]+pop["T2D", cycle=j, treatment=i])
      halfpop["Dead", cycle=j, treatment=i]<- 0.5*(pop["Dead", cycle=(j-1),
treatment=i]+pop["Dead", cycle=j, treatment=i])
      cycle_qalys[i, ] <-
        (state_q_matrix[treatment = i, ] %*% halfpop[, , treatment = i]) *
(1/(1 + cDr))^(-1:(n_cycles-2))
      age <- age + 1
    }
    #Half cycle correction applied
    cycle_state_costs[i, ] <-
      (state_c_matrix[treatment = i, ] %*% halfpop[, , treatment = i]) *
(1/(1 + cDr))^(-1:(n_cycles-2))
   # discounting at _previous_ cycle
    cycle_costs[i, ] <- cycle_state_costs[i, ] #+ cycle_trans_costs[i, ]
    total_costs[i] <- sum(cycle_costs[treatment = i, -1])
    total_QALYs[i] <- sum(cycle_qalys[treatment = i, -1])
  }
  #adding in costs of DPP participation
  total_costs["with DPP"] <- total_costs["with DPP"] + dppcost_sum
  #adding in benefits from DPP participation
  #multipling by 0.5 as this is utility gain to convert to QALY
  total_QALYs["with DPP"] <- total_QALYs["with DPP"] + 0.5*utility_sum

  costs[n, ] <- total_costs
  qalys[n,] <- total_QALYs
```

```
}
agecat1 <- data.frame(costs, qalys)
names(agecat1)[names(agecat1) == "without.DPP.1"] <- 
"Without.DPP.qalys.AGECAT1"
names(agecat1)[names(agecat1) == "with.DPP.1"] <- "With.DPP.qalys.AGECAT1"
names(agecat1)[names(agecat1) == "without.DPP"] <- 
"Without.DPP.costs.AGECAT1"
names(agecat1)[names(agecat1) == "with.DPP"] <- "With.DPP.costs.AGECAT1"
writexl::write_xlsx(agecat1, "age category1 - 58.xlsx")
save(agecat1,file="agecat1.Rda")
#example population trace
popagecat1 <- data.frame(pop)
writexl::write_xlsx(popagecat1, "age category1 trace.xlsx")
save(popagecat1,file="agecat1trace.Rda")

########################
# Analysis of Results  #
########################
set.seed(9009)
load("agecat6.Rda")
load("agecat5.Rda")
load("agecat4.Rda")
load("agecat3.Rda")
load("agecat2.Rda")
load("agecat1.Rda")
#willingness to pay threshold
wtp <- 20000
#creating a common variable ID so they can be merged 'trials'
```

```
n_trials<-10000

trials<-c(seq(1, n_trials, by=1))

agecat1$trials <- trials

agecat2$trials <- trials

agecat3$trials <- trials

agecat4$trials <- trials

agecat5$trials <- trials

agecat6$trials <- trials

#merging the two data frames

age12 <- merge(agecat1, agecat2, by="trials")

age123 <- merge(age12, agecat3, by="trials")

age1234 <- merge(age123, agecat4, by="trials")

age12345 <- merge(age1234, agecat5, by="trials")

whole_cohort <- merge(age12345, agecat6, by="trials")

writexl::write_xlsx(whole_cohort, "whole cohort.xlsx")

save(whole_cohort,file="whole_cohort.Rda")

#create a variable that's the sum of all the rows

#create datasets for QALYs & COSTS according to DPP / without DPP

#sum across the rows - using rowSums

#create a dataset from this

#With DPP - costs

DPP_costs
=data.frame(whole_cohort$With.DPP.costs.AGECAT6,whole_cohort$With.DPP.costs.A
GECAT5,whole_cohort$With.DPP.costs.AGECAT4,whole_cohort$With.DPP.costs.AGECAT
3,whole_cohort$With.DPP.costs.AGECAT2,whole_cohort$With.DPP.costs.AGECAT1)

DPP_costs$total_costs=rowSums(DPP_costs)

#without DPP - costs

withoutDPP_costs
=data.frame(whole_cohort$Without.DPP.costs.AGECAT6,whole_cohort$Without.DPP.c
osts.AGECAT5,whole_cohort$Without.DPP.costs.AGECAT4,whole_cohort$Without.DPP.
costs.AGECAT3,whole_cohort$Without.DPP.costs.AGECAT2,whole_cohort$Without.DPP
.costs.AGECAT1)

withoutDPP_costs$total_costs=rowSums(withoutDPP_costs)

#With DPP - qalys

DPP_qalys
=data.frame(whole_cohort$With.DPP.qalys.AGECAT6,whole_cohort$With.DPP.qalys.A
GECAT5,whole_cohort$With.DPP.qalys.AGECAT4,whole_cohort$With.DPP.qalys.AGECAT
3,whole_cohort$With.DPP.qalys.AGECAT2,whole_cohort$With.DPP.qalys.AGECAT1)

DPP_qalys$total_qalys=rowSums(DPP_qalys)

#without DPP - qalys

withoutDPP_qalys
=data.frame(whole_cohort$Without.DPP.qalys.AGECAT6,whole_cohort$Without.DPP.q
alys.AGECAT5,whole_cohort$Without.DPP.qalys.AGECAT4,whole_cohort$Without.DPP.
qalys.AGECAT3,whole_cohort$Without.DPP.qalys.AGECAT2,whole_cohort$Without.DPP
.qalys.AGECAT1)

withoutDPP_qalys$total_qalys=rowSums(withoutDPP_qalys)

PSA_results =data.frame(DPP_costs$total_costs, withoutDPP_costs$total_costs,
DPP_qalys$total_qalys, withoutDPP_qalys$total_qalys)

writexl::write_xlsx(PSA_results, "PSA results from mixed age cohort.xlsx")

save(PSA_results,file="PSA_results.Rda")

#creating a variable that calculates the difference between costs / qalys

PSA_results$cost_dif <- PSA_results$DPP_costs.total_costs -
PSA_results$withoutDPP_costs.total_costs

PSA_results$qaly_dif <- PSA_results$DPP_qalys.total_qalys -
PSA_results$withoutDPP_qalys.total_qalys

#calculating the incremental net monetary benefit at different WTP thresholds

#

PSA_results$inmb0 <- ((PSA_results$qaly_dif * 0) - PSA_results$cost_dif)/1000

PSA_results$inmb1 <- ((PSA_results$qaly_dif * 1000) -
PSA_results$cost_dif)/1000

PSA_results$inmb2 <- ((PSA_results$qaly_dif * 2000) -
PSA_results$cost_dif)/1000

PSA_results$inmb3 <- ((PSA_results$qaly_dif * 3000) -
PSA_results$cost_dif)/1000
```

```
PSA_results$inmb4 <- ((PSA_results$qaly_dif * 4000) -
PSA_results$cost_dif)/1000

PSA_results$inmb5 <- ((PSA_results$qaly_dif * 5000) -
PSA_results$cost_dif)/1000

PSA_results$inmb6 <- ((PSA_results$qaly_dif * 6000) -
PSA_results$cost_dif)/1000

PSA_results$inmb7 <- ((PSA_results$qaly_dif * 7000) -
PSA_results$cost_dif)/1000

PSA_results$inmb8 <- ((PSA_results$qaly_dif * 8000) -
PSA_results$cost_dif)/1000

PSA_results$inmb9 <- ((PSA_results$qaly_dif * 9000) -
PSA_results$cost_dif)/1000

PSA_results$inmb10 <- ((PSA_results$qaly_dif * 10000) -
PSA_results$cost_dif)/1000

PSA_results$inmb11 <- ((PSA_results$qaly_dif * 11000) -
PSA_results$cost_dif)/1000

PSA_results$inmb12 <- ((PSA_results$qaly_dif * 12000) -
PSA_results$cost_dif)/1000

PSA_results$inmb13 <- ((PSA_results$qaly_dif * 13000) -
PSA_results$cost_dif)/1000

PSA_results$inmb14 <- ((PSA_results$qaly_dif * 14000) -
PSA_results$cost_dif)/1000

PSA_results$inmb15 <- ((PSA_results$qaly_dif * 15000) -
PSA_results$cost_dif)/1000

PSA_results$inmb16 <- ((PSA_results$qaly_dif * 16000) -
PSA_results$cost_dif)/1000

PSA_results$inmb17 <- ((PSA_results$qaly_dif * 17000) -
PSA_results$cost_dif)/1000

PSA_results$inmb18 <- ((PSA_results$qaly_dif * 18000) -
PSA_results$cost_dif)/1000

PSA_results$inmb19 <- ((PSA_results$qaly_dif * 19000) -
PSA_results$cost_dif)/1000

PSA_results$inmb20 <- ((PSA_results$qaly_dif * 20000) -
PSA_results$cost_dif)/1000

PSA_results$inmb21 <- ((PSA_results$qaly_dif * 21000) -
PSA_results$cost_dif)/1000

PSA_results$inmb22 <- ((PSA_results$qaly_dif * 22000) -
PSA_results$cost_dif)/1000
```

```
PSA_results$inmb23 <- ((PSA_results$qaly_dif * 23000) -
PSA_results$cost_dif)/1000

PSA_results$inmb24 <- ((PSA_results$qaly_dif * 24000) -
PSA_results$cost_dif)/1000

PSA_results$inmb25 <- ((PSA_results$qaly_dif * 25000) -
PSA_results$cost_dif)/1000

PSA_results$inmb26 <- ((PSA_results$qaly_dif * 26000) -
PSA_results$cost_dif)/1000

PSA_results$inmb27 <- ((PSA_results$qaly_dif * 27000) -
PSA_results$cost_dif)/1000

PSA_results$inmb28 <- ((PSA_results$qaly_dif * 28000) -
PSA_results$cost_dif)/1000

PSA_results$inmb29 <- ((PSA_results$qaly_dif * 29000) -
PSA_results$cost_dif)/1000

PSA_results$inmb30 <- ((PSA_results$qaly_dif * 30000) -
PSA_results$cost_dif)/1000

PSA_results$inmb31 <- ((PSA_results$qaly_dif * 31000) -
PSA_results$cost_dif)/1000

PSA_results$inmb32 <- ((PSA_results$qaly_dif * 32000) -
PSA_results$cost_dif)/1000

PSA_results$inmb33 <- ((PSA_results$qaly_dif * 33000) -
PSA_results$cost_dif)/1000

PSA_results$inmb34 <- ((PSA_results$qaly_dif * 34000) -
PSA_results$cost_dif)/1000

PSA_results$inmb35 <- ((PSA_results$qaly_dif * 35000) -
PSA_results$cost_dif)/1000

PSA_results$inmb36 <- ((PSA_results$qaly_dif * 36000) -
PSA_results$cost_dif)/1000

PSA_results$inmb37 <- ((PSA_results$qaly_dif * 37000) -
PSA_results$cost_dif)/1000

PSA_results$inmb38 <- ((PSA_results$qaly_dif * 38000) -
PSA_results$cost_dif)/1000

PSA_results$inmb39 <- ((PSA_results$qaly_dif * 39000) -
PSA_results$cost_dif)/1000

PSA_results$inmb40 <- ((PSA_results$qaly_dif * 40000) -
PSA_results$cost_dif)/1000

PSA_results$inmb41 <- ((PSA_results$qaly_dif * 41000) -
PSA_results$cost_dif)/1000
```

```
PSA_results$inmb42 <- ((PSA_results$qaly_dif * 42000) -
PSA_results$cost_dif)/1000

PSA_results$inmb43 <- ((PSA_results$qaly_dif * 43000) -
PSA_results$cost_dif)/1000

PSA_results$inmb44 <- ((PSA_results$qaly_dif * 44000) -
PSA_results$cost_dif)/1000

PSA_results$inmb45 <- ((PSA_results$qaly_dif * 45000) -
PSA_results$cost_dif)/1000

#calculating number of trials that are cost-effective given willingness to
pay

wtprange<-c(seq(0,45000,1000))

#per_ce<-c(rep(0,length(wtprange)))

#manually changed length to allow for additional CEAC estimates to ensure
there is not a 'kink' in the graph

per_ce<-c(rep(0,length(67)))

ceac<-data.frame(wtprange, per_ce)

ceac$per_ce[1] <- (sum(PSA_results$inmb0 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[2] <- (sum(PSA_results$inmb1 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[3] <- (sum(PSA_results$inmb2 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[4] <- (sum(PSA_results$inmb3 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[5] <- (sum(PSA_results$inmb4 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[6] <- (sum(PSA_results$inmb5 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[7] <- (sum(PSA_results$inmb6 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[8] <- (sum(PSA_results$inmb7 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[9] <- (sum(PSA_results$inmb8 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[10] <- (sum(PSA_results$inmb9 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[11] <- (sum(PSA_results$inmb10 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[12] <- (sum(PSA_results$inmb11 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[13] <- (sum(PSA_results$inmb12 >= 0, na.rm=TRUE)/n_trials)

ceac$per_ce[14] <- (sum(PSA_results$inmb13 >= 0, na.rm=TRUE)/n_trials)
```

```
ceac$per_ce[15] <- (sum(PSA_results$inmb14 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[16] <- (sum(PSA_results$inmb15 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[17] <- (sum(PSA_results$inmb16 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[18] <- (sum(PSA_results$inmb17 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[19] <- (sum(PSA_results$inmb18 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[20] <- (sum(PSA_results$inmb19 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[21] <- (sum(PSA_results$inmb20 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[22] <- (sum(PSA_results$inmb21 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[23] <- (sum(PSA_results$inmb22 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[24] <- (sum(PSA_results$inmb23 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[25] <- (sum(PSA_results$inmb24 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[26] <- (sum(PSA_results$inmb25 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[27] <- (sum(PSA_results$inmb26 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[28] <- (sum(PSA_results$inmb27 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[29] <- (sum(PSA_results$inmb28 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[30] <- (sum(PSA_results$inmb29 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[31] <- (sum(PSA_results$inmb30 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[32] <- (sum(PSA_results$inmb31 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[33] <- (sum(PSA_results$inmb32 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[34] <- (sum(PSA_results$inmb33 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[35] <- (sum(PSA_results$inmb34 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[36] <- (sum(PSA_results$inmb35 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[37] <- (sum(PSA_results$inmb36 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[38] <- (sum(PSA_results$inmb37 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[39] <- (sum(PSA_results$inmb38 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[40] <- (sum(PSA_results$inmb39 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[41] <- (sum(PSA_results$inmb40 >= 0, na.rm=TRUE)/n_trials)
```

```
ceac$per_ce[42] <- (sum(PSA_results$inmb41 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[43] <- (sum(PSA_results$inmb42 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[44] <- (sum(PSA_results$inmb43 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[45] <- (sum(PSA_results$inmb44 >= 0, na.rm=TRUE)/n_trials)
ceac$per_ce[46] <- (sum(PSA_results$inmb45 >= 0, na.rm=TRUE)/n_trials)
par(mar=c(1,1,1,1))
#par("mar") 5.1 4.1 4.1 2.1
plot(ceac$wtprange,ceac$per_ce, type="l", ylim=c(0,1), col="black", lwd=2, xlab="Willingness to Pay (£)", ylab="Probability Cost-effective", main="Cost-effectiveness Acceptability Curve")
#export to excel
writexl::write_xlsx(ceac, "ceac.xlsx")
percent_CE_20 <- (sum(PSA_results$inmb20 >= 0, na.rm=TRUE)/n_trials)*100
percent_CE_30 <- (sum(PSA_results$inmb30 >= 0, na.rm=TRUE)/n_trials)*100
NE <-sum(PSA_results$cost_dif>= 0 & PSA_results$qaly_dif>= 0, na.rm=TRUE)
NW <- sum(PSA_results$cost_dif>= 0 & PSA_results$qaly_dif<= 0, na.rm=TRUE)
SE <- sum(PSA_results$cost_dif<= 0 & PSA_results$qaly_dif>= 0, na.rm=TRUE)
SW <- sum(PSA_results$cost_dif<= 0 & PSA_results$qaly_dif<= 0, na.rm=TRUE)
PSA_results$qaly_difpp <- (PSA_results$qaly_dif)/1000
PSA_results$cost_difpp <- (PSA_results$cost_dif)/1000
#exporting the results
writexl::write_xlsx(PSA_results, "incremental_nmb.xlsx")
#plotting per person incremental
#find out range of x variable to ensure plot is correct
range(PSA_results$qaly_difpp)
#-0.02891131  0.10872687
ceacplot <- ggplot(data=PSA_results,aes(y=cost_difpp))+
  geom_point(aes(x=qaly_difpp),color="blue")+
  geom_vline(xintercept=0,color="grey", size=1)+
  geom_hline(yintercept=0,color="grey", size=1)+
  xlim(-0.03,0.2)+
  xlab("Incremental QALYs")+
  ylab("Incremental Costs")+
  ggtitle("")+
  theme(plot.title = element_text(hjust = 0.5))
ceacplot + geom_abline(intercept = 0, slope = 20000, color="black",
linetype="solid", size=1)
#Average costs:
mean(DPP_costs$total_costs)
#31862639
mean(withoutDPP_costs$total_costs)
#31998394
#Average QALYs:
mean(DPP_qalys$total_qalys)
#10806.61
mean(withoutDPP_qalys$total_qalys)
#10765.8
#Average incremental net benefits
#20,000 WTP
mean(PSA_results$inmb20)
#951.9416
#30,000 WTP
mean(PSA_results$inmb30)
#1360.035
#Number of monte carlo simulations that are cost-effective
ceac$per_ce[21]
```

```
#98.08%
ceac$per_ce[31]
#98.42%
```

Table A4.7: Completed CHEERS (2022) checklist.

	Item	Guidance for Reporting	Reported in section
TITLE			
Title	1	Identify the study as an economic evaluation and specify the interventions being compared	Title (manuscript version)
ABSTRACT			
Abstract	2	Provide a structured summary that highlights context, key methods, results and alternative analyses.	Abstract (manuscript version)
INTRODUCTION			
Background and objectives	3	Give the context for the study, the study question and its practical relevance for decision making in policy or practice.	4.1. Introduction, 4.2.4. Context for model development
METHODS			
Health economic analysis plan	4	Indicate whether a health economic analysis plan was developed and where available.	Availability of data and materials
Study population	5	Describe characteristics of the study population (such as age range, demographics, socioeconomic, or clinical characteristics).	4.2.5.1. Interventions Analysed, 4.2.5.3. Model Parameters
Setting and location	6	Provide relevant contextual information that may influence findings.	4.1. Introduction, 4.2.4. Context for model development
Comparators	7	Describe the interventions or strategies being compared and why chosen.	4.2.5.1. Interventions Analysed
Perspective	8	State the perspective(s) adopted by the study and why chosen.	4.2. Methods
Time horizon	9	State the time horizon for the study and why appropriate.	4.2.5.6. Cost-effectiveness Analysis
Discount rate	10	Report the discount rate(s) and reason chosen.	2.5 NHS DPP Effectiveness

Table A4.7: Completed CHEERS (2022) checklist. *(Continued)*

	Item	Guidance for Reporting	Reported in section
Selection of outcomes	11	Describe what outcomes were used as the measure(s) of benefit(s) and harm(s).	4.2.5.4. Costs & Outcomes
Measurement of outcomes	12	Describe how outcomes used to capture benefit(s) and harm(s) were measured.	4.2.5.4. Costs & Outcomes
Valuation of outcomes	13	Describe the population and methods used to measure and value outcomes.	4.2.5.4. Costs & Outcomes
Measurement and valuation of resources and costs	14	Describe how costs were valued.	4.2.5.4. Costs & Outcomes
Currency, price date, and conversion	15	Report the dates of the estimated resource quantities and unit costs, plus the currency and year of conversion.	4.2.5.4. Costs & Outcomes
Rationale and description of model	16	If modelling is used, describe in detail and why used. Report if the model is publicly available and where it can be accessed.	4.2. Methods Model code available from Github
Analytics and assumptions	17	Describe any methods for analysing or statistically transforming data, any extrapolation methods, and approaches for validating any model used.	4.2.5.3. Model Parameters, Supplementary Material
Characterizing heterogeneity	18	Describe any methods used for estimating how the results of the study vary for sub-groups.	Not applicable
Characterizing distributional effects	19	Describe how impacts are distributed across different individuals or adjustments made to reflect priority populations.	Not applicable
Characterizing uncertainty	20	Describe methods to characterize any sources of uncertainty in the analysis.	4.2.5.8. Sensitivity Analyses

Table A4.7: Completed CHEERS (2022) checklist. *(Continued)*

	Item	Guidance for Reporting	Reported in section
Approach to engagement with patients and others affected by the study	21	Describe any approaches to engage patients or service recipients, the general public, communities, or stakeholders (e.g., clinicians or payers) in the design of the study.	Availability of data and materials: wider project protocol (manuscript version)
RESULTS			
Study parameters	22	Report all analytic inputs (e.g., values, ranges, references) including uncertainty or distributional assumptions.	Table 4.2
Summary of main results	23	Report the mean values for the main categories of costs and outcomes of interest and summarise them in the most appropriate overall measure.	Table 4.3, Table 4.4
Effect of uncertainty	24	Describe how uncertainty about analytic judgments, inputs, or projections affect findings. Report the effect of choice of discount rate and time horizon, if applicable.	Table 4.6 4.3.1.2. Sensitivity Analyses
Effect of engagement with patients and others affected by the study	25	Report on any difference patient/service recipient, general public, community, or stakeholder involvement made to the approach or findings of the study.	Not reported
DISCUSSION			
Study findings, limitations, generalizability, and current knowledge	26	Report key findings, limitations, ethical or equity considerations not captured, and how these could impact patients, policy, or practice.	4.4.1. Strengths and limitations
Source of funding	27	Describe how the study was funded and any role of the funder in the identification, design, conduct, and reporting of the analysis	Author's declaration Funding statement (manuscript version)
Conflicts of interest	28	Report authors conflicts of interest according to journal or International Committee of Medical Journal Editors requirements.	Conflicts of Interest (manuscript version)

Journal permissions

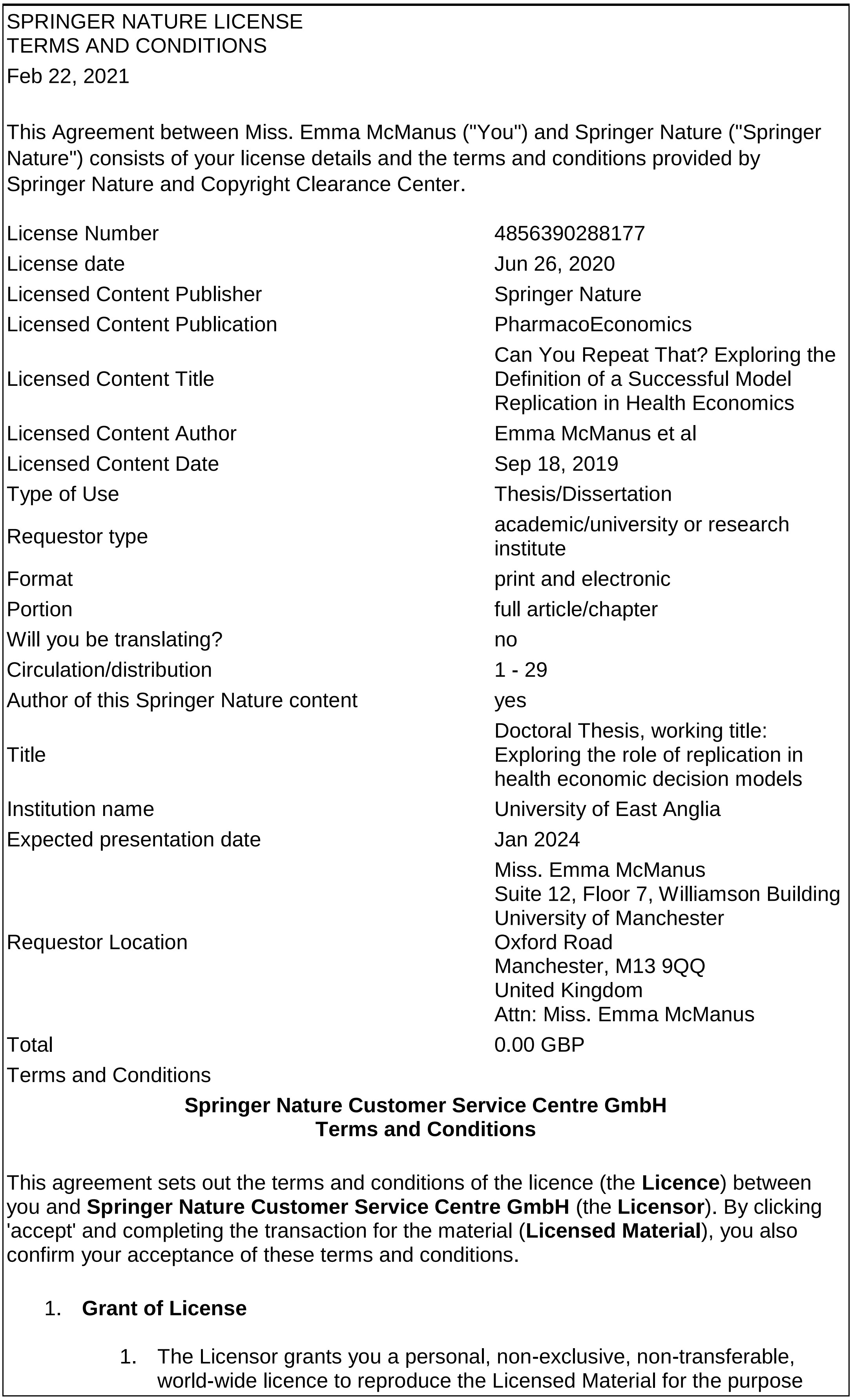
Chapter 1:

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Jul 06, 2021

This Agreement between Miss. Emma McManus ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number 5103100089480
License date Jul 06, 2021
Licensed Content Publisher John Wiley and Sons
Licensed Content Publication Journal of the European Academy of Dermatology and Venereology Licensed Content Title An introduction to the methods of decision-analytic modelling used in economic evaluations for Dermatologists
Licensed Content Author N.J. Levell, T.H. Sach, E. McManus
Licensed Content Date Jun 27, 2019 Licensed Content Volume 33
Licensed Content Issue 10
Licensed Content Pages 8 Type of use Dissertation/Thesis
Requestor type Author of this Wiley article
Format Print and electronic
Portion Text extract
Number of Pages 2
Will you be translating? No
Title Doctoral Thesis, working title: Exploring the role and value of replication in decision models Institution name University of East Anglia
Expected presentation date Jan 2024
Portions Sections describing modelling methods
Requestor Location Miss. Emma McManus Suite 12, Floor 7, Williamson Building University of Manchester Oxford Road Manchester, M13 9QQ United Kingdom Attn: Miss. Emma McManus Publisher Tax ID EU826007151
Total 0.00 GBP

Chapter 2:



SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Feb 22, 2021

This Agreement between Miss. Emma McManus ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4856390288177
License date	Jun 26, 2020
Licensed Content Publisher	Springer Nature
Licensed Content Publication	PharmacoEconomics
Licensed Content Title	Can You Repeat That? Exploring the Definition of a Successful Model Replication in Health Economics
Licensed Content Author	Emma McManus et al
Licensed Content Date	Sep 18, 2019
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title	Doctoral Thesis, working title: Exploring the role of replication in health economic decision models
Institution name	University of East Anglia
Expected presentation date	Jan 2024
Requestor Location	Miss. Emma McManus Suite 12, Floor 7, Williamson Building University of Manchester Oxford Road Manchester, M13 9QQ United Kingdom Attn: Miss. Emma McManus
Total	0.00 GBP

Terms and Conditions

Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

1. **Grant of License**

    1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose

specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

3. If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. **Scope of Licence**

1. You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

2. A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

3. Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

4. Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

5. An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

□ **Duration of Licence**

A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

Scope of Licence	**Duration of Licence**
Post on a website	12 months
Presentations	12 months
Books and journals	Lifetime of the edition in the language purchased

□ Acknowledgement

The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

□ Restrictions **on use**

Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

1. You must not use any Licensed Material as part of any design or trademark.

2. Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

□ **Ownership of Rights**

1. Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

□ **Warranty**

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

□ **Limitations**

BOOKS ONLY: Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

□ **Termination and Cancellation**

Licences will expire after the period shown in Clause 3 (above).

1. Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

Appendix 1 — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT]** (year of publication)

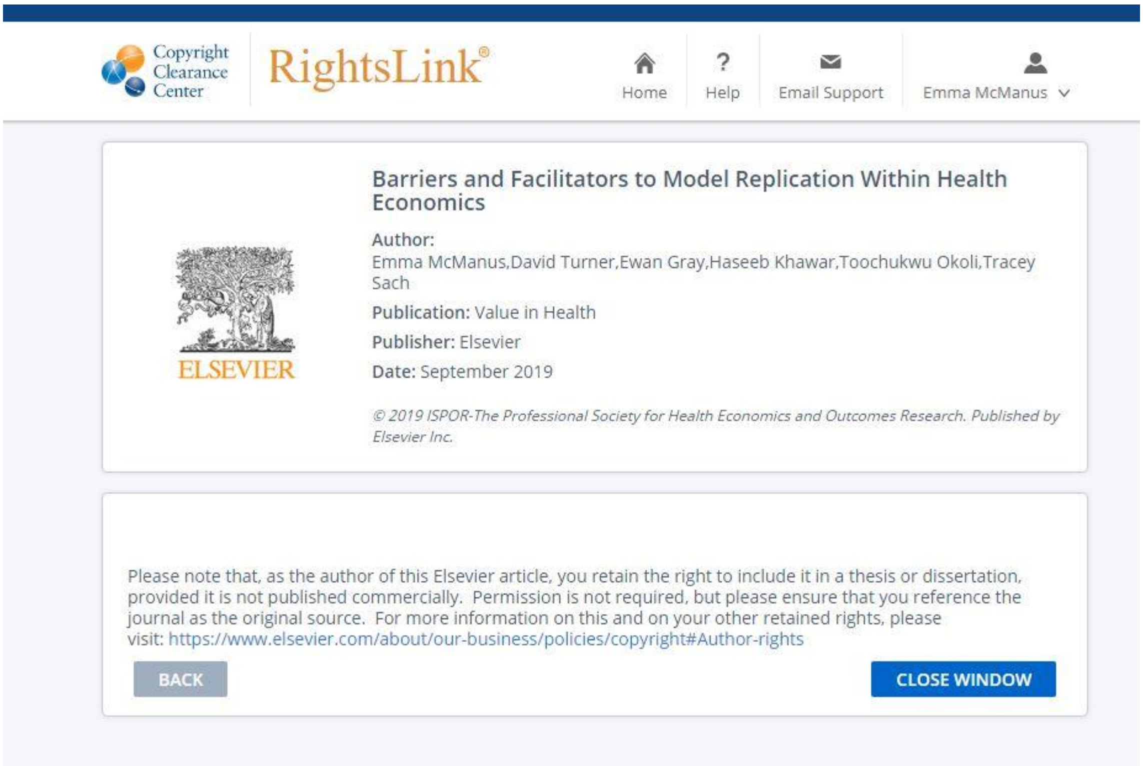
For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.2

Chapter 3:



Barriers and Facilitators to Model Replication Within Health Economics

Author:
Emma McManus,David Turner,Ewan Gray,Haseeb Khawar,Toochukwu Okoli,Tracey Sach

Publication: Value in Health

Publisher: Elsevier

Date: September 2019

© 2019 ISPOR-The Professional Society for Health Economics and Outcomes Research. Published by Elsevier Inc.

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights