Author:
Moreno-Stokoe, Chris

Title:
Visual and gamified approaches to understanding complex causal networks in human health

Visual and gamified approaches to understanding complex causal networks in human health

Chris Moreno-Stokoe

School of Psychological Sciences

University of Bristol

January 2023

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences.

Word count: 64,545

Abstract

Recent advances allow researchers to study the complex relationships between hundreds or thousands of health factors in a single study but interpreting their results is threatened by high levels of network complexity. Previous research demonstrates that network data complexities can be understood by developing technology to visualise, simulate and gamify it.

In this thesis I present research developing ways of understanding the complexity in the network of factors related to human health. First, I introduce and analyse the complex relationship between psychological and physical health factors, particularly wellbeing and insomnia. Using a technique called Mendelian Randomisation (MR) I test the causal pathways between sleep and wellbeing (chapter 2), and then use network MR to conceptualise the wide-ranging potential factors that influence the network of effects between sleep and wellbeing (chapter 3). The resulting network dataset provides the foundation of the second part of the thesis, where I explore and test different methods to help researchers better understand the complexity in network MR datasets. Chapter 4 explores the development and use of causal network visualisation and chapter 5 describes the creation of a data game that allows participants to interact with the data and see the predicted impact of intervening on different nodes of the network. In the final empirical chapter (chapter 6), I build on previous experiments, and investigate whether the inclusion of game features results in greater participant understanding of the complexity of the data compared to using an interactive visualisation control.

My findings indicate that physical and mental health factors exist as part of large and complex network structures and that researchers can better understand these to some degree with visualisations, and perhaps to a greater extent with interactive and game mediums.

Acknowledgements

Firstly, I would like to thank my family including my brother, my dad and my mum (to whom I dedicate this work). You have supported me in every way over the last years and this would not have been possible without you.

Second, my supervisors. Claire and Oliver, you have been fantastic; your standards for excellence have shaped part of who I am today, and I owe that to you both. We have been on a journey together and I couldn't be happier about it, what I've learned, and what we have managed to accomplish.

Third, my friends in Bristol and London. From everyone in the Dynamic Genetics Lab group including Jess, Nina, Benji, Jacks, and Valerio, to the School of Psychological Science, including Anca who kept me sane through the pandemic, and the volunteers who gave up their time to play a space game about MR- Thank you! Lastly, I would like to thank my current colleagues and managers at PwC for creating the most positive and supportive atmosphere that allowed me to finish this thesis alongside work.

Author declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed:  **Date: 14th January 2023**

Table of contents

Contents

List of tables

List of figures

Table of appendices

1 Introduction

1.1 Thesis motivation

My motivation for this thesis was to design ways of understanding the complexity of the network of human health. My education is in experimental Psychology and my previous work is in designing patient-facing interfaces, so I have affinity for approaches combining conventional statistics with visual and interactive media. The endpoint in this thesis is developing a playable simulation of public health interventions. The notion of understanding complexity through play is particularly interesting to me because I am a fan of videogames and, as a child of the 90s, I grew up playing so-called "educational games". Some were good, some were not so good, but they were memorable. I remember playing a game as a class in school, working together to identify solutions and solve problems. It struck me then that one of the powerful aspects of play is encouraging people to engage with a topic, and my experience in this thesis has confirmed that to me. The influence of the educational game craze of the 90s and 00s can be felt in the present day. These educational games have even inspired me to design some of my own, for communicating serious issues in health to the public and policy makers. However, developing games proves to be difficult and time intensive. So, does the educational potential of games justify this effort?

1.2 Thesis overview

In this thesis I present research developing ways of understanding the complexity in the network of factors related to human health.

In the first half of this thesis, I introduce and analyse the real-world complexity that exists between psychological and physical health factors, particularly wellbeing and sleep, and obtain a network dataset describing the relationships between them. In my first two chapters, Chapters 2 and 3, I introduce "Mendelian randomisation" (MR) and extend causal research methods in a network analysis approach—which is often theorised but less often applied—to make causal inferences. A

key aspect of this thesis is the use of "network MR" which is a new and fast-evolving method of probing the network complexity in the network of human health. The resulting network dataset builds on precursor projects to develop the framework for network MR analysis (Hemani, Bowden, et al., 2017) and to start transforming this data into interactive experiences to help understand its complexity (Free Ice Cream & Davis, 2018).

In the second half of this thesis, I experiment with developing new technology to help researchers better understand the complexity in network MR datasets. I made several contributions to MR software allowing researcher to visualise network datasets in a novel web tool (Chapter 4, www.morenostok.io/mirana), simulate and gamify their data (Chapter 5, https://github.com/CMorenoStokoe/mr-game-webapp), and systematically reviewing available software to build a more comprehensive list of software available for MR researchers (www.morenostok.io/mrsoftwarelist.html). By publishing software along with the open-source code I have allowed researchers to modify, improve and expand upon my software in future research projects. In Chapter 6, I demonstrate how my simulation and game software can be used to perform an experimental study and found that an MR data game successfully promoted longer engagement and facilitated learning of the underlying dataset. This game has since been played by a wide range of players, including attendees at the research without borders festival (Bristol, 2020), and is featured on the MRC IEU public engagement website (http://www.bristol.ac.uk/integrative-epidemiology/engagement/#dropdown-heading0-3).

1.3 Epidemiology

Understanding complexity in the network of human health is part of the broader discipline of epidemiology. Epidemiology is "the art and science of preventing disease, prolonging life and promoting health" (World Health Organisation, 2021) and investigates the occurrence, distribution, effects, and determinants of disease. Research in epidemiology forms an important evidence base which is used to understand how diseases work, who might be affected by ill health, and what can

be done to prevent and treat disease (Burgess & Thompson, 2015; Webb et al., 2020). In this section, I will introduce what is meant by illness, what epidemiologists aim to understand about it, and what the domains and methods of epidemiology are. I will close this section by explaining what causal research methods are and why I will use them in this thesis to identify the causes of a variety of physical and psychological health factors.

1.3.1 What is illness?

In this thesis I adopt broad definitions of health and ill health intended to capture a range of factors related to our physical and psychological health. The World Health Organisation defines health as "a state of complete physical, mental and social wellbeing and not merely the absence of disease" (www.who.int/about/governance/constitution). I define "illness" as the opposite of "health". Illness is a factor, or combination of factors, which detriment one's experience of life in any domain. I do not restrict my investigation of illness to diseases with biological origins, such as heart disease, and include factors which affect our cognitive processes, such as negative thought patterns in depression, and our ability to live a happy and fulfilled life.

1.3.2 The aims of epidemiology

Epidemiology has three main aims (Webb et al., 2020) and these can be understood as three related streams of work:

> **Observational epidemiology** seeks to measure the frequency of disease, along with describing its associated symptoms and risk factors. It is important to identify the risk factors correlated with susceptibility to disease, so that they can be informed, monitored by healthcare services, and prevention programs may be designed for avoiding a risk factor.
>
> **Causal epidemiology** seeks to understand the causes of disease. While risk factors may indicate who is at risk of a disease, it may not tell epidemiologists how this factor produces

the disease, or how to treat it. Causal epidemiologists use a class of causal research methods that are able to obtain stronger evidence for causal pathways than observational research.

Public health epidemiology applies observational and causal knowledge to design, evaluate and implement preventative measures (e.g., vaccination) and treatments (e.g., medicine, therapy) on a society-wide scale.

This thesis focusses on data from causal epidemiology, but all three streams contribute to healthcare by providing an evidence base for improving health across a variety of domains, from cancer research to understanding loneliness.

1.3.3 Domains of epidemiology

Epidemiology is a broad discipline encompassing many areas of work (Webb et al., 2020). Some researchers focus on one type of disease, such as cancer, and others focus on one type of cause, such as metabolomics. Some causes have large and clear effects, such as the main symptoms of viral infections, while others have small effects which are harder to identify, such as genetic vulnerabilities to common, complex disease. Detailed investigations of a small number of individuals can help identify how, whether and when factors produce disease at an individual level, while population-level investigations involving many individuals can identify trends which exist in the wider population. Different methods are often combined in multi-disciplinary studies to provide stronger evidence through triangulation (Munafò et al., 2021). I will focus on using methods of genetic epidemiology to gain population-level insights into the factors which influence our wellbeing and mental health.

It is increasingly recognised that wellbeing and mental health have large impacts on society. Poor wellbeing and mental illness are not traditional diseases caused by biological pathogens like bacteria, nor are they communicable in the usual sense, but they have profound impacts on an individuals' everyday life consistent with definitions of ill health (World Health Organisation, 2021). Mental illness accounts for 21% of the total burden of disease in England with an estimated yearly

cost to the economy of £105 billion (Public Health England, 2020). Recently there has been a particular increase in focus on mental health, both in terms of public recognition and research funding. For example, the charity YoungMinds (https://www.youngminds.org.uk/) conducts observational epidemiology, understanding the prevalence of wellbeing and mental health issues among young people in the UK, and several major funders have launched initiatives to address what has been described as a "mental health crisis".

Genetic epidemiology seeks to understand genetic influences on disease. Some variations in our genetic material, known as alleles, predispose us to certain diseases, or protect us from them. For example, a genetic variation common in sub-Saharan Africa alters the shape of red blood cells and this has a protective effect against Malaria, but introduces a risk of sickle cell anaemia (Luzzatto, 2012). This knowledge is valuable to epidemiologists because it helps identify sub-Saharan ancestry as a protective factor for malaria and risk factor for sickle cell anaemia. The human genome contains approximately 3 billion base pairs of DNA containing a sequence of nucleotides, either Adenine and Thymine (A-T), or Cytosine and Guanine (C-G), and even a single nucleotide change (polymorphism) can be associated with disorder. For example, sickle cell disease is caused by a single nucleotide polymorphism in a gene involved in producing red blood cells (*HBB*)(Tozatto-Maio et al., 2020). However, large genome-wide association studies with hundreds of thousands of human participants have shown that common, complex disorders such as cardiovascular disease or mental health difficulties are much more likely to be influenced by large numbers of variants, each of very small but additive effect (Jansen et al., 2019; Okbay et al., 2016b). Beyond their direct influence on disease, these polymorphisms can be used as a tool for understanding other, non-genetic causes.

1.3.4 Causality in epidemiology

In epidemiology a "cause" is a factor which produces disease (Webb et al., 2020). A disease could have a single cause, such as a virus, or it may be produced by a combination of causes (some known and some unknown), such as obesity caused by overeating and a lack of exercise. It is relatively easy

to identify which factors are correlated or associated with a disease, but identifying whether, and how, they truly cause a disease is more difficult. Without clear understanding of the true cause of disease, putative treatments can be ineffective or even harm instead of heal. For example, hormone-replacement therapy was once recommended as a treatment for breast cancer on the basis of observational evidence which associated it with reduced mortality but in trials it did not reduce mortality because the observational evidence did not indicate a casual relationship (Burgess & Thompson, 2015). Establishing causation is a key challenge in epidemiology and this is in part due to the difficulties in identifying factors which might confound an association.

When researching the causes of disease, the terms “exposure” and “outcome” are often used; a causal relationship is expressed as the effect which a risk factor, an exposure, has in producing a disease, an outcome (Figure 1.1).

Figure 1.1 Causal epidemiology aims to identify the causal effects of exposures (X) on outcomes (Y)

Confounding (Webb et al., 2020) occurs when the relationship between an exposure and outcome is, at least in part, due to a third factor associated with both exposure and outcome (Figure 1.2). This is a mixing of effects where the exposure is mixed up with another factor and this makes interpreting causality difficult.

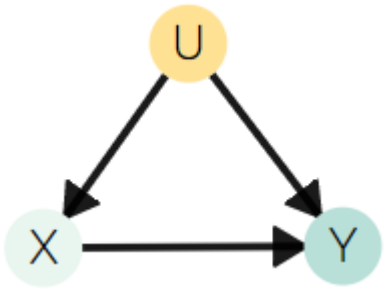


Figure 1.2 It is important for causal epidemiologists to be wary of confounding factors (U) since they make it more difficult to interpret the true effect of exposures (X) on outcomes (Y). Confounded relationships induce an association by acting on both exposure and outcome.

The consequences of confounding are that it can cause overestimation or underestimation of the size of a real effect, or completely hide the association. A confounding relationship can even reverse the direction of the effect to make it appear that a factor protects against a disease when it really causes it. For example, a high caloric intake is sometimes associated with a reduced risk of heart disease. The true causal relationship is that lower caloric intake protects against heart disease, but the reverse relationship is seen when exercise confounds it by acting on both factors, increasing caloric intake and reducing heart disease independently (Burgess & Thompson, 2015). It is therefore important to use methods which go beyond observing associations and test for causal relationships that are not produced by confounding.

1.3.5 Observational research methods

Observational research methods are well established in epidemiology and provide epidemiologists with descriptions of the incidence, prevalence and risk factors associated with disease. This information helps identify individuals who are at risk of a disease, would benefit from preventative measures and may require future treatment, and foreknowledge may improve outcomes. Furthermore, the observation that certain factors are associated with a disease can provide an indication that they may contribute to producing it.

Observational methods involve calculating the frequency with which a disease occurs (Webb et al., 2020), often described in terms of incidence and prevalence. Incidence describes the rate at which a disease occurs in a population, for example during the 2020-22 COVID-19 pandemic the UK Government published the number of new cases daily (https://coronavirus.data.gov.uk). Prevalence describes the number of individuals who currently have or have previously had a disease over a given time period. For example, the number of individuals who caught COVID over a year.

Observational methods also involve identifying the factors associated with the risk of a disease (Webb et al., 2020), measured using odds ratios. Odds ratios describe the relative increase in risk of disease that a risk factor conveys and are calculated as the incidence of disease in an at-risk population over the incidence of disease among the general population.

The advantage of observational research is that it is a practical approach to describing disease. Observational research methods are well established and their use dates back far beyond the modern era of epidemiology. Additionally, they can be simple to conduct. For example, during the Crimean war (c. 1850), Florence Nightingale convinced the British army of a severe disease epidemic by simply counting and recording the number of ill soldiers per day (Magnello, 2012). Such methods are typically inexpensive and relatively quick (Webb et al., 2020).

The disadvantage of observational research is that although it measures the association between an exposure and outcome, this does not necessarily represent a causal relationship. Factors such as confounding prevent causal inference. An association between exposure and outcome can appear for several reasons including confounding, as well as simple random chance. Observational epidemiology therefore helps provide evidence that factors are associated with disease but causal research methods are required to understand exactly how they are associated.

1.3.6 Causal research methods

Causal research methods complement observational methods and can provide strong evidence that an exposure is instrumental in causing an outcome (Burgess & Thompson, 2015). Causal methods

investigate the effect of exposure on outcome as closely as possible in order to obtain stronger evidence for causal relationships.

The strength of evidence for causality can be assessed using criteria such as the six Bradford-Hill criteria (Hill, 1965). First, a cause must happen before its effect (#1). Next, associations which are strong (#2), consistent (#3), and specific (#4) to the factors of interest present stronger evidence. Furthermore, an association where the level of the outcome is directly related to the level of the exposure is a better indication of cause and effect (dose-response relationship, #5). Finally, plausible effects are more likely (#6). Meeting these criteria does not provide a guarantee of causality, but they can be used by researchers as guidelines to interpret the strength of evidence in various domains demonstrating a plausible causal relationship (Fedak et al., 2015; Ioannidis, 2016). No method can eliminate all possible sources of bias, so often different methods are combined to approach an issue from multiple angles and compare results in a process of triangulation (Munafò et al., 2021).

Randomised control trials (RCTs) are often regarded as the gold standard of evidence for causality (Kaptchuk, 2001). RCTs involve comparing outcomes between two groups who either received an exposure or did not. Strength of evidence is ensured by making a direct manipulation of the exposure, so changes in the outcome are less likely to be unrelated to the changes in the exposure. Furthermore, individuals who receive the exposure are selected to be similar and are treated identically to those who do not receive the exposure so it is less likely that changes in the outcome are due to differences between groups. Similarly, individuals are allocated to groups randomly so there is no systematic bias when selecting groups. Through these methods RCTs are able to provide strong evidence for causality which meets many of the Bradford Hill criteria (Bradford Hill, 1965). For example, manipulation of the exposure occurs prior to changes in the outcome (temporality), and are less likely to be related to changes in other factors (specificity), and it can be assessed whether changing levels in the exposure result in similar changes in the outcome (dose-response

relationship). However, it is not always possible to conduct RCTs since they are expensive, time intensive and it is not always possible, or ethical, to manipulate the exposure of interest (Burgess & Thompson, 2015; Webb et al., 2020). Therefore, other methods are used to help build evidence about the causes of disease.

Instrumental variable (IV) analysis is an inexpensive and practical alternative to RCTs. An instrumental variable is a measurable factor assumed to have a causal influence on an exposure, but no direct influence on the outcome, other than through the exposure (Burgess & Thompson, 2015). The effect of an exposure can be measured by observing the effect that different levels of an instrumental variable have on an outcome (Figure 1.3). For example, the tobacco tax can be used as an instrument for smoking, since higher taxes reduce the prevalence of smoking, and lower taxes increase its prevalence. Whereas RCTs make a direct manipulation of the exposure, IV analyses use a pseudo-experimental process of observing the effects of a factor which manipulates the exposure of interest. This is both its advantage and disadvantage since direct randomisation of the exposure strengthens RCTs' ability to demonstrate causality, but it is not always practical or ethical manipulate exposures. For example, it would be unethical to manipulate smoking as this would cause harm. Any factor which causes the exposure can be used as an instrument in instrumental variable analysis and analyses using genetic instrumental variables are a fast-evolving area of research.

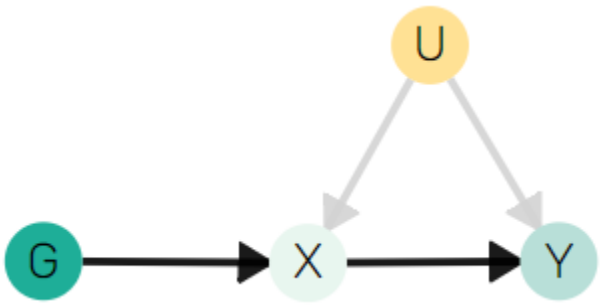


Figure 1.3 Instrumental variable analyses are a practical alternative to RCTs which observe the effects of an instrument (G) which has a causal effect on an exposure. The effect of exposure (X) on outcome (Y) can be

measured by comparing the outcome at different levels of the instrument, independent of unmeasured confounder U.

In the next chapters, two and three, I will introduce a type of instrumental variable analysis known as Mendelian randomisation (MR)(Davey Smith & Hemani, 2014b) and show how it can be used to investigate the causal relationships between factors in the network of human health. MR uses genetic variants as instruments. This is possible because our genetics predisposes us to various exposures, such as disease, and Genome-Wide Association Studies (GWAS) can be used to identify which genetic variants are associated with higher or lower levels of an exposure. For example, one GWAS (Driscoll et al., 2008) identified single nucleotide polymorphism genetic variants which are associated with an increased likelihood to smoke. These genetic variants can then be used as instruments. The use of genetic variants in an IV framework is particularly beneficial for two reasons. First, the random allocation of risk alleles in the population at conception means that individuals are randomly assigned to exposure and control groups. Second, our genetic make-up cannot be changed by environmental exposures, so it is not possible that the exposure directly affects the instrument. Taken together, these factors improve strength of evidence since it mitigates some traditional problems like confounding and reverse-causality. Furthermore, the process of conducting MR is now very practical since there is a wide range of freely available GWAS datasets that have been summarised ready for MR on sites such as MR Base (www.mrbase.org). However, researchers must be sure that genetic instruments are not just associated with an exposure but are instrumental in causing it. Ascertaining the causal influence of single nucleotide polymorphisms on an exposure is difficult, so researchers use a variety of methods to ensure instrumental variables meet criteria for valid causal inference (Haycock et al., 2016). MR is therefore a method of causal research which is complementary to traditional methods and comes with its own opportunities and challenges (Burgess & Thompson, 2015; Webb et al., 2020).

1.4 The network of human health

The field of human health is well placed to benefit from these developments in causal analysis. I will investigate how human health is a network comprising many factors which have complex effects on each other. In this section I will introduce the concept of network complexity, explain how this relates to our understanding of health, and outline recent developments in modelling health as a network.

1.4.1 Networks

At a basic level, a network is a collection of inter-related factors (Lima, 2013). Factors in a network are often referred to as "nodes" and the relationships they share with each other can be direct from one factor to another, or an indirect pathway comprising multiple steps. Networks make it more difficult to predict the effects of factors because they are mixed up with many other effects (Figure 1.1). Different relationships can have additive effects on a single factor, or negate each other's effects, as well as complex interactions including multiplication and division effects.

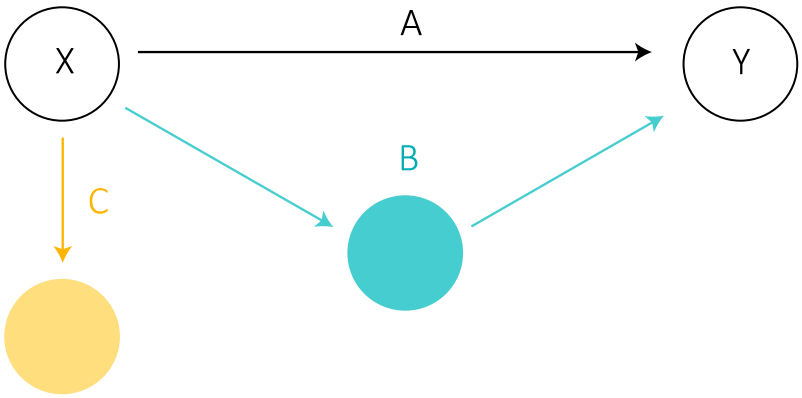


Figure 1.4 Network effects make it more difficult to understand the effects of a factor, or exposure (X), on an outcome (Y). Understanding the effect is complicated because in this example there is both a direct effect (A) and an effect mediated through a third variable (B), along with side-effects (C).

The concept of a network arose as a by-product of our growing scientific knowledge and pursuing increasingly complex research questions that involve more and more factors that inevitably relate to each other (Weaver, 1948). Networks can exist in many forms, but our modern-day concept of

networks was first described by French philosophers Gilles Deleuze and Félix Guattri (Deleuze & Guattri, 1972) in their book *Capitalism and Schizophrenia*. They describe the concept of a "rhizome" by analogy with a root network. A rhizome is a system where factors share relationships with each other in a decentralised manner. In a rhizomatic system complexity emerges in a non-hierarchical manner, through the natural inter-relations between nodes. Figure 1.14 above is an example of a rhizomic network comprised of individual nodes which have formed relationships with each other.

Networks are represented in a variety of ways including computer models (Ganesh et al., 2005), formulas and theories (Keeling & Eames, 2005), diagrams and visualisations (Lima, 2013). Modelling scientific problems as networks can be helpful because it helps researchers to document the nodes and relationships in a network (Lima, 2013), as well as to predict the effects of changes in a network (Christakis & Fowler, 2013; Ganesh et al., 2005). In health research, the former advantage is useful for understanding the causes of diseases, and the latter is particularly important to health researchers seeking to understand the likely effects of interventions.

1.4.2 Health as a complex system

Our health is comprised of many interacting factors, so it is natural to investigate it using a network approach. Taking a systems thinking view (Klabbers, 2003; Meadows, 2008) our health can be seen as a system in which many inter-related factors are associated with disease. For example, obesity is related to many factors (Wright & Aronne, 2012) including biological (e.g., weight), social (e.g., social eating behaviours), behavioural (e.g., mealtime conditioning), developmental (e.g., childhood portion sizes), and affective factors (e.g., emotional eating). Over the next two chapters I will demonstrate how this network complexity makes it more difficult for us to understand the effects of everyday factors like sleep and wellbeing.

Understanding the topology of health can help us understand what causes diseases and how to treat them. The "topology" of a network describes what nodes are present and what the arrangement of relationships between them is. The topology of networks is often investigated in health research, for

example to understand how infectious diseases spread through social networks (Ganesh et al., 2005). This is important because it can help researchers identify all the factors related to a disease. Furthermore, it can help specify the exact causal mechanisms responsible for producing a disease, identify the steps in causal pathways, and identify targets for intervention.

Health is also dynamic and operates as a "complex adaptive system" (Holland, 1992). The relationships between health factors are constantly changing and adapting to our environment, and to human behaviours. For example, a smoking cessation program may reduce smoking at first but over time people will fatigue and it will become less effective (Liu et al., 2013). Understanding the human factor, how people behave, is particularly important to public health where interventions can be designed around how people behave in certain situations. Public health is therefore best understood from a viewpoint that health is dynamic, there are human factors, and researchers might consider a multidisciplinary approach incorporating a diversity of perspectives to help better capture public sentiment and values (Webb et al., 2020).

Intervening in human health is often considered a "wicked problem". Wicked problems are a class of problem for which "no single computational formulation is sufficient, for which different stakeholders do not even agree on what the problem is, and for which there are no right or wrong answers" (Kunz & Rittel, 1972; Rittel & M. Weber, 1973). The dynamic complexity makes predicting health outcomes extremely difficult. Understanding the aspects of health involving both social and biomedical causes is particularly challenging (Schrier, 2016) and requires a multi-disciplinary approach incorporating human factors with an understanding of the topology of the network of health. Researchers are developing new ways to address the challenges presented by complex health networks and one approach is using MR to investigate entire networks of causal relationships simultaneously.

1.4.3 Towards a network model of health

Network MR is increasingly being applied in human health to map between factors numbering in the hundreds (Brown & Knowles, 2020), or thousands (Hemani, Bowden, et al., 2017), using publicly available data.

The "human phenome" is the collection of all human characteristics. This includes every aspect from the biology of our cells and organs, to our habits and social behaviours. A human phenome project (Freimer & Sabatti, 2003) would seek to replicate the success of the human genome project (https://www.genome.gov/human-genome-project) but instead of identifying and sequencing the genes in human DNA, it would help understand the nature and causes of all human characteristics, phenotypes.

"Mapping" the human phenome and identifying the relationships between phenotypes may help us start to tame the wicked complexity of health by revealing causes. Given the complexity of health and its dynamic nature, a fully realised map of the human phenome remains in the realm of science-fiction for now, but researchers are already producing prototypic maps of a great many human phenotypes (Brown & Knowles, 2020; Hemani, Bowden, et al., 2017). Such "maps of the human phenome" (Figure 1.5) could have extremely valuable applications including identifying novel targets for drugs (Hemani, Bowden, et al., 2017).

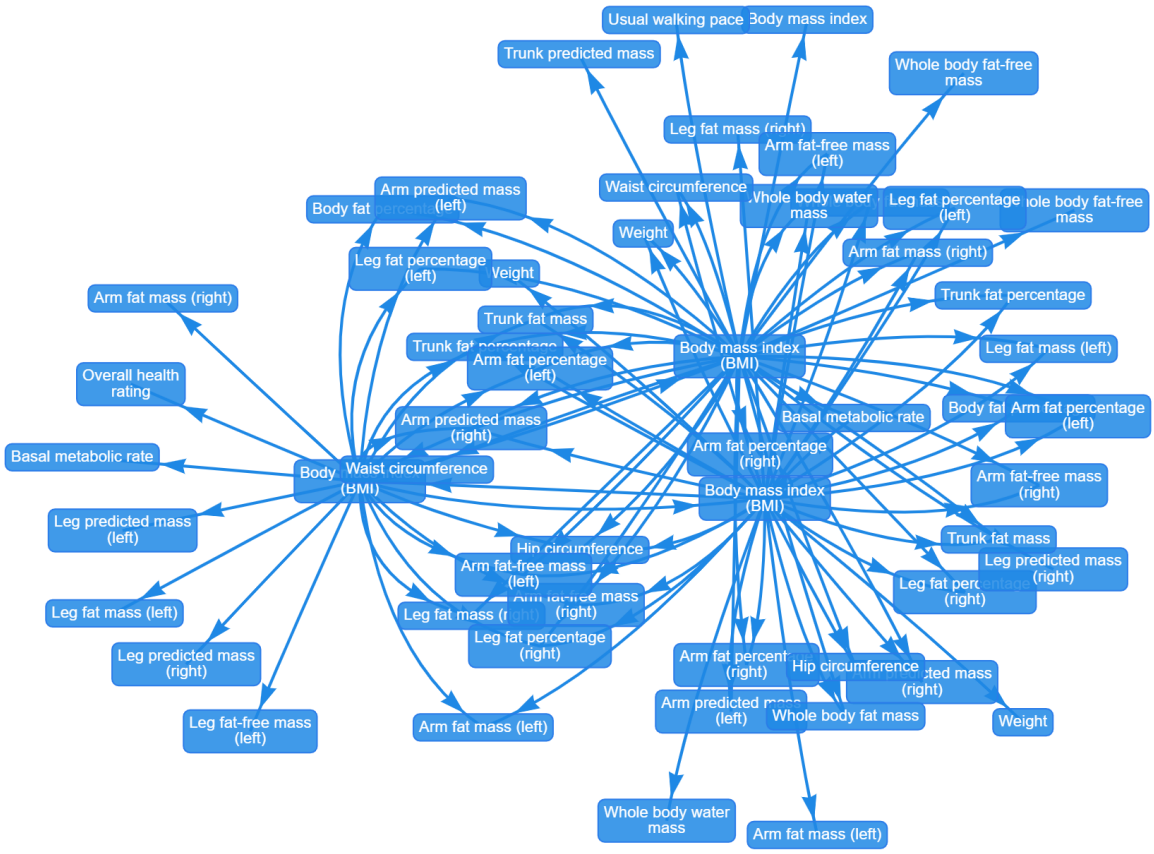


Figure 1.5 Maps of the human phenome strive to capture the network complexity of health. Pictured above is an example (Hemani, Bowden, et al., 2017) to demonstrate the large number of relationships that exist among health traits. This is a small slice showing 3% of the relationships from a phenome-wide analysis, in a map of the factors related to body mass index (BMI). BMI appears multiple times as it was measured using different ways and shows a lot of relationships with other factors including walking pace, metabolic rate, and overall health. The full map contains relationships not just with BMI but 2,406 other factors.

Recent developments in the availability of large-scale biomedical data makes it possible to test many phenotypes simultaneously. Commentators note that hypothesis-free approaches are viable for use with summary information of genetic correlations given the modern availability of large-scale and freely accessible datasets (Evans & Davey Smith, 2015). For example, the UKBiobank biomedical database, with its information on a range of physical and psychological measurements from half a million UK participants, will be a core source of data throughout this thesis. This availability of rich large scale datasets allows researchers to compare information on many factors in a single study, such as using MR to investigate entire networks of health factors.

1.4.4 Investigating health networks

Network MR (Burgess et al., 2015a) is an approach to generating and interrogating causal networks. I will apply it in chapter 3 so I will explain the opportunities and challenges it presents. It can be used in a hypothesis-free manner to test every possible relationship between a set of factors, is used in mediation analyses and can be used to produce plausible networks, but network results are often difficult to understand, and relatively little software exists to support its interpretation.

Hypothesis-free analyses do not test a specific hypothesis but rather a multitude of possibilities are tested with no clear expectations of outcomes. They are becoming increasingly popular in epidemiology and can help identify new causes of disease, and causal pathways, which may not have been known before (Evans & Davey Smith, 2015). Network MR can be used in a hypothesis-free manner since genetic instruments for many different factors can be used to test for the causal effects of each factor in a network on every other factor (Hemani, Bowden, et al., 2017).

Network MR is most often used to understand the wider context and interrogate causal effect estimates. The causal effect of an exposure on an outcome can involve additional factors which lie on the causal pathway and are instrumental in facilitating the causal effect. These are known as mediating factors and network MR is often used in mediation analysis to identify these factors and quantify their effects (Burgess et al., 2015a). Mediating factors are distinct from confounding factors because they represent a true causal effect and identifying them can help understand the mechanistic pathways of disease. Various methods of mediation network MR exist including methods intended for mediation in small networks (two-step MR)(Burgess et al., 2015a) as well as for large networks where multiple relationships are used to identify the most likely direct, unmediated, pathway between two factors (bi-directional mediated MR)(Brown & Knowles, 2020), and to offer mediation solutions for entire networks (graph MR)(Hemani, 2022).

Network MR can also be used to generate entire networks of effects and construct phenome maps. MR has only been conducted on the phenome-wide scale recently, following previous successes in

comparing genetic groups on the phenome-level (Evans & Davey Smith, 2015). For example, one study (Evans et al., 2013) demonstrated that genetic risk factors for body mass index show a plausible network of correlations with other health factors that replicates known relationships including heart disease and diabetes. Network MR has since been used to estimate the causal relationships between 2407 phenotypes (Hemani, Bowden, et al., 2017) including physical factors such as red blood cell count, behaviours such as exercise, and psychological factors such as wellbeing. A more recent study (Brown & Knowles, 2020) produced a slightly different map which uses mediation analysis to suggest the most likely causal effects between 405 phenotypes.

Using MR to obtain a map of the human phenome comes with advantages over observational approaches, but this is a novel area of research and methods supporting the inference in this area are still under active development. Network applications of MR benefit from the ability to determine causal pathways; applied at a large scale, this increases the confidence that resulting network estimates are not confounded. Network MR is also often combined with approaches for reducing the false-positive error rate associated with multiple testing (Hemani, Bowden, et al., 2017). This reduces the likelihood that causal links detected are caused by random chance, although it also reduces the power to detect smaller effects. For example, a Bonferroni correction for multiple testing (Sedgwick, 2012) lowers the *P*-value threshold considered good evidence for an effect by dividing it by the number of independent tests (e.g., an alpha level of *P*=.05 in single testing becomes *P*=.025 when making two tests). As an area of active research, network MR methods are currently limited in their practical uses. Authors (Brown & Knowles, 2020; Hemani, Bowden, et al., 2017) acknowledge that they are prototypes to demonstrate that network methods are possible, and further that testing and development is required to validate models.

Although results are preliminary, network MR helps us understand the topology of the network of health and understand the mechanisms of disease. However, the results of network MR and phenome mapping can be complex and difficult to understand. Even a small slice of a phenome map,

for example the one shown in Figure 1.5, can involve many factors, and the relationships between them are not always immediately clear because of overlapping lines and multiple indirect routes between nodes. Network analyses capture the complexity of health, and this is potentially valuable epidemiological information, but applying it to, for example, public health policy to develop treatments or preventative measures, requires distilling complex information into a format which is easier for researchers, policy makers and the public to understand. In the second half of this thesis I describe the development of methods intended to improve understanding of these complex networks by transforming them into different formats, such as interactive data visualisations and, beyond that, games.

1.5 Visualising networks

Philosophers and researchers have been developing methods to help describe and understand the complexity of life since early history. One of the first uses of imagery to convey a complex system of information is believed to have been in 270 AD where tree imagery was used to convey the relations between categories of life as branches on a tree (Figure 1.6). The use of tree imagery was influential in the classical era and became a particularly popular method of conveying networks, such as family trees, in the medieval era (Lima, 2013). Since then, significant advancements have been made in visualising complex information and the dedicated field of network visualisation has grown out of this historical experimentation. In this section I will introduce how visualisation is used in public health, describe specialised visualisations for conveying networks, and explain how they are used to convey health networks. I will end this section by introducing interactivity in visualisations.

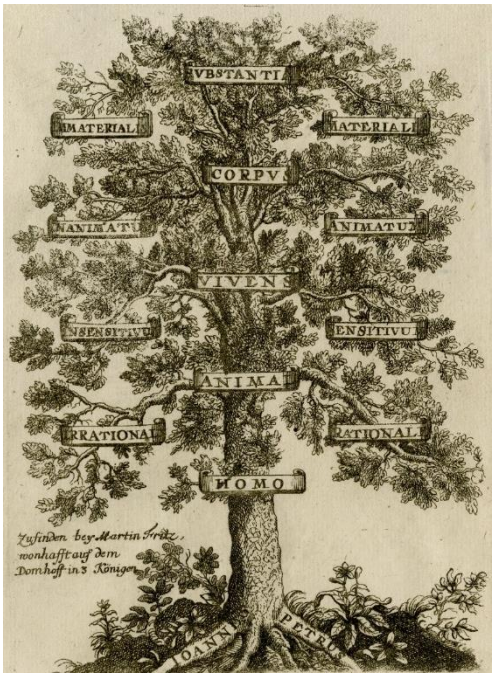


Figure 1.6 Visualisation has been used to understand complex information since early history. This artistic impression of Porphyrys of Tyre's "tree of life" depicts Aristotle's categories of life as branches on a tree (Martin Fritz, c.1680).

1.5.1 Health visualisation

Visualisation is used in epidemiology to understand the causes and effects of disease and communicate them to a wide range of academic and non-academic audiences (Greenland et al., 1999). I will demonstrate this by drawing on a range of examples from history and in modern use during the 2020-2022 COVID-19 pandemic.

There are two historical examples that are often regarded as particularly important for the development of data visualisation in epidemiology. Florence Nightingale's polar area graph (Figure 1.7) is an iconic representation of disease prevalence (Magnello, 2012). Nightingale produced a visualisation of the deaths the British army in the Crimean war (1853-56) were sustaining due to various causes and discovered that disease was the leading cause of death. This visualisation helped her to first understand, and then demonstrate this fact to the British army and policy makers who consequently revised their treatment of field medicine (Magnello, 2012). Similarly, anaesthetist John Snow produced a map (Figure 1.7) of cholera in London and was able to identify the Broad Street pump as the source of a localised outbreak. While it is contested how much of a role this map played

in this instance, his maps are credited with the advent of geographically mapping epidemics (McLeod, 2000).

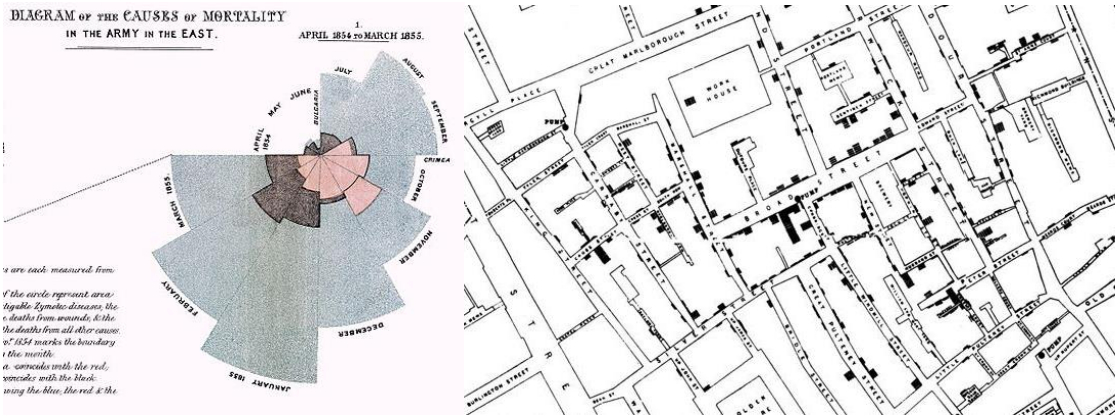


Figure 1.7 These iconic representations of disease prevalence shaped the way we visualise health today. The Nightingale polar area graph (left) representing causes of mortality for the British army in the Crimean war were predominately disease not injury. John Snow's cholera map of London (right) which demonstrated that cases, in black, were concentrated in one neighbourhood with an evidently contaminated water supply.

The influence of these early examples can be seen in the development of methods to visualise the vast amounts of biomedical data now available (Buckingham, 2008) and communicating the results of analyses in a clear and understandable way to the general public and to policy makers. One example is the current trend for "infographics" (McCrorie et al., 2016) which aim to distil data into a concise and engaging format suited for general readership. They are often simple, aesthetically pleasing, and colourful (Figure 1.8)(Harrison et al., 2015) and may be combined with icons to quickly and effectively communicate important ideas in population health science (Neurath & Kleinschmidt, 1939), an approach that came into its own during the COVID-19 pandemic (2020-2022).

Figure 1.8 Visualisations are used to communicate key health messages. Infographics (left) use simple and colourful designs to communicate messages in a simple and engaging way, for example this population pyramid demonstrates that homeless individuals tend to be young males (J. Song et al., 2022). Icons (right) are often used to help convey aspects of complex messages without the need for additional text, for example this is an excerpt of a public health poster (Neurath & Neurath, 1955) which shows that individuals will progressively recover from leprosy with treatment.

In the narrower field of causal epidemiology, drawing diagrams is standard practice to communicate and interpret causal effect estimates (Greenland et al., 1999). These are especially helpful for documenting the factors related to exposure and outcome, and for identifying additional factors which may confound the relationship between exposure and outcome (Suttorp et al., 2015). Diagrams are often used by researchers to present hypotheses (Textor et al., 2011). However, causal diagrams tend to be simple (Figure 1.9) and other methods of visualisation are required to convey complex network information.

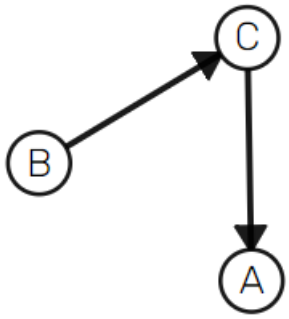


Figure 1.9 Causal diagrams are used to represent the relationships between factors. In this example, factor B causes factor C which in turn causes factor A.

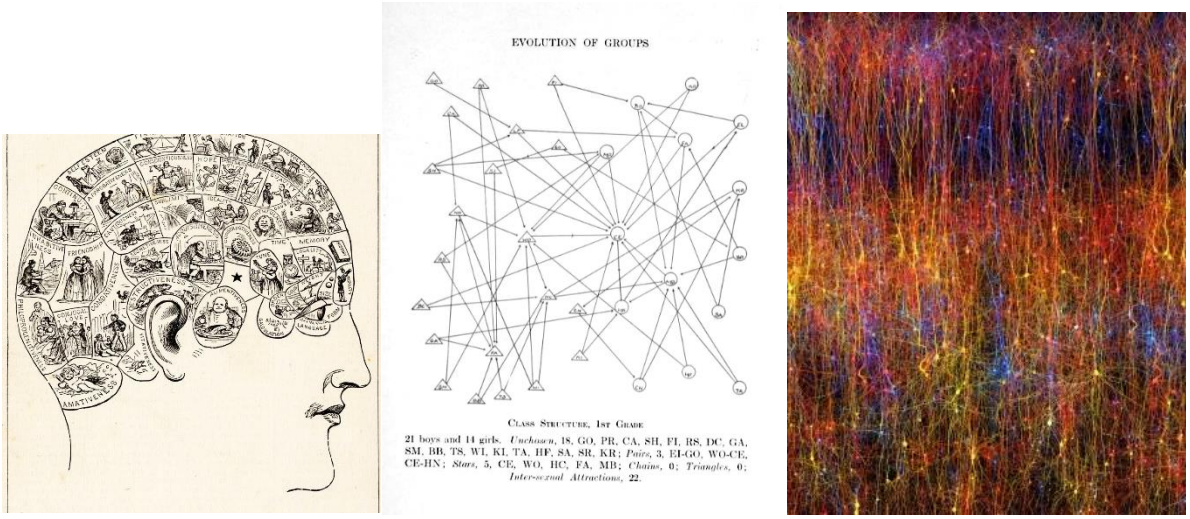
1.5.2 Mapping networks

Networks are often visualised using maps which help identify the factors in a network and the relationships between them. Network visualisations are argued to help our understanding of

organised complexity in several ways, including documenting, clarifying, and revealing information (Lima, 2013). They help achieve understanding on three levels:

> **The micro-level** describes the individual factors in a network. For example, the pseudoscientific movement Phrenology involved mapping special functions to different areas of the brain and this was often conveyed visually (de Puy, 1883).
>
> **The relationship level** describes the relationships between factors. For example, Social Psychologist Jacob Moreno mapped the social relationships between children using 'sociograms' (Moreno, 1933).
>
> **The macro-level** describes the overall pattern of interaction between factors. For example, the Blue Brain Project (Kaviya, 2014) generated a computer model which demonstrates a pattern of massive interactivity between ten thousand neurons.

Network visualisations therefore help researchers understand network complexity in terms of the factors contained within a network, the relationships between them, and the general pattern of effects (Figure 1.10).



Micro level Relationship level Macro level

Figure 1.10 Network visualisations are used to convey details at various levels, for the factors contained within a network (left)(de Puy, 1883), the relationships between factors (middle)(Moreno, 1933), and the overall pattern of inter-relation (right)(Kaviya, 2014).

In epidemiology, network graphs are often used to visualise complex health networks. These visualisations map diseases which have many causes and effects, or map a wider network of factors related to a causal effect which may confound or mediate it (Greenland et al., 1999).

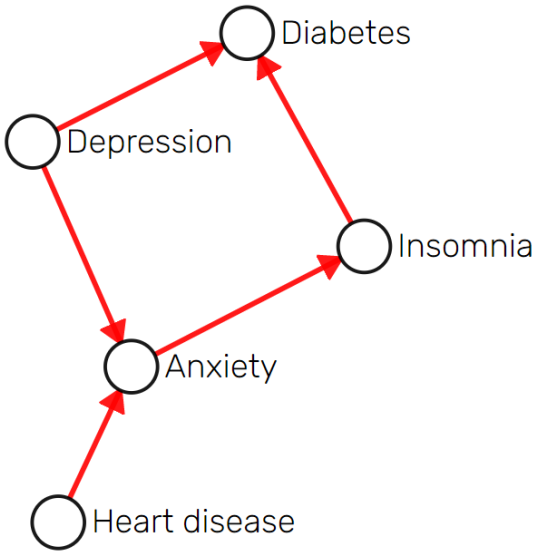


Figure 1.11 Network graphs help epidemiologists understand the factors involved in a network as well as the relationships between factors. This example visualises complex inter-related effects between heart disease, anxiety, depression, insomnia and diabetes.

Drawing these diagrams helps researchers to understand epidemiological networks at the three levels mentioned above (Lima, 2013), to document the factors related to a disease, to understand

precisely how they are related and if relationships are causal. They also help clarify the overall pattern, whether it indicates that factors are highly inter-related, are largely separate, or whether there are "hubs" of factors which are much more closely related than others. In chapter three I review causal epidemiology literature and find that this type of visualisations is often published in academic papers. However, despite the demand, there is currently relatively little specialist software available for researchers to visualise network results.

1.5.3 Interactive visualisation

Visualising a complex network is not always sufficient to understand it and interactivity is often used as a way of further immersing a reader with complex information. Interaction is used to provide several functions to users including presenting an overview and details on demand, as well as zoom and filter abilities (Shneiderman, 2003). Providing users control over the information the view can help them restrict their investigation to the factors they are most interested in. In a similar way varying levels of detail allow users to control the amount of information they are presented with and avoid information overload. A recent example of a widely viewed interactive visualisation is the UK Government's COVID-19 dashboard (https://coronavirus.data.gov.uk) where users can switch between various data views, filter information, and select different levels of detail.

With interactive visualisations, users do not have to be a passive recipients of information and can be involved as active participants. In interactive simulations the user is given an active role exploring, making changes and experimenting with data (Joldersma & Geurts, 1998). Complex adaptive system simulations further situate the user as an agent within a system where humans play a role, such as inhabitants of a world undergoing a pandemic (Lofgren & Feff, 2007). For example, one simulation (van Bilsen et al., 2010) immerses players in information about the logistics and management challenges of operating a shipping port in the Netherlands. However, interacting with a visualisation requires more engagement than viewing it, and the freedom for the user to view and change data how they like might result in a less focussed experience. People do not always feel strongly

motivated to interact with data, therefore some researchers are adding game features to motivate users and encourage them to engage in a specific way, for example by setting goals to achieve (Zyda, 2005a). This approach has been established as a valuable tool for understanding and communicating complex topics in science with many examples of success (Schrier, 2016).

1.6 Towards playable networks

Games are often considered niche and frivolous, certainly not a part of scientific research. However, recent developments have led many to question this assumption. Games are now a mainstream interest and they have made valuable contributions to serious fields such as education and research. In this section I will define what games are and outline the current state of gaming. Games are emerging as valuable tools for engaging people with serious tasks, and I will explain how certain gameplay mechanisms help structure and motivate a player's engagement. I will close this section by highlighting the opportunities for new experimental research in this area, and argue that this is important to evaluating whether adding gameplay to science projects is worth the investment.

1.6.1 The state of play

The classic game "Tetris" (Pajitnov, 1984) asks players to combine blocks to solve puzzles for fun but now games such as "FoldIt" (Eiben et al., 2012) ask players to reprise these skills to help researchers analyse scientific data, for example discovering the functions of different protein configurations. This is indicative of a shift from viewing games as simply for entertainment and towards the notion that gameplay can be useful for serious purposes as well. It is important to set the background of games in science; they are now mainstream and are being applied in serious contexts but there are some outstanding controversies surrounding the wider gaming industry.

Games now have mainstream appeal, and videogames are particularly popular. In Europe, 50% of the population play videogames (ages 6-64)(ISFE, 2021), and some games are particularly popular with player counts in the hundreds of millions, including Fortnite (Epic Games, 2017) and Minecraft

(Mojang Studios, 2011). Furthermore, the traditional perception that games only appeal to young males is not true in the current market since 47% of gamers are women and 31-43% of the adults over the age of 35 play games (ISFE, 2021). Arguably this stereotype was never true but rather was the product of gendered marketing in the 90's (Etchels, 2019). Games therefore have tremendous appeal and the potential to engage a large proportion of the population.

Games are no longer associated with a childish activity which trivialises or exploits, and are instead increasingly being used as a medium for engagement with serious topics (Schrier, 2016). In her book *Knowledge Games,* Karen Schrier (Schrier, 2016) coined the term to describe games which create insight, solve problems, or create change. These games help achieve positive change in the world in a variety of ways from educating us about serious topics in health, contributing to scientific research, or influencing government and policy making. For example, "Go Viral" (Basol et al., 2021) was developed by Cambridge University to help develop "psychological herd immunity" against COVID-19 misinformation. It effectively improved players' confidence in spotting misleading information spread on social media platforms and reduced their likelihood to propagate and share it. This effect was present two weeks after and demonstrates that games can achieve real-world impact about serious topics. In chapter 5 of this thesis I develop a similar type of game which is focussed on aiding on-going research in epidemiology. I use gameplay to direct and motivate players to engage with problems of network complexity in epidemiology, learn about them, and suggest solutions.

The public image of games as good forces is, however, marred by controversies which affect the general perception of games. Women and marginalised group have often been poorly portrayed in games in ways which can be stereotypical, misogynistic, sexist and racist (Larsson & Goldberg, 2015). The videogame industry is also known for engaging in exploitative labour practices (Schreier, 2017), as well as under-representing and marginalising female and minority voices (Larsson & Goldberg, 2015).

There is also a pervasive but unsubstantiated, belief that certain videogames lead to violent behaviour in real-life situations. The best publicised example of this was the mission “No Russian” in Call of Duty Modern Warfare 2 (Activision, 2017). In this mission the player views, and can participate in, a terrorist attack on an airport. At the time this grabbed headlines, and will not feature in a re-make (Modern Warfare 2, Activision, 2022), but this demonstrates an example where violence in games is misunderstood. This mission continued the game series’ legacy for portraying war in a realistic and horrific manner, to make players uncomfortable, which was something games were not doing at the time. Some research into videogame violence (Ferguson et al., 2020; Mathur & VanderWeele, 2019) has highlighted correlations between time spent playing games, particularly violent games, and violent thoughts, but there is no strong evidence that this relationship is causal. It is generally considered that this association reflects selection bias, that individuals with more violent thoughts are more likely to play violent types of games, and that there is a degree of desensitisation and normalisation meaning that violence in games has become distanced from real-life violence. .

Modern videogames are an established mainstream interest where the majority of individuals engage in healthy ways. However, it is important to note, and avoid contributing to, the wider ethical issues with the genre, and to be aware that individuals’ perception videogames may affect the application of games in science. The second half of this thesis focuses on answering questions around how we *can* use games for good purposes in society, but I will also discuss the wider value of games in society, and whether we *should* use them in science. Before diving deeper into the details of how games can be used for science, I will explore a formal definition of games.

1.6.2 What are games?

Games are characterised as a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome (Salen & Zimmerman, 2003). The different genres of games can give a more general idea of the play experience. For example, shooters involve shooting, strategy games involve strategizing, and team games involve playing as a team. In this section I will

expand on this definition of games as systems of play which consist of “rules”, the experience of “play”, and the wider “context” for play.

“Rules of Play” (Salen & Zimmerman, 2003) is a seminal work which is used by many as an academic framework for designing, developing and researching games. In this framework games are considered systems bound by explicit rules. The “rules” of a game establish the boundaries within which the player can freely play. For example, chess has pieces, such as knights, which can move in a certain way and interact with each other, for example by jumping around the board and taking another piece. A key aspect of game systems is that the player can interact with them, change them, and in turn the system responds back. A “gameplay loop” describes the sequence of actions which players engage in when they play a game. For example, in chess, a player will strategize, move a piece on the board, their opponent will respond, and gameplay loop continues until the game ends. Rules therefore comprise the objective and observable properties of a game (Deterding et al., 2011) and these facilitate play.

If rules are an objective characteristic of games, “play” is the subjective experience of these rules. Play is free and explorative (Groh, 2012) and involves experimenting within the rules of a game (Salen & Zimmerman, 2003; Susi et al., 2007). For example, players can make any move in chess as long as it is legal. A player’s individual experience of play within rules is subjective and what is fun to one individual is not necessarily fun to another (Marczewski, 2018). An important aspect of play is that the experience of play is real even if players are experimenting and exploring a make-believe world.

Play in games is often misunderstood as frivolous, particularly in games which explore fantasy worlds. However, whether a game models the real world or not, a player’s experience of play exists in the real world and it can evoke real feelings, thoughts and beliefs. A “magic circle” is the ritualistic belief that there is a world of magic and you can draw portals between this and the real world, and this is often used as a metaphor for the real experience of an imagined, fictional, or virtual game

world (Etchels, 2019; Goldberg, 2015). This allows players to have one foot safely in reality and another in experiencing a fictional world, and this is important since it allows players to experience scenarios which would be dangerous, impractical, or impossible in the real world. For example, the French army medic corps are exposed to real-world battlefield conditions in the educational game "3D-SC1" (Pasquier et al., 2016). Games can model the real world in different ways and while games like 3D-SC1 offer a more literal model of the world, other games can model real problems in a more analogous manner (Schrier, 2016). For example, the videogame series "The Witcher" (CD Projekt Red, 2015) represents historic struggles in Eastern Europe as tensions and conflict between different races of humans, elves and dwarves in a grim fantasy setting. The magic circle therefore allows players to play in virtual worlds which are not necessarily bound by the rules of reality, yet they can model elements of reality. While commercial videogames are played for fun there is a different context for playing games in science.

The "context" of a game describes the wider culture, meanings and purpose for play (Salen & Zimmerman, 2003). Games both reflect the culture of the real world, and can change and create new culture. For example, competitive games such as "Dota 2" (Valve, 2013) have a culture of sportsmanship which reflects that found in football or rugby and builds on it with new terms such as "GG" (which stands for "good game"). In contrast to playing games for entertainment, citizen science has emerged as a new context for playing games (Schrier, 2016). In citizen science games, play does not fulfil a solely entertaining purpose, but rather it is a way for player to contribute to science. For example, "FoldIt" is played partly because players want to help biochemists understand the role of proteins in diseases like cancer (Eiben et al., 2012). A similarly "serious" context for play is presumed by the designers of educational games, where the ultimate aim is to teach specific learning outcomes.

1.6.3 Knowledge games

Knowledge games use gameplay to engage players in serious contexts. They situate players as an active agent in the creation of new knowledge, rather than the passive recipients of educational materials, or providers of data points. I will explain what is meant by "knowledge", demonstrate that these games have been applied in many serious contexts such as health, and focus on fully integrating gameplay in a meaningful way.

The "knowledge" in knowledge games does not refer to the data a scientist might obtain but rather the meaning that results from analysis and understanding (Schrier, 2016). The knowledge that players create could be the conclusion from analysing scientific data, but it could also be learning or communicating information, which is socially, culturally or personally valuable. For example, in "GoViral" (Lewsey, 2020) players assume the role of a fake news social media influencer and identify how fake news is spread by spreading it themselves. Knowledge games empower the player to make meaning and this is a distinct way of using gameplay for serious purposes.

Knowledge games are a type of "serious game" (Sailer et al., 2017) which broadly have the aim of achieving outcomes in serious topics (Clapper, 2018). Research has reviewed and evaluated serious games across range of fields including politics, military science, organisation, logistics (Geurts et al., 2007), computer science, psychology and sociology (Schrier, 2016). Health science is perhaps the most common application of knowledge games (Schrier, 2016; Susi et al., 2007; Wardaszko, 2018) and they have achieved a range of outcomes in healthcare delivery, education and research (Table 1.1). For example, one study (Firth et al., 2017) found that gamified breathing exercises helped anxiety patients reduce levels of panic.

Table 1.1 Serious games are made for a range of applications in different areas. This table summarises the number of games covered in five of the largest reviews across various domains.

Area	Johnson and colleagues (2016)	Sardi and colleagues (2017)	Ricciardi and de Paolis (2014)	Schrier (2016) Peplow (2016)
Treatment and management games				
Physical exercise	7	2		
Nutrition	3	2		
Managing chronic disease	2	10	2	
Wellbeing	2			
Mental health	2	5[1]		
Addiction	1			
Pain management	1			
Education games				
First aid			13	
Surgery			6	
Dieticians			6	
Pain management			5	
Nursing			3	
Odontology			2	
Cardiology			2	
Psychology			2	
Research games				

Genetics	3[2]
Cellular biology	2
Cancer	2[2]
Malaria	1

Note: [1]Authors do not distinguish between mental health and wellbeing. [2] *Play to Cure: Genes in Space* is counted for both genetics and cancer biology.

In addition to situating the player as an active agent in the creation of knowledge, knowledge games are also distinct in that they involve a relatively intensive process of gamification. "Gamification" is the act of adding gameplay to non-game tasks (Deterring et al., 2011) and this can be conducted to varying degrees. Integrating many gameplay mechanics in a meaningful way is more likely to result in a "full game", whereas the addition of one or two mechanics in a superficial manner results in "gamified apps". Knowledge games are most often full games which fully integrate real-world problems into gameplay to produce a cohesive and engaging experience. Gamification is often represented as a spectrum from non-game to game (Qin et al., 2010), and knowledge games lie closer towards being a game (Figure 1.12).

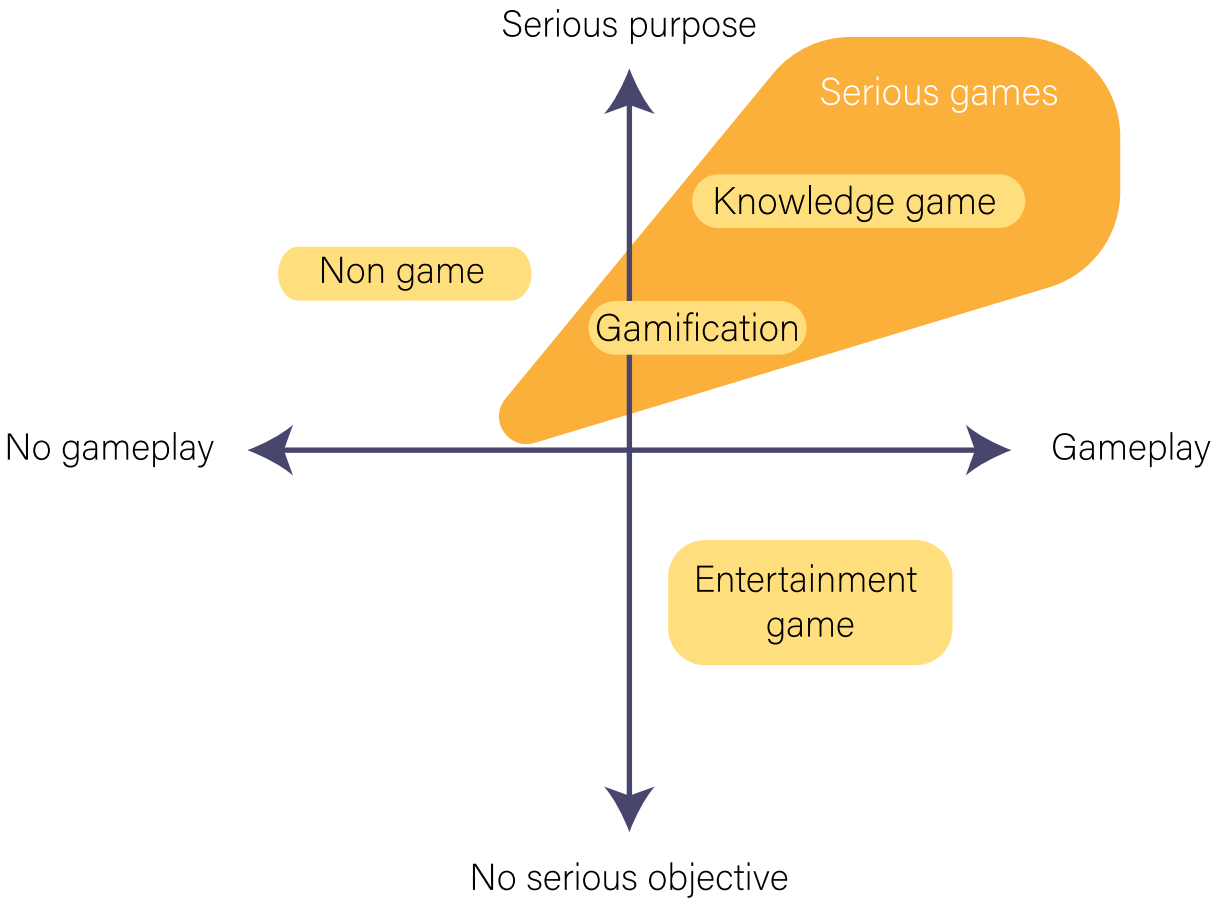


Figure 1.12 Knowledge games are a type of serious game which uses substantial gameplay for a serious purpose.

There are many reasons why researchers might choose to not produce a game, for example resources or expertise may be limited, or it might not be desirable to completely transform a task. For example, in one study (Lumsden et al., 2016b), a scoring system was added to a cognitive test of attention, and the authors deliberately chose this so as not interfere with the psychometric properties of the test. Knowledge games, however, are deliberately designed to achieve outcomes through the course of a substantial play experience (Schrier, 2016). This is not to say that every game should or could be as entertaining as commercial mega-hits such as Fortnite (Etchels, 2019; Schrier, 2016), this is a lofty goal to achieve after all, but there is a focus in knowledge games for using and investigating gameplay as mechanisms for creating insight and change.

1.6.4 Gameplay mechanisms

There is a pervasive notion that games are fun and this is misleading in the context of understanding how knowledge games work. While this is of course true to some degree, the idea that fun is inseparable from games has contributed to confusion in the academic literature of serious games.

"Fun" is a nebulous term and classically difficult to define (Bisson & Luckner, 1996). Since it is subjective the idea of fun to one person may not be true for another. One games researcher's perspective (Marc Prensky, 2001) is that fun gives enjoyment and pleasure. A broader entertainment perspective (McKee, 2016) similarly defines fun as "pleasure without purpose". A different perspective is taken in education (Bisson & Luckner, 1996) where fun is defined adjacent to educational tasks, as a positive experience of flow characterised by motivation, engagement and affinity for a task (Csikszentmihalyi, 1990).

The term "serious game" was coined to deliberately contrast the fun nature of games as an oxymoron, but this has resulted in misinterpretations and being re-defined many different ways (Djaouti et al., 2011). For example, in one academic debate (Clapper, 2018) the notion that games are fun drew focus towards the question "are serious games fun?" and away from answering arguably the more important question "should they be fun?". A game is bound by rules but the experiences that players have during play are diverse (Lucero et al., 2013). For example, a game could give a player an experience which is most conventionally "fun", such as thrill and excitement, or it could give feelings of exploration and awe. In this thesis I focus on two knowledge game mechanisms that are not thrilling, but instead aim to structure and motivate players to engage with serious tasks, producing feelings of affinity, ability and a desire to continue.

Structure

Knowledge games direct players' attention and behaviours in ways which are conducive to problem solving. I will explain how gameplay can be crafted to encourage players to explore problem spaces in ways that foster understanding or help generate solutions. Problem spaces model a problem in a

realistic manner and allow players to experiment with solutions in a way which is constructively aligned to the real-world problem.

In David Jonassen's theory of problem solving (Jonassen, 2000) a problem is defined as "an unknown entity in a situation whose solution has some social, cultural or intellectual value". The rules in knowledge games are deliberately designed to guide, limit, structure and nudge players to engage with these problems, learn about them, and even suggest solutions through play (Schrier, 2016).

Knowledge games model a real-world problem in a world which helps players understand it. A complex problem can be communicated to the player through gameplay (Schrier, 2016) and user interfaces (Wardaszko, 2018) in a manner which makes it understandable and more accessible without over-simplifying it to the point where it is inaccurate (Cannon et al., 2010). Games also give players tools to understand the world. Play can take the form of active learning where players learn by trial and error of what the game allows or rewards (Salen & Zimmerman, 2003; Schrier, 2016). This is argued to have value as a method of "implicit" or "experiential" learning (Cannon et al., 2010) where players develop an implicit knowledge by engaging with tasks that are parallel to learning outcomes (Seger, 1994). Furthermore, play is argued to take the form of "problem-based" learning (Schrier, 2016) where learners accumulate knowledge in order to solve a problem (Wood, 2003).

Constructive alignment (Biggs, 2003) is a term used in education to describe making the modalities of learning and testing as similar as possible. For example, an intended learning outcome for properly applying wound dressing would be best tested in a practical test of wound dressing rather than a paper exam. In a similar way gameplay can be aligned such that players interact with a problem in a manner as close to real-life as possible. Complex simulation games provide some good examples of this. Complex topics such as politics (Positech, 2013), government (Paradox Interactive, 2013) and military strategy (Paradox Interactive, 2016) are simulated in enormous detail and gameplay is designed to allow players to interact with this simulation in a manner as close to real life as possible. For example, in "Democracy" (Positech, 2013) political power is bought using currency

and this restricts players to play from a perspective of having limited financial resources. In knowledge games, constructively aligning gameplay can help achieve educational outcomes by better modelling real-world complexities to players, and research outcomes can be achieved when gameplay coaxes players to offer solutions. For example, the players of FoldIt (Eiben et al., 2012) can suggest viable protein combinations to the researchers because the gameplay, arranging molecules in a 3D model, is directly aligned with the output they hope to receive, 3D models of new proteins.

Motivation

A second mechanism of knowledge games is that they motivate players to engage with a topic for longer (Schrier, 2016). Motivation is one reason why gameplay is engaging (Sailer et al., 2017) and is especially helpful in understanding how it can help achieve behavioural outcomes, such as contributing data to research and these types of serious games are often designed to motivate (Johnson et al., 2016; Susi et al., 2007). Additionally, motivation is a key aspect of what makes game play engaging and many game design frameworks guide designers to understand how game features achieve feelings of motivation (Bartle, 1996; Chou, 2014; Marczewski, 2018; Mora et al., 2015; Yee, 2006).

Motivation helps individuals start or continue behaviours (Schunk et al., 2012) and games are argued to be a source of motivation (Schrier, 2016). Self-determination theory (Deci & Ryan, 2012) has become a core theory for understanding the role of motivation in gameplay (Johnson et al., 2016). In this model we are motivated to meet basic "intrinsic" psychological needs including competence, autonomy and relatedness. Games can help meet these basic needs (di Tomasso, 2011; Marczewski, 2018), for example games give players choices and this can give feelings of competence when good choices are made, or autonomy when freedom of choice is provided. Another theory is that games induce a state of immersive concentration known as flow (Csikszentmihalyi, 1990). Similarly, setting objectives in goal-oriented play may also help motivate players to achieve outcomes in-game and in the real-world (i.e., goal-setting theory)(Locke & Latham, 1994). These theories of motivation are

used by game designers since play experiences are often crafted to appeal to the intrinsic motivations of players (Mora et al., 2015; Tondello et al., 2018).

There is some evidence that in practice gameplay motivates individuals to engage across a range of contexts including learning (Looyestyn et al., 2017), commerce, healthcare and data gathering (Hamari et al., 2014). A systematic review of 22 studies (Hamari et al., 2014) found that each study reported that gamification improved motivation in either a self-report or objective measure of engagement. Although this success rate is likely inflated by a file drawer effect where failures are not published, it is supported by a similar meta-analysis of 15 e-learning courses (Looyestyn et al., 2017) which found that 80% of courses benefitted from gameplay and the effect size improvements to motivation were medium to large (0.4 to 1.1 standard deviation improvement).

1.6.5 Digital and physical games

In this thesis I decided to make a computer game, and while the underlying theory behind play is largely similar regardless of the modality of play, this decision came with some practical advantages and disadvantages in terms of harnessing computer processing power but complicating development.

The modality of play does not necessarily affect the rules of a game but may affect the play experience (Salen & Zimmerman, 2003). The rules of games are often broadly similar across physical and digital modalities. For example, digital and physical chess have identical rules. However, whether players interact with a computer screen or are playing a physical board can affect the player experience. For this reason I will draw on evidence from computer games where possible as the play experience in these games is more similar and most relevant to my thesis.

Computer games are played on a computer which offers the opportunity to model complex mathematical simulations of a problem, but developing a computer program is more complicated than making a card game. The advantage of computer processing is considerable and has been credited before as a key reason why games are able to simulate the complexity of real-world

problems (Wardaszko, 2018). Furthermore, a computer provides functionality which would be cumbersome or not possible without a computer. Knowledge games such as FoldIt offload computations from player to computer, so the player can concentrate on a task humans are better at, pattern matching, while the computer quickly processes, modifies and analyses vast amounts of data for the player (Schrier, 2016). The same is true of rules, which can be applied automatically in a computer game, but require mental effort in a physical game. A computer is also able to provide players with immediate feedback which makes for more responsive gameplay (Salen & Zimmerman, 2003). Additionally, a computer game uploaded online can be accessed by anyone anywhere (this naturally become more important during the pandemic lockdowns of 2020 and 2021). However, the decision to make a computer game came with drawbacks, and particularly complicated development. In chapter 4 I will discuss the importance of frequent testing sessions where a select group of individuals help identify issues, refine gameplay and review subsequent versions of the game intended to fix any problems. This is much more difficult with a digital game because software takes longer to alter than a physical (e.g. paper) game does. Developing a computer game is therefore a considerable undertaking and it can be argued that while it is popular to ask the question "can a game do this?" we should also ask "is it worth making a game to do this"?

1.6.6 Is adding gameplay to non-game tasks worth it?

The games research literature is popular, influential, and has drawn a lot of funding but it remains unclear whether investing in gameplay is worth it. I will now give some background on games research, explain that there is little understanding of the specific value of gameplay, but that developing a game is a considerable investment, and argue that it is important to understand this so that we can ensure resources are spent on the most effective courses of action.

Gamification is applied in a considerable amount of research. Games research draws on many fields (Salen & Zimmerman, 2003) and is applied in many fields as well (Susi et al., 2007). For example, gamification is incorporated into education (de Sousa Borges et al., 2014), marketing (Zichermann &

Linder, 2010), wellbeing and health research (Sardi et al., 2017). Furthermore, games research journals exist that are specialised to certain fields of study, including simulation (SAGE Simulation and Gaming), disease prevention, healthcare delivery (Games for Health Journal), education, assessment, data collection and computation (Journal of Medical Internet Research: Serious Games). Some research also achieves high impact publication in journals with general readership such as *Nature* (Eiben et al., 2012).

It is unclear, however, what role gameplay plays in the success of applied games research. Commentators note that games research suffers from a lack of consensus over definitions, since games and specific mechanisms of gameplay are often referred to with overlapping terms (Susi et al., 2007) and this makes it more difficult to understand the underling constructs researchers are referring to. It is also often noted that games studies are designed in different ways, and do not necessarily follow best research practices. For example, in the literature assessing the potential for videogames to cause violence, reviewers note that heterogenous and poor-quality study designs make reviews and meta-analysis particularly difficult because large differences between studies reduce the validity with which they can be compared (Mathur & VanderWeele, 2019). A final issue, which I specifically address in chapter five, is the absence of experimental research comparing games to non-game controls. For example, in the literature assessing the motivational value of gameplay, reviewers note that more experimental studies would increase the strength with which researchers can attribute the positive outcomes of games to the gameplay (M. Brown et al., 2016). Just as in observational epidemiology, it is possible that observed outcomes of gamification are due to some other aspect, such as making a non-interactive task interactive. Therefore, while there is a considerable amount of research effort focussed on *applying* gamification, relatively little research focusses on *understanding* gameplay, and the latter is necessary to inform the former.

It is important to establish whether games in science really work because the resources spent on developing them would be inefficiently spent if gameplay does not have a large effect and could

instead be put towards better uses. Creating a game is a considerable endeavour that requires extensive resources in terms of time, personnel, and often financing (Etchels, 2019; Salen & Zimmerman, 2003). Game design is not the act of simply combining ingredients, but rather requires the careful consideration of the rules, play and context of the game. Interviews with professional game development teams (Schreier, 2019) indicate that even high-profile, well-funded, experienced teams can fail to produce compelling games even when they have a previous example of success to guide development. There exists no list of requirements for producing the success of hit commercial games such as Fortnite (Cai et al., 2019; M. Carter et al., 2020; Jiang, 2020; Marlatt, 2020); creating engaging gameplay remains challenging, although like reproducibility in research, the process can be made less risky through following best practice guidelines such as regular playtesting (Fullerton, 2018). Knowledge games do not necessarily need to reach the same level of engagement as commercial blockbusters, but their gameplay does need to be carefully designed and integrated into their real-world context to achieve their goals (Schrier, 2016). Therefore, designing gameplay is a resource intensive process and researchers and practitioners alike need to be able to justify that this is a valuable use of resources in research, education and healthcare. In chapter five I conclude that although developing a knowledge game took a long time in my case it was greatly effective in engaging learners with education and participants with research.

1.7 Chapter overview

This chapter introduced the problem of network complexity in epidemiology and methods of understanding this complexity. In this thesis I develop and apply a combination of techniques to demonstrate three main conclusions and contributions to our understanding of network complexity in human health. First, in chapters 2 and 2 I apply Mendelian randomisation to obtain a network dataset that highlights that physical and psychological health factors are causally related through often complex paths, and this makes understanding them difficult. Second, in chapter 4 I highlight the relative lack of network visualisation software for Mendelian randomisation and develop novel

software to fill this niche. Third, in chapters 5 and 6 I demonstrate that a custom-built game can engage players with a model of network complexity in a way that a non-game simulation control does not. My thesis will end with a discussion of the many challenges of conducting this type of work, during a pandemic or otherwise, and highlight the next steps for collecting further evidence of the specific benefits of gameplay.

2 Testing the causal relationship between insomnia and wellbeing: A Mendelian randomisation Study

2.1 Introduction

This chapter acts as a foundation for understanding the Mendelian randomisation (MR) method used in this thesis. I will apply MR to investigate the relationship between sleep and wellbeing. There is some evidence that poor wellbeing and sleep disorders such as insomnia are associated, but there is limited evidence about a causal relationship. MR will be used as a method of overcoming some limitations of previous research and reveal information about a causal pathway.

2.1.1 Wellbeing and sleep are important to our functioning as healthy individuals

Wellbeing is central to our experience of life, and sleep is a necessity, yet they are both complex and poorly understood. I will outline empirical research on wellbeing and sleep in turn, explain the detrimental effects that poor wellbeing and sleep disorders can have, and highlight the importance of understanding them.

Wellbeing is a complex phenomenon that influences our everyday experience of life. Aristotle first described eudemonia as a state of wellbeing that arises from living a virtuous life. Since then, wellbeing has been defined in different ways to describe a range of phenomena with different causes and effects (Diener, 1984). In this thesis I use a definition of subjective wellbeing which was pioneered by Ed Diener which describes a state of happiness, satisfaction with life, and the absence of negative emotions and feelings (Diener, 1984; Diener, Oishi, et al., 2018). In this view, wellbeing is an individual's subjective experience and evaluation of their life according to one's own criteria, beliefs and values. Despite its individual and personal nature, population-level factors have been found to reduce wellbeing, particularly unemployment and perceived low socioeconomic status, and to improve wellbeing, including perceived social support and quality social relationships (Diener, Oishi, et al., 2018).

The effects of good and poor wellbeing are not yet fully understood but there is mounting evidence that poor wellbeing can have profound effects on our health and everyday lives. Good wellbeing can have protective effects on physical health while research has demonstrated detrimental effects from poor wellbeing (R. T. Howell et al., 2007) . For example, higher levels of wellbeing are associated with a lower risk of cardiovascular disease, stronger immune response, and healthy behaviours including exercising and not smoking (Diener et al., 2017). Furthermore, poor levels of wellbeing in non-depressed individuals is associated with a greater likelihood of developing depression later in life (Lewinsohn et al., 1991). Although wellbeing and mental illness are highly inter-related it is important to distinguish the two since they are commonly confounded (Diener, 1984); poor wellbeing does not necessarily produce mental illness (Iasiello & Joep van, 2020), and individuals with a diagnosis of mental illness can experience good wellbeing in their day-to-day lives (Greenspoon, P.J., Saklofske, 2001). More generally wellbeing is important to a functioning society since an unhappy society is an unproductive one (Marks & Shah, 2004). Wellbeing is therefore a pervasive and important aspect of life that is not yet fully understood.

Sleep is similarly essential to functioning in our everyday lives. Sleep is part of a wider pattern of physiological and psychological changes which occur throughout the day: our circadian rhythm (Deak & Epstein, 2009). A balance of “wakefulness” and “sleepiness” produces sleep (Daan et al., 1984). The balance is scheduled to tip towards sleepiness at night by an internal 24-hour body clock, and towards wakefulness during the day.

A large proportion of the population experience difficulties sleeping (20-48%) and symptoms like daytime sleepiness (4-26%)(Stallman & Kohler, 2016). Sleep can become disordered in a number of ways and the International Classification of Sleep Disorders (Sateia, 2014) recognises five main disorders which affect the duration, quality or schedule of sleep (Table 2.1). The most commonly diagnosed sleep disorder is insomnia which negatively impacts the duration and/or quality of sleep (Ohayon, 2011).

Table 2.1 Prevalence of the most common sleep disorders

Disorder	Example	Estimated prevalence
Difficulties initiating or maintaining sleep	Insomnia	6% to 15%
Disorders of excessive daytime sleepiness	Narcolepsy	0.02% to 0.07%
Sleep breathing disorders	Obstructive sleep apnoea	2% to 4%
Sleep movement disorders	Restless Leg Syndrome	3.25% to 8.5%
Parasomnias	Sleep walking	1.5%[2]

Sleep can become disordered in a number of ways but the most prevalent sleep disorder is insomnia. Prevalence estimates were sourced from the Stanford Sleep Epidemiology Research Center (Ohayon, 2011) as well as a study of sleep walking (Stallman & Kohler, 2016), and describe European and American populations.

In this thesis I define sleep disorders as sleeping difficulties which impact the duration, quality or schedule of sleep. Compared with clinical diagnostic criteria of sleep disorders, for example the International Classification of mental and behavioural Disorders (ICD-10) (World Health Organization, 1993), this definition includes a broader range of individuals who may experience sleeping difficulties which are transient or do not result in daytime impairment, or that would meet diagnostic criteria but have not yet sought a diagnosis. Accounting for undiagnosed individuals is important because the large discrepancy between the prevalence of sleep disorder symptoms and diagnoses (Table 2.1) is often interpreted as evidence that sleep disorders are under-diagnosed (Saddichha, 2010). This wider criteria for inclusion increases sample size which is relevant for this chapter since I will rely on existing datasets and need to ensure adequate data is available for analysis. For example, one biomedical database (UKBiobank: https://www.ukbiobank.ac.uk) with information on over half a million individuals identified 165,433 cases of self-reported insomnia(30% prevalence), compared with 598 cases of hospital inpatient diagnoses for severe insomnia (0.2% prevalence). Taken together this definition ensures that a large sample size of individuals is available for analysis.

Sleep disorders can be caused by a range of variables including physical and mental illness, as well as unhealthy behaviours. Pre-existing physical conditions, such as skin conditions (Nowowiejska et al., 2021), and mental health conditions, such as depression (Jermann et al., 2022) can make it more difficult to fall asleep or cause waking during the night. Additionally, several substances, such as caffeine (Snel & Lorist, 2011) and alcohol (Cappuccio et al., 2010), disrupt the natural balance of metabolites involved in sleep, including hormones such as cortisol (Payne & Nadel, 2004) and neurotransmitters such as GABA which has an inhibitory and sedative action (Gottesmann, 2002). Furthermore, our daytime behaviours can affect sleep quality, in particular, maintaining good sleep hygiene is important for ensuring our 24-hour body clocks are correctly calibrated to the natural cycle of light during the day (e.g., not using screens before bed)(Vhaduri & Poellabauer, 2018). Natural changes in the day-light cycle can also de-synchronise our internal body clocks, for example some individuals are more acutely sensitive to the change in daylight as the seasons change (Seasonal Affective Disorder)(Magnusson & Boivin, 2003). Similarly, jetlag is a response to the abrupt change in local time caused by long-distance air travel (Foster et al., 2013). Sleep is therefore affected by a number of variables and this may explain how sleep disorders such as insomnia have such a high prevalence.

Sleep disorders have far-reaching consequences on our physical health (Knutson et al., 2007), as well as our cognitive and emotional health (Walker, 2009). The function of sleep is not fully understood although there is evidence to suggest it serves essential roles in cellular maintenance, conserving energy and consolidating short-term memories into long-term memories (Zielinski et al., 2016). The full range of sleep disorders are accordingly associated with increased risk of physical illnesses including cardiovascular disease, immune suppression, and risk of mortality (Garbarino et al., 2016a). Sleep deprivation studies have also shown that excessively short sleep is associated with mental health outcomes including neurological disorders (Bishir et al., 2020), inattention and poor memory (Alhola & Polo-Kantola, 2007). Furthermore, difficulties initiating and maintaining sleep are also associated with an increased risk of developing mental illnesses such as depression (Fernandez-

Mendoza & Vgontzas, 2013). More generally, sleep deprivation presents a significant threat to society since it is a leading cause of poor workplace performance and accidents on the road and in the workplace (Wade, 2010).

Sleep and wellbeing are therefore central to our everyday lives and our physical and mental health. However, it is not clear whether sleep disorders causally influence our wellbeing, or whether poor wellbeing can cause sleep disorders. This is an important question to answer because it helps inform epidemiologists as to what the targets of treatment and prevention programs ought to be, as well as to predict the effects of poor sleep and wellbeing. .

2.1.2 Exploring evidence for an association between sleep and wellbeing

The body of literature on the relationship between sleep disorders and wellbeing is considerable, approaching 69,000 articles on SCOPUS and PubMed (Appendix 2.1 for search terms). These studies have been summarised in a series of recent reviews and meta-analyses. These include observational research on sleep disorders, inadequate and low-quality sleep (Garbarino et al., 2016a; Kyle et al., 2010; Reimer & Flemons, 2003; Sella et al., 2021), sleep deprivation experiments (Haack & Mullington, 2005), and trials of medicines and therapies (Boggiss et al., 2020; Krystal, 2007; Perach et al., 2019). I will review this literature, drawing out strong evidence that wellbeing and sleep disorders are associated, as well as discussing experimental evidence suggesting that sleep disorders may causally reduce wellbeing, and presenting opportunities for future causal research to add to this evidence. The current literature does not arrive at strong conclusions as to the relationship(s) between sleep disorders and wellbeing, likely because heterogenous definitions and measurements are used. Often insomnia is assessed as a measure of sleep disorder, using self-reported difficulties imitating or maintaining sleep, and wellbeing is measured using self-report happiness and life satisfaction items.

Measuring wellbeing and sleep

Researchers use a range of methods to measure sleep and wellbeing. Sleep disorders are measured using diagnoses, self-reported symptoms, and objective measures including polysomnography, a gold-standard battery of physical tests and brain activity measurement (Rundo & Downey, 2019), and accelerometery based measures of sleep activity (e.g., limb movements)(van de water et al., 2011). Wellbeing is assessed using self-report questionnaires (Diener, Lucas, et al., 2018) measuring positive affect (e.g., subjective happiness scale)(Lyubomirsky, S., & Lepper, 1995) and life satisfaction (e.g., satisfaction with life scale)(Diener et al., 1985). In the literature investigating sleep and wellbeing, it was common to assess wellbeing as part of health-related quality of life assessments intended to measure the impact that illnesses have on various domains of life, including subjective wellbeing (Bulpitt, 1997). The most popular measures have sub-scales measuring subjective wellbeing. For example, the RAND Short-Form survey 36 (SF-36)(Ware & Sherbourne, 1992) has a five-item "emotional wellbeing" sub-scale with items measuring positive affect, negative affect, life satisfaction . Similarly, the WHO Brief Quality of Life survey (WHOQOL-BREF)(World Health Organisation, 2012) has a six-item "psychological" sub-scale which measures wellbeing. Findings using these measures will be drawn upon to evaluate the evidence for a causal pathway between sleep disorders and wellbeing.

Observational research

Reviews of research using self-report questionnaires find strong evidence that sleep and wellbeing are associated. Kyle and colleagues (Kyle et al., 2010) reviewed eight studies of insomnia in otherwise healthy individuals. 7 of 8 studies showed that individuals with insomnia score lower on wellbeing than individuals who do not have insomnia. Furthermore, five of these studies showed a dose-response relationship where individuals with more severe symptoms of insomnia scored worse on wellbeing than individuals with less severe symptoms. These findings are supported by a recent meta-analysis of 23 studies investigating the impact of self-reported sleep quality on wellbeing (Sella

et al., 2021). 14 studies (61%) indicated an association between sleep quality and wellbeing with a moderate effect size of r = .21 (P<.001). However, there is little association between self-reported sleep duration and wellbeing (Jean-Louis et al., 2000). Although sleep quality and duration are slightly different from disorders such as insomnia, an analysis of self-reported insomnia symptoms and positive affect in UKBiobank participants produces a similar effect size of r = .14 (using Neale Lab summary data: https://ukbb-rg.hail.is/rg_browser/). Furthermore, reviews of other sleep disorders are similarly associated with either poor wellbeing, including excessive daytime sleepiness, sleep breathing difficulties, narcolepsy (Garbarino et al., 2016a; Reimer & Flemons, 2003), or low life satisfaction, including restless leg syndrome and a snoring partner interrupting sleep (Reimer & Flemons, 2003). Taken together, these reviews provide evidence that there is a strong and replicable association between a range of sleep disorders and wellbeing.

Observational research using objective measures of sleep duration, however, do not find a consistent association. Objective accelerometery based measures of sleep duration (Jean-Louis et al., 2000) and polysomnography measures of sleep quality (Driscoll et al., 2008) are often not associated with wellbeing. The difference between objective and self-report findings could be due to self-report inaccuracy (Lauderdale et al., 2008), or that duration and quality are only some of the ways sleep can become disordered and objective measures may not be sensitive to other ways sleep can be disordered (e.g., unhealthy sleep schedule). The difference could also be explained by a subjective factor common to both sleep and wellbeing, such as sleep satisfaction, which may drive the true effect or bias these measurements. It is therefore important for researchers to acknowledge that subjective and objective measures produce different results and the reason for this is currently unclear but they may assess different properties of sleep.

Generally, it is difficult to infer causality from observational studies (Smith & Hemani, 2014). It is possible that the association between sleep disorders and wellbeing is produced by a confounding variable which jointly produces an effect on both. For example, it is known that physical and mental

illness, such as cardiovascular or respiratory illness and depression, can cause both disordered sleep and poor wellbeing, and demographic variables such as gender, age and socioeconomic status are associated with greater prevalence and adverse effects of sleeping difficulties (Garbarino et al., 2016a). There is some evidence that in practice observational estimates are not entirely produced by confounding variables since the association between sleeping disorders and wellbeing often remains when demographic variables, physical and mental health are controlled for (Garbarino et al., 2016a; Reimer & Flemons, 2003). While these designs only control for "known" confounders, a degree of "residual confounding" often exists (Webb et al., 2020), research controlling some confounding variables suggest that the relationship between sleep disorders and wellbeing is specific and not confounded by the most relevant demographic, physical and mental health variables. However, it is difficult to control for all possible variables which could influence the outcome of an observational study.

The observational studies I have reviewed thus far have confirmed that there is a general pattern of association between sleep and wellbeing, however it remains unclear whether this represents a causal relationship. These findings set the scene for further work that tests whether there is a causal relationship, and if so, whether disordered sleep causes poor wellbeing or vice-versa. This is a critically important step because it has implications on designing treatments and prevention by helping identify the true cause, be it sleep disorders, poor wellbeing, or a third variable. I will now review some evidence from studies using experimental methods with the aim of understanding the causal pathway.

Experimental research

One strength of experimental research methods is that they investigate whether changing an exposure produces a change in the outcome, and so can provide stronger evidence of a causal pathway and infer directionality (Webb et al., 2020). Additionally, comparing an experimental condition with a control condition can help demonstrate that an effect is specific and unaffected by

confounding variables. Experimental research conducted in the present context, with sleep and wellbeing, suggests that a causal pathway does exist between them and that both variables may exert bi-directional effects on each other, but relatively little research has investigated an effect of wellbeing on sleep.

Reviews of experimental sleep studies indicate that insomnia and a lack of sleep may reduce wellbeing while good sleep may improve wellbeing. A sleep deprivation study (Haack & Mullington, 2005) reported that participants wellbeing was reduced following a period of restricted sleep (Haack & Mullington, 2005). This complements evidence that treating insomnia can increase wellbeing (Krystal, 2007; Kyle et al., 2010; Reimer & Flemons, 2003a). (Reimer & Flemons, 2003) In total, eleven distinct pharmaceutical trials for insomnia have investigated the effect on wellbeing. Four studies (36%) reported improved wellbeing following use of non-benzodiazepine sedatives including zolpidem (Hajak et al., 2002), zopiclone (Hajak et al., 1994; Hindmarch, 1995) and its Z-isomer eszopiclone (Scharf et al., 2005). However, findings are mixed because four non-benzodiazepines trials reported null effects (Goldenberg et al., 1994; Omvik et al., 2008; Walsh et al., 2000, 2007) and trials with other sedatives, including hypnotics (Hajak et al., 1995), and other classes of drugs (Morin et al., 2005), including anti-depressants (Fava et al., 2006), produced no effects on wellbeing. It is therefore possible that insomnia medication improves wellbeing by improving sleep, but these findings are inconsistent and could be caused by other mechanisms, such as an agonistic effect on GABA which can treat mental illness (Kalueff & Nutt, 2007). Two pharmaceutical trials for restless leg syndrome report similar results improving wellbeing (Perach et al., 2019). Dopamine agonists levodopa, benserazide produce no effect on wellbeing (Beneš et al., 1999) while cabergoline does (Stiasny et al., 2000). These results may also be explained by side-effects since increasing dopamine is associated with happiness (Fibiger & Phillips, 1988). Non-pharmaceutical trials give further indication that improving insomnia can improve wellbeing (Kyle et al., 2010; Perach et al., 2019). Eleven studies have specifically measured the effects that non-pharmaceutical trials have had on wellbeing. Three trials (27%) found positive effects on wellbeing as a result of Tai chi (Chan et al.,

2016) and cognitive behavioural therapy (CBT)(Dixon et al., 2006; Espie et al., 2007; Verbeek et al., 2006) programs, and a further three trials (27%) found improvements to variables related to wellbeing including stress, mindfulness and mood following similar non-pharmacological therapies (Alessi et al., 2016; Li et al., 2004; Martin et al., 2017). However, remaining studies (46%) found null effects following Tai chi (Li et al., 2004), CBT (Espie et al., 2008; Omvik et al., 2008; Soeffing et al., 2008), and following a sleep intervention program (Martin et al., 2017). Furthermore, treatments intended for insomnia do not always improve sleep quality or duration (Perach et al., 2019). This reduces the strength of evidence they provide for a causal pathway, as it is less clear that the change in the exposure (sleep disorders) actually produces the observed change in the outcome (wellbeing). Taken together, evidence from sleep deprivation, drug and therapy trials suggest that sleep disorders have the potential to reduce wellbeing but further research is required to strengthen evidence for a causal pathway.

There are fewer experimental studies of wellbeing but they indicate that good wellbeing has a protective effect on sleep, reducing the likelihood of sleep disorder. Positive psychology interventions are used to improve wellbeing (Carr et al., 2021; White et al., 2019) and some researchers use these to investigate the effects of wellbeing on sleep (Jackowska et al., 2016; Mitchell, 2010; Pahlavan & Ghasem, 2022). In particular, a meta-analysis of sleep outcomes from positive psychology gratitude exercises (Boggiss et al., 2020) found improvements to self-reported sleep quality (5 of 8) and sleep duration (1 of 2) over a range of time-frames (1-10 weeks) with longer interventions tending to report more positive effects on sleep. This suggests that improving wellbeing improves sleep. This is supported by pseudo-experimental studies (Ong, 2010) which find that poor sleep in the future is predicted by earlier wellbeing, and by variables related to wellbeing including earlier levels of optimism (Lau et al., 2017), loneliness (Hom et al., 2020), depression (Nutt et al., 2008) and bereavement (Lancel et al., 2020).

Experimental research therefore demonstrates that insomnia can have detrimental effects on wellbeing, and that good wellbeing can act as a protective factor reducing the likelihood of insomnia. While the results of insomnia treatments are inconsistent, and there is a lack of research experimentally manipulating wellbeing, overall these studies provide strong evidence that it is possible wellbeing and insomnia exert bi-directional effects on on-another.

Methodological challenges

Research investigating wellbeing and sleep disorders has faced a number of methodological challenges which should be addressed in future research. Researchers use a range of definitions for wellbeing (Diener, Lucas, et al., 2018) and sleep disorders (Garbarino et al., 2016a), so it is not clear what the underlying construct they are trying to measure is. Additionally, researchers use a range of measures for sleep and wellbeing (Ziporyn et al., 2017) and different measures intended to measure the same broad construct, such as wellbeing, may in practice capture slightly different constructs. For example, wellbeing sub-scales on the SF-36 (Ware & Sherbourne, 1992) and WHOQOL-BREF (World Health Organisation, 2012) include items which measure variables like worries (SF-36) and self-esteem (WHOQOL-BREF). Consequently these differences between measurements could produce different scores which are not entirely compatible with one-another. Researchers will also have been constrained by the ethics of manipulating wellbeing and insomnia so experimental methods are of limited use when investigating these variables. In sum, future research should provide stronger evidence by using consistent definitions, precise measures, and taking advantage of ethical methods.

Interim summary

The literature demonstrates that there is an association between sleep disorders and wellbeing. Experimental studies provide some evidence for a causal pathway implicating poor wellbeing in causing sleep disorders, but research is hampered by difficulties defining and measuring these variables, and the possibility of an effect of wellbeing on sleep is under-studied. I will use a method

known as Mendelian randomisation to overcome some of these limitations and add to the body of evidence for a causal pathway between wellbeing and sleep.

2.1.3 Investigating causal pathways using Mendelian randomisation

Mendelian randomisation (MR) is a method of estimating the causal effect of an exposure on an outcome. It is predicated on the randomised, Mendelian, process in which we inherit genetic variations from our parents. Genetic variants associated with an exposure are used as instruments to understand the effects of the exposure on an outcome. Compared with observational methods, MR can overcome traditional issues of confounding and reverse-causality, and compared with experimental methods, it can be used to investigate a wide range of variables which may be unethical to manipulate in experimental trials. I will use MR to estimate the causal effects, if any, between insomnia and wellbeing. This will add to a base of causal evidence which can be triangulated with other methods, for example a randomised control trial follow-up, to make strong causal inferences (Munafò et al., 2021). In this section I review the principles behind MR and discuss the opportunities and challenges with this approach.

Mendelian inheritance refers to the randomisation of genetic variants passed from parents to children (Smith & Hemani, 2014). Parents' alleles randomly segregate such that half their gametes carry each allele, in a process of meiosis that happens before conception. We can therefore be said to have all been assigned to genetic groups by chance. Some genetic variants pre-dispose us to certain disorders and in MR these are used as an instrument in an instrumental variable analysis to understand the effects of an exposure of interest (Burgess & Thompson, 2015). For example, the effect of insomnia on wellbeing could be investigated by comparing measures of wellbeing between two groups who either are or are not genetically predisposed to have insomnia. The underlying mechanisms of MR has been compared to randomised control trials (for a comparison see Figure 2.1). It is owing to this Mendelian inheritance that MR reduces bias from traditional limitations of epidemiological research including confounding and reverse-causality. Our DNA sequence is

immutable so it is not possible that the outcome has a traditional reverse-causal effect on the genetic instruments for the exposure. MR helps overcome traditional confounding issues in the same manner that a randomised control trial does, by observing the effects that a difference in levels of the exposure between groups has on an outcome. Only in MR, this change in levels of the exposure is the result of randomised genetic variants, rather than randomised treatments.

Additionally, since no manipulation is made it can be used to investigate a wide range of phenomena which may ordinarily be considered unethical to subject participants to, such as poor wellbeing and sleep disruption.

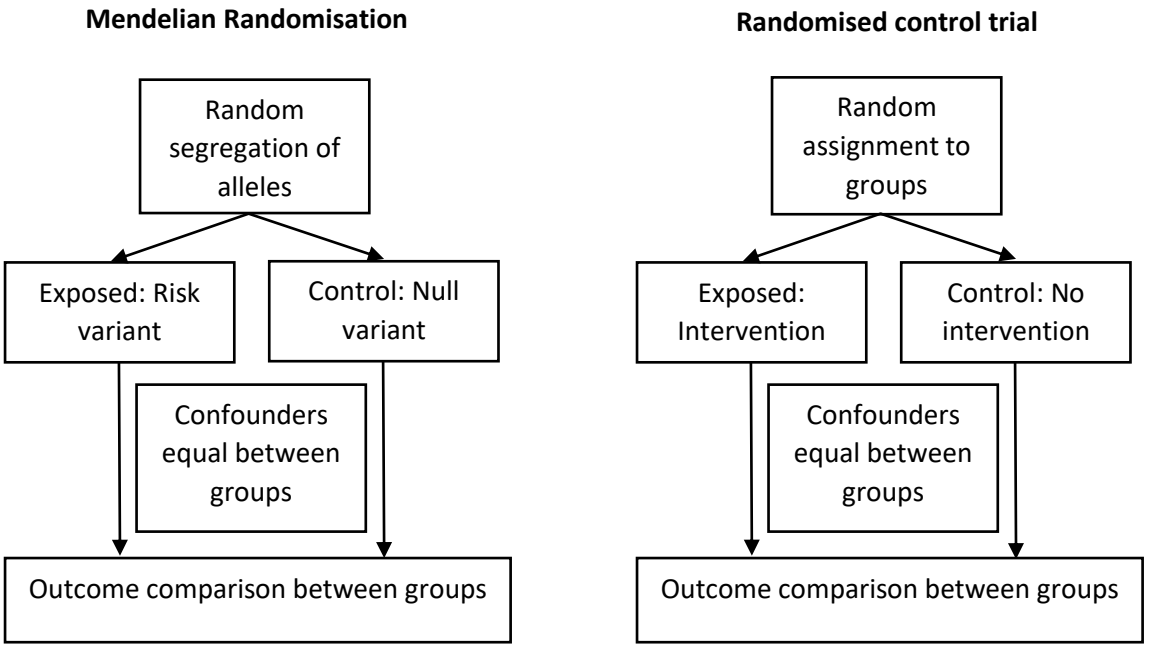


Figure 2.1 Mendelian randomisation (left) works in a manner similar to a randomised control trial (right) with an analogous assignment to control and exposure groups and outcome comparison at the end. This diagram shows a single risk variant for simplicity but in practice the exposure group is often investigated for multiple variants.

2.1.4 Opportunities and challenges in using MR

There are a variety of opportunities for using MR, and advantages compared with traditional methods of observational research. I review opportunities including using large-scale genetic studies (genome-wide association studies) to identify variants robustly associated with exposures of interest, combining multiple variants to produce strong instruments (inverse-variance weighted

estimation), as well as methods of comparing a broad range of phenotypes across different GWAS (two-sample MR), and inferring the direction of causal effects (bi-directional MR). There are also challenges to applying MR in a valid manner and I will review three core instrumental variable assumptions which have to be met for valid causal inference, explain the consequences of not meeting them, and explain how a combination of sensitivity tests can be used to investigate whether these have been met.

Genome-wide association studies (GWAS) are performed to identify genetic variants which are associated with phenotypes so that they can be used as instruments. It is common to use single nucleotide polymorphisms (SNPs) as genetic instruments in MR and these comprise a single base-pair difference in an individual's DNA at a specific locus. Complex traits like sleep (Jansen et al., 2019) and wellbeing (Okbay et al., 2016b) are often influenced by many SNPs of small effect (i.e. they are highly polygenic traits) so multiple variants (SNPs) are combined to observe sufficiently strong genetic influences. The Inverse-Variance Weighted method of MR (IVW)(Burgess et al., 2013) allows researchers to meta-analyse multiple genetic variants and obtain an overall effect estimate. The overall effect of the variants is estimated using a meta-analysis method which returns a single effect estimate describing their average effects on the outcome.

Two-sample MR (2SMR) gives the opportunity to compare a wide range of variables. 2SMR is a method of MR where the researcher compares the instrument-exposure and instrument-outcome associations across two independent GWAS, assuming that the same individuals did not participate in both (Lawlor, 2016). This is the most popular method of performing MR since often a single genetically-informed sample does not contain information on both the exposure and outcome a researcher may want to investigate with large enough sample size. This is especially useful for research comparing different types of variables since consortia providing one type of data (e.g., physical) do not necessarily provide another (e.g., psychological). 2SMR is therefore a valuable tool which allows researchers to perform analyses using datasets which ordinarily would not be

comparable using traditional one-sample methods of MR. The basic logic of 2SMR is that you take the genetic variants associated with the exposure from the first GWAS, and then you look at the results for these specific genetic variants in the second GWAS on your outcome. This is possible because it is standard for GWAS publications to include summary information on all variants included in their study.

Bi-directional MR (BDMR)(Richmond & Davey Smith, 2019) gives the opportunity to infer directionality, particularly where an outcome-exposure effect is likely. Uni-directional MR involves obtaining a single effect estimate using genetic instruments for a single exposure on an outcome, whereas BDMR involves obtaining a second effect estimate in the reverse direction using genetic instruments for the outcome on the exposure. This is valuable because it is possible to determine the direction of a causal effect by comparing effect estimates in both directions.

A key challenge in MR is selecting valid genetic instruments. This is important because causal effect estimation can be biased if genetic instruments are not properly selected (Haycock et al., 2016; Smith & Hemani, 2014). Three core instrumental variable assumptions (Figure 2.2) will now be outlined, along with methods of assessing whether instrumental variables meet these, commonly known as sensitivity tests (Burgess et al., 2017).

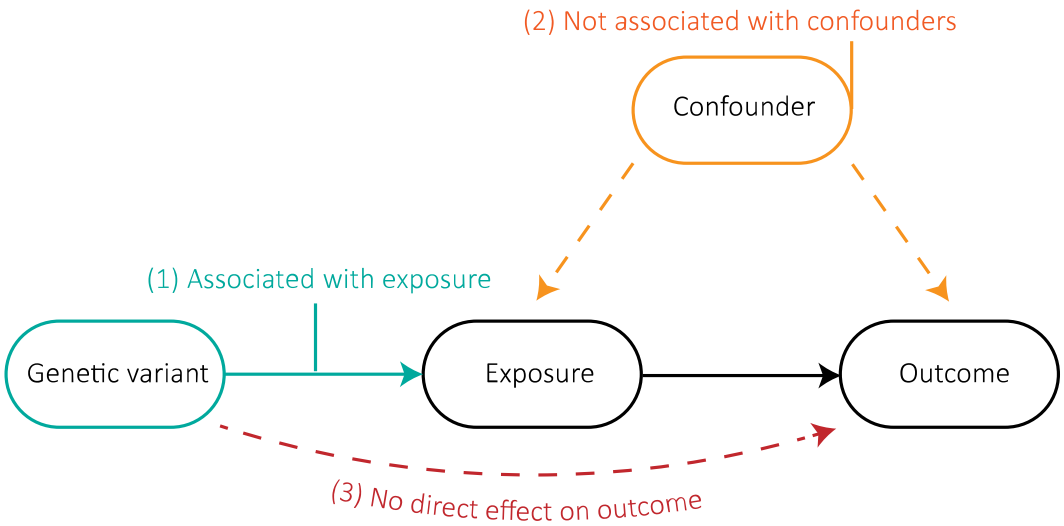


Figure 2.2 In Mendelian randomisation genetic variants are used as instruments to manipulate an exposure, which in turn is investigated for an effect on the outcome. However, there are several assumptions that genetic variants are (1) associated with the exposure, (2) not associated with a confounder of the exposure-outcome association, and (3) do not affect the outcome through pathways other than via the exposure.

(1) Genetic instruments are associated with the outcome

Researchers must use "strong" instruments which are robustly associated with the exposure and avoid "weak" instruments which are not (Burgess & Thompson, 2011). Strong instruments are typically defined as being associated with an exposure at genome wide significance ($P<5x10^{-08}$). Weak instruments do not robustly predict the exposure and are likely to bias the effect estimate towards the confounded observational estimate. Instruments need to contribute a strong enough signal to overcome this risk of bias and weak instruments are identified as those contributing more noise than signal to an analysis. An F statistic is often calculated as a ratio of signal to noise, where a value of 10 or more is taken as a rule of thumb indication of lower, but not eliminated, risk of bias (Burgess & Thompson, 2011). Weak genetic associations are particularly problematic in psychology because variables like wellbeing are difficult to measure, have heterogenous measures, and individual variants have smaller effect sizes, so it can be more difficult to identify associated genetic variants (the challenges in identifying genetic contributions to wellbeing is further discussed in Bartels, 2015, and the relevance to GWAS in Okbay et al., 2016).

Measurement error is a related concept which contributes noise to an analysis reducing the "quality" of instruments. When calculating effect estimates using regression methods, which I will rely on in this chapter as part of my analysis, it is assumed that instrument-exposure associations are estimated accurately, without substantial measurement error (Bowden, del Greco, et al., 2016). Measurement error can dilute the regression and risk of bias towards the null, so this is often measured using the I^2GX statistic (Bowden, del Greco, et al., 2016) where a value above 90% indicates low measurement error.

Selecting strong instruments is therefore an important process to prevent weak instrument bias and the strength of instruments must be assessed using F statistics and I^2GX statistics.

(2) Genetic instruments share no common cause with the outcome

It is assumed that any chances observed in the outcome are due to changes in the exposure and its instruments. However, one case where this does not hold true is where there is a confounding variable which affects both the distribution of genetic variants and the outcome (Davies et al., 2018). This would link the instruments with the outcome in an invalid manner and can occur as a result of sampling biases and statistical corrections. For example, "population stratification" occurs when demographic differences, such as ancestry and culture, affect both the frequency of genetic variants as well as the outcome variable of interest. This can be controlled for by selecting datasets which best represent the general population and comparing samples with similar ancestry to ensure similar distribution of genetic variants.

(3) Genetic instruments do NOT have a direct effect on the outcome other than through the exposure

"Vertical pleiotropy" is the essence of MR and describes the association of a genetic variant for an exposure with other variables which represent a valid causal pathway. "Horizontal pleiotropy" (Davey Smith & Hemani, 2014a) however, refers to a genetic instrument which is associated with a third variable which acts on the outcome independent of the exposure. Horizontal pleiotropy, often simply referred to as "pleiotropy", is problematic. Pleiotropy biases the effect estimate with the effect of another variable on the with the outcome, rather than measuring purely the effect of exposure on outcome.

The first defence against pleiotropy is a functional biological knowledge of the genetic instruments. It is important to identify the biological pathways through which genetic instruments manipulate the exposure (and only the exposure), however this becomes increasingly difficult when variants act

through complex pathways (e.g., on wellbeing)(Røysamb & Nes, 2018) and researchers investigate multiple genetic variants simultaneously (Smith & Hemani, 2014).

The second defence is sensitivity testing after the main analysis has been conducted (Burgess et al., 2017). A battery of indicators and tests are used to formally identify signs of pleiotropy and measure their impact. Sensitivity tests often take advantage of multiple genetic variants, so it is additionally important for researchers to use multiple variant methods of MR. The estimates produced by different genetic instruments are inspected for indications of pleiotropy. When different instruments produce greatly different effect sizes for exposure on outcome, this is taken as an indication different variants act on the exposure through different and potentially pleiotropic pathways (Hemani, Bowden, et al., 2018). Cochrane's Q statistic for heterogeneity is used to quantify differences in effect sizes across instruments where a statistically significant *P* value can identify substantial heterogeneity which warrants further investigation (Burgess et al., 2017). The Q statistic is used as a first warning sign of global pleiotropy but does not give information on whether pleiotropy biases the effect estimate in a certain direction. For example, pleiotropy which biases the effect estimate away from the null can cause over-estimation of the causal effect estimate, while bias towards the null may cause under-estimation. Directional pleiotropy is investigated by comparing effect estimation methods which make different assumptions about how pleiotropy might affect the effect estimate.

The MR Egger (Bowden et al., 2015) method helps identify and measure directional effects of pleiotropy. Valid instrumental variables act on the outcome purely through an exposure; some estimators, such as IVW, assume this is true and that the average effect of pleiotropy is zero, whereas the MR Egger relaxes this assumption that all instruments are valid. The MR Egger relies on the assumption that pleiotropic effects of instruments are independent of instrument strength (InSIDE)(Burgess et al., 2017). This assumption is weaker and easier to meet which means that MR Egger estimates are more robust to produce accurate causal effect estimates with invalid

instruments, and when this assumption is met stronger instruments should give more reliable estimates of the true causal effect. An overall effect estimate is obtained by accounting for the average pleiotropic effect, using an unfixed intercept term, to produce a dose-response relationship describing the effect of the exposure on the outcome through the true causal pathway. The MR Egger intercept term can therefore be used as a measure of bias from directional pleiotropy, and the overall model can be compared for compatibility with the IVW estimate. Agreement among estimators with different assumptions is evidence of a reliable effect estimate and a large pleiotropic influence is less likely (Lawlor, 2016).

Two other estimators, weighted median (Bowden, Davey Smith, et al., 2016) and weighted mode (Hartwig, F. P., Davey Smith, G. & Bowden, 2017), are also commonly compared with the IVW estimate as further indications that pleiotropy is less likely. As their names suggest these methods calculate the average exposure-outcome effect by taking the median or mode effects (while considering the strength of instruments). The median estimator assumes that 50% of the instrument weight comes from valid instruments which means that it is robust even when a minority of instruments act through pleiotropic pathways (minority directional pleiotropy). The mode estimator assumes the largest number of similar causal effect estimates comes from valid instruments and is consistent with the true effect, even when most instruments are pleiotropic (majority directional pleiotropy). Considerably different effect estimates from this method may indicate that instruments in the original IVW estimate act through pleiotropic pathways.

Managing pleiotropy is therefore extremely important to obtaining an accurate effect estimate, but it is difficult to identify so researchers use a range of methods to detect various warning signs of pleiotropy

Lastly, a final invalid pathway is the direct action of a genetic variant on an outcome. If a genetic instrument is directly associated with the outcome an effect estimate would not describe the effect of the exposure on the outcome but rather the direct association of a genetic instrument. It is

therefore important to select instruments which are not strongly associated with the outcome and this is assessed by comparing the instrument-outcome and instrument-exposure associations in a Steiger test (Hemani, Tilling, et al., 2017) where instruments which are statistically significantly better predictors of the exposure are considered more valid.

Example

For an example of an MR study where instruments are used for wellbeing (Wootton et al., 2018), please see Figure 2.3.

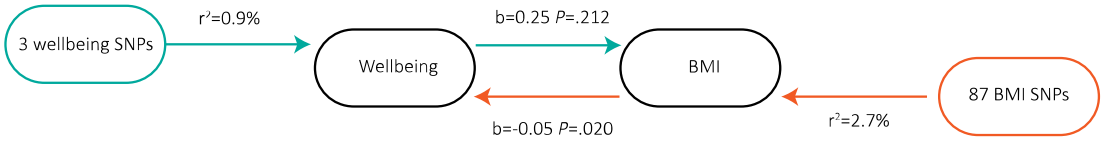


Figure 2.3 As an example to demonstrate how MR works, I present a previous study seeking to understand the effect of wellbeing on Body Mass Index (BMI), and vice-versa (Wootton et al., 2018). In this study a two-sample MR design was used to compare measures of subjective wellbeing and BMI in different genomics datasets. Genetic variants robustly associated with wellbeing and BMI at genome wide significance ($P<5x10^{-8}$) were selected. Numerous strong instruments were selected for BMI (n=87, r^2=2.7% variance explained) but only a few, relatively weak instruments were available for wellbeing (n=3, r^2=0.9% variance explained). Measurement information and variants were entered into a bi-directional IVW analysis which found a causal effect suggesting higher BMI causally reduces wellbeing (b=-0.05, *P*=0.020), but there was no strong evidence for a causal effect of wellbeing on BMI (b=0.25, *P*=.212). Sensitivity testing produced comparable estimates from MR Egger, weighted median and weighted mode estimators, indicating that a significant bias from directional horizontal pleiotropy is less likely. Part of the authors' conclusion is that the strength of currently available instruments for wellbeing present challenges to making strong conclusions about the causal effect of wellbeing in MR analyses.

2.1.5 Understanding the causal pathway between sleep and wellbeing using MR

I will use MR to investigate the causal pathway between difficulties initiating or maintaining sleep (insomnia) and subjective wellbeing. Insomnia is only one way in which sleep can be disrupted but I focus on this because it is the most common sleeping disorder and has clear effects reducing the quantity and quality of sleep (Fernandez-Mendoza & Vgontzas, 2013). Additionally, much more data has been collected on insomnia which ensures good instrumental variables and measures are available for analysis.

The causal pathway between wellbeing and insomnia could exist in various states and I investigate these as four distinct hypotheses (Haack & Mullington, 2005). Treating insomnia has been found to improve wellbeing (Krystal, 2007) so the first hypothesis I will investigate is that insomnia causes poor wellbeing. Second, it has been shown that wellbeing-improving interventions also improve insomnia symptoms (White et al., 2019), so it is possible that poor wellbeing may instead cause insomnia. Third, wellbeing and insomnia may exert bi-directional causal effects on eachother. This is based on observations that earlier levels of wellbeing and insomnia predict future levels of each-other (Ong, 2010). Lastly, it is possible that the association between wellbeing and insomnia does not represent a causal relationship (i.e., the null hypothesis).

To date three MR studies have investigated these hypotheses. One study (O'Loughlin et al., 2021) used self-report measures of diurnal preference and wellbeing in UKBiobank. They found that being a "morning person" causally increased happiness by 6%. Although this is not specific to insomnia it adds to evidence that sleep traits exert causal effects on the positive affect component of wellbeing.

Two recent MR studies specifically investigated wellbeing and insomnia (Jansen et al., 2019; Zhou et al., 2021). The first (Jansen et al., 2019) studied the effects of self-reported insomnia on a range of variables related to, and including, wellbeing. While not the primary objective of their paper, the authors present bi-directional effect estimates in their supplementary materials suggesting that insomnia causally reduces wellbeing, and that good wellbeing also causally reduces insomnia symptoms. This indicates a bi-directional effect, but interrogation by sensitivity testing is required to gage the reliability of this estimate since, for example, they select more weakly associated instruments for wellbeing ($P<5x10^{-05}$) than is convention ($P<5x10^{-08}$). The present study builds upon this study, contributing interrogation and sensitivity testing for this relationship, and since this was conducted, a second and more recent study has been performed finding more evidence of a bi-directional effect (Zhou et al., 2021). The advantage of their study is that they draw on a higher-powered discovery GWAS (Turley et al., 2018) to identify stronger instruments for wellbeing. This

contribution will be relevant for discussing the impact of weak instruments on the strength of evidence in MR.

The present study acts as a base analysis providing a foundation for expansion in the next chapter (3) which contributes a novel network dataset building upon evidence for a causal pathway between insomnia and wellbeing. I conducted a bi-directional MR study intended to corroborate previous effect estimates by way of investigating four possible hypotheses:

H_1: Insomnia causes poor wellbeing

H_2: Poor wellbeing causes insomnia

H_3: Wellbeing and insomnia exert bi-directional causal effects

H_0: No causal pathway exists between wellbeing and insomnia

2.2 Methods

In the methods section I start by defining wellbeing, explaining how I selected genomics consortia to source data and measurements consistent with these definitions, and identified variants associated with wellbeing to act as genetic instruments. I then provide details for insomnia and outline the statistical procedures for using bi-directional two-sample MR methods to estimate the effect of wellbeing on insomnia and vice-versa.

Throughout the process of obtaining data and performing MR I used the MR Base "TwoSampleMR" package for *R* (https://mrcieu.github.io/TwoSampleMR/)(Hemani, Zheng, et al., 2018a). MR Base www.mrbase.org is a database and analysis platform for MR consisting of analysis software and a GWAS catalogue. The TwoSampleMR package allows researchers to obtain MR estimates from summaries of GWAS studies. Its GWAS catalogue has information from over 40,000 GWAS summary datasets and allows researchers to explore and compare a wide range of phenotypes.

2.2.1 Wellbeing

Wellbeing was defined as the presence of positive affect and life satisfaction and the absence of negative affect. This is consistent with the definition of subjective wellbeing (Diener, 1984) that most researchers use when investigating wellbeing. The ideal measure of wellbeing for the present study would therefore arrive at an overall score by asking individuals to self-report wellbeing across the three domains.

In order to source data on wellbeing consistent with this definition, I searched MR Base to discover and obtain relevant GWAS summary datasets(Hemani, Zheng, et al., 2018). I selected the largest sample size dataset with information on wellbeing in MR Base (Okbay et al., 2016b). This GWAS involved a meta-analysis of 298,420 individuals of mixed genders and European ancestry participating across 49 cohorts in the Social Science Genetic Association Consortium (SSGAC: www.thessgac.org).

I selected a measure for wellbeing from this GWAS consistent with my definition. This measure is a meta-analysis of responses on questionnaires used across different GWAS samples. Items assessing positive affect, life satisfaction and the absence of negative affect were extracted from a range of wellbeing questionnaires including the Satisfaction With Life Scale (Diener et al., 1985) and Subjective Happiness Scale (Lyubomirsky, S., & Lepper, 1995). Individuals' wellbeing was scored relative to the rest of their cohort, measured using standard deviations (outliers excluded: +/-3 SD). The resulting composite therefore measures wellbeing in the three different domains, and was selected over datasets with less specific measures (e.g., including mental illness)(Baselmans et al., 2019) and measures of one aspect of wellbeing, such as positive affect (Elsworth et al., 2019).

Genetic instruments for wellbeing were selected from robustly associated genetic variants ($p<5x10^{-8}$) in GWAS (genotyped variants=2,264,177)(Okbay et al., 2016b). Three variants (Appendix 2.2) reached genome-wide significance for wellbeing. Together these variants have been found to explain 0.9% of the variation in measurements of subjective wellbeing (Okbay et al., 2016b). This is a

well-known GWAS and its data on variant-phenotype associations is often used in wellbeing research (Baselmans et al., 2019; Turley et al., 2018; Wootton et al., 2018).

2.2.2 Insomnia

Insomnia was defined as difficulties initiating or maintaining sleep. Ideally, this definition would include only symptoms accurately diagnosed by a clinician (e.g., by ICD-10 criteria)(World Health Organisation, 2018) but the under-diagnosis of insomnia highlighted in the introduction presented a challenge to obtaining large enough sample sizes for MR analysis.

The largest GWAS on insomnia (Jansen et al., 2019) combines information from two large biomedical datasets, UKBiobank which contains information on the sleeping habits of 462,341 mixed gender individuals of European ancestry (www.ukbiobank.ac.uk), and 23AndMe which contains similar information on 944,477 individuals (www.23andme.com). Due to access restrictions measurement information was only available for the UKBiobank portion of this GWAS (Elsworth et al., 2019).

This GWAS includes a self-report measure of insomnia. Participants responded to the question "Do you have trouble falling asleep at night or do you wake up in the middle of the night?" with a response of "never/rarely", "sometimes" or "usually". Responses were scored relative to the rest of the sample on a discrete scale from never-to-usually experiences difficulties, using standard deviation units (outliers excluded: +/- 3 SD). This measure therefore captures difficulties initiating and maintaining sleep consistent with my definition of insomnia. This GWAS was selected over another dataset including a measure of sleep duration (Dashti et al., 2019) to capture difficulties impacting sleep quality as well (e.g., waking during the night).

Genetic instruments were selected from the UKBiobank and 23AndMe dataset (Jansen et al., 2019). 116 variants were robustly associated with self-reported difficulties initiating and maintaining sleep (variants genotyped = 9,851,867). These variants are implicated in circadian rhythm control and have been previously found to explain as much as 2.6% of the variance in measurements of self-reported

insomnia (Jansen et al., 2019). Further information on the instruments in my study are given in Appendix 2.2.

2.2.3 Data preparation

Two sample MR uses SNP-exposure estimates from a GWAS of the exposure and SNP-outcome effects from a separate GWAS of the outcome. Additional steps have to be taken to ensure data is comparable (Lawlor, 2016), including ensuring that variants are labelled the same way (harmonisation) and information for instruments is available in both GWAS.

Harmonisation is an essential process in performing MR since it relies on identifying and comparing variants which produce effects on phenotypes (affect alleles). Variants can either be labelled on the forwards strand or the reverse strand, and care must be taken to ensure that GWAS label variants in the same way otherwise it can be unclear what variants are being referred to. Harmonisation is the process of ensuring all variants are labelled in the same way (on the forward strand). In cases where the method of labelling is not known, the orientation of labels can be estimated by comparing the minor effect allele frequencies. In the present study harmonisation was used where minor affect alleles with low frequencies (R=<0.3) were aligned and in cases of ambiguity variants were excluded (e.g., palindromic variants).

The sample of genetic variants genotyped in a GWAS varies and one GWAS may contain information on a variant which another did not test. This information is necessary for comparing the effects of variants across GWAS, so missing variants in one GWAS are substituted with "proxy" variants which are in linkage disequilibrium so occur very often with the other (Katikireddi et al., 2018). Proxies have similar genotype-phenotype associations and are used as instruments in place of missing variants. The TwoSampleMR package for R (Hemani, Zheng, et al., 2018a) includes a function which searches the 1000 genomes project for proxies (www.internationalgenome.org). In the present study 70 variants in the insomnia GWAS were not available in the wellbeing GWAS, but 34 highly correlated (R^2=.8) variants were selected in place of them. No proxies were found for the other 36

variants, so the final instrument for insomnia included 80 variants (with 34 of these being proxies). Proxies were not required for wellbeing.

2.2.4 Main effect estimation

The main effect estimate in MR is obtained by comparing the ratio of instrument-exposure and instrument-outcome associations. I will use the two-sample MR method, so r the instrument-exposure association are taken from one GWAS and instrument-outcome association from another (Lawlor et al., 2008). It is also important to perform a power calculation to ensure an analysis is sufficiently powered to detect a range of plausible effect estimates.

When using one genetic variant as an instrument the Wald method of estimation is used. A Wald ratio effect estimate is calculated by dividing the instrument-outcome association by the instrument-exposure association (Lawlor et al., 2008):

$$effect = \frac{r_{instrument-outcome}}{r_{instrument-exposure}}$$

Since I am combining multiple genetic variants to use as instruments their effects will be meta-analysed to arrive at an overall effect estimate. Wald ratios will be calculated for each individual instrument and effect estimates will be regressed using the IVW method (Burgess et al., 2013). IVW will be used as the main method of estimation.

I conducted a power calculation to estimate power to detect a range of effects over a range of effect sizes using the MR power calculator (https://shiny.cnsgenomics.com/mRnd/)(Brion et al., 2013). Various parameters are required for power estimation including error rate (*P*=.05), the variance in measurements for wellbeing (1SD) and insomnia (1SD), as well as an observational estimate for the association between them (r=.14, estimated for UKBiobank in the introduction). Additional parameters were input to estimate the power to detect effects for insomnia on wellbeing and wellbeing on insomnia in bi-directional MR. Analysis revealed that the present design would have

power to detect small effects of insomnia on wellbeing (b>0.05). Given the sample size of outcome GWAS (n=298,420) and the variance explained by insomnia instruments (r^2=2.6%), this gives power over a range of effect sizes: b=0.1 (>99%), b=0.05 (99%), b=0.025 (60%). Analysis in the reverse direction suggests similar power to detect small effects of wellbeing on insomnia (b>0.05). In the direction of wellbeing to insomnia, the sample size of the outcome GWAS (n=462,341) and variance explained by wellbeing instruments (r^2=0.9%) gives varying power at a range of effect sizes: b=0.1 (>99%), b=0.05 (93%), b=0.025 (40%). The true effects could be larger, so would require even less power to detect, but I tested conservative effect estimates intended to represent the minimum meaningful effect sizes. The units of wellbeing and insomnia are the same, z-scores, so these effect sizes correspond to a 100% change in the exposure causing changes in the outcome of 10%, 5%, 2.5% respectively. Detecting very small effects (b<0.025) would be unlikely, given both directions have low chances of distinguishing them from random variance (60% and 40% respectively).

2.2.5 Sensitivity testing

Sensitivity testing was conducted to investigate whether the genetic instruments likely met the valid instrument assumptions. The three most important assumptions outlined in the introduction were ensuring the instrumental variables: (1) are associated with the exposure, (2) but not confounders (3) and did not act on the outcome through a pathway other than the exposure. The first assumption will be tested by assessing the strength of instruments with respect to their signal-to-noise F ratios (Burgess & Thompson, 2011) and I^2GX regression dilution due to measurement error (Bowden, del Greco, et al., 2016). The second and third assumptions will be tested by identifying indicators of pleiotropy. The Q statistic (Hemani, Bowden, et al., 2018) will be used to indicate heterogeneity and the first sign of pleiotropy. This will be followed up with an MR Egger intercept test as a measure of directional pleiotropy, and three complementary methods of MR effect estimation will be compared including the MR Egger (Bowden, Fabiola Del Greco, et al., 2016), weighted median and mode based

estimators (Bowden, Davey Smith, et al., 2016). Steiger testing will be used to compare per-SNP effect estimates on both exposure and outcome (Hemani, Tilling, et al., 2017).

2.3 Results

I will examine evidence for an effect of insomnia on wellbeing, and vice-versa, drawing on results from main effect estimation and sensitivity testing. The main results are presented in Table 2.2 and all results and code for analysis can be found on the Open Science Framework repository (https://osf.io/43dw6/). Additional plots can be found in Appendix 2.2 (for "leave-one-out" and "funnel" plots). In this section I will explain my results and in the discussion I will interpret them in relation to my three hypotheses.

Table 2.2 Main MR effect estimation

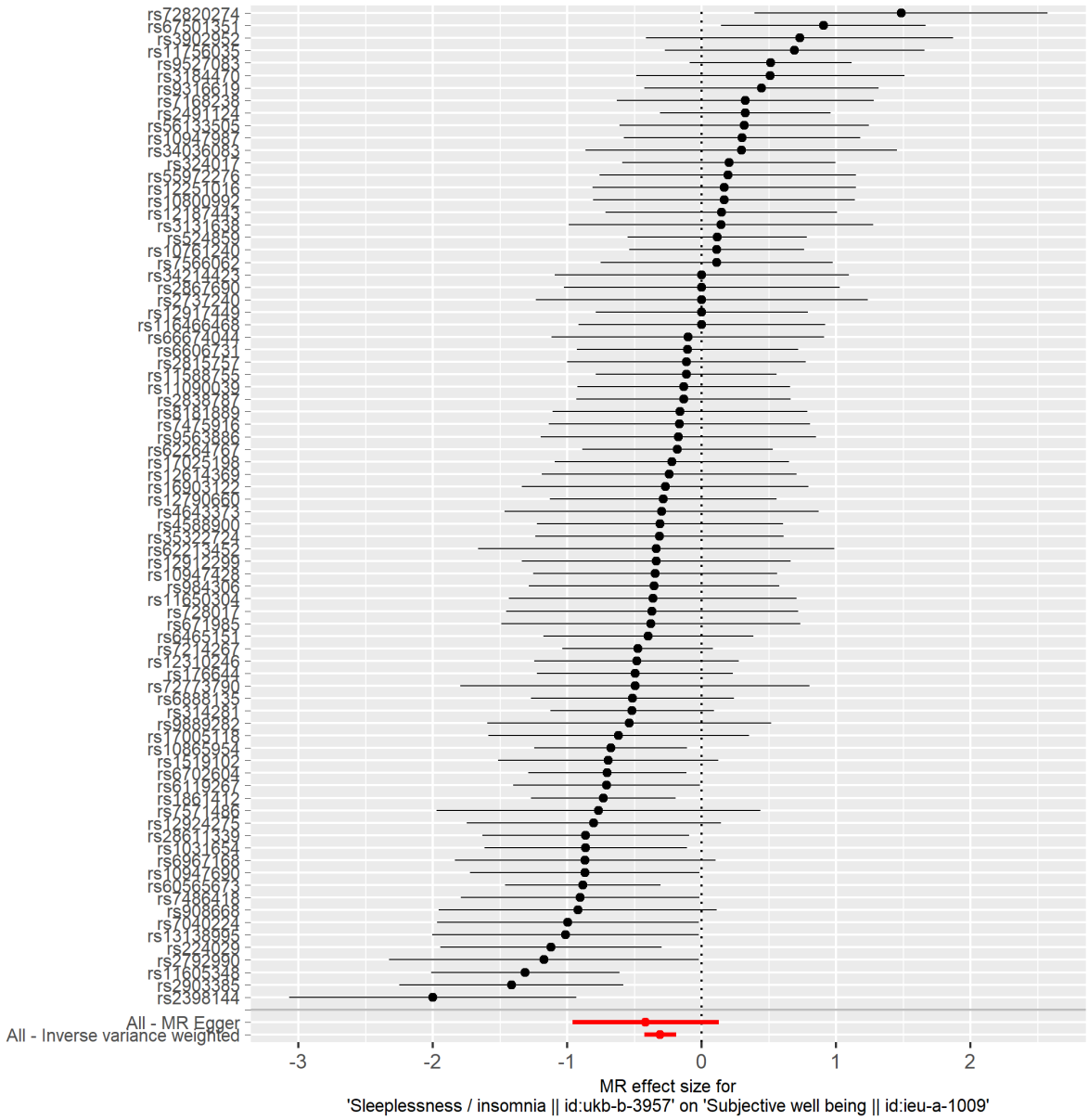
	Method	SNPs	B	CI	se	P
Effect of insomnia on wellbeing	**Main effect (IVW)**	80	-0.309	[-0.320 , -0.298]	0.06	$2.66x10^{-07}$
	MR Egger	80	-0.418	[-0.469 , -0.366]	0.278	0.137
	Egger intercept	80	-0.001		0.001	0.454
	Weighted median	80	-0.298	[-0.311 , -0.284]	0.08	$2.00x10^{-04}$
	Weighted mode	80	-0.282	[-0.321 , -0.242]	0.199	0.159
Effect of wellbeing on insomnia	**Main effect (IVW)**	3	-0.068	[-0.165 , 0.029]	0.058	0.239
	MR Egger	3	-0.318	[-0.806 , 0.171]	0.29	0.471
	Egger intercept	3	0.013		0.012	0.479
	Weighted median	3	-0.082	[-0.192 , 0.029]	0.066	0.218
	Weighted mode	3	-0.094	[-0.226 , 0.038]	0.077	0.343

Note: Confidence intervals (CI) indicate the 95% confidence interval where the true value lies and smaller intervals indicate greater certainty.

2.3.1 Effect of insomnia on wellbeing

The present study finds evidence to support a large causal effect where insomnia reduces wellbeing. IVW estimation (Figure 2.4) revealed a risk effect (b=-0.31). The confidence intervals for this effect are small and that this effect estimate is reliable (Sterne, 2001), indicating great certainty that the effect size lies within a small range (b=0.29-0.32). This is further supported by a highly statistically significant *P* value of $2.66x10^{-07}$.

Sensitivity testing supports this inference by finding little indication that this estimate is unreliable. Instruments (n=80) were robust predictors of insomnia (assumption 1) since in the present study they predicted a sizable proportion of variance in the outcome measurement of insomnia (r^2 = 0.40%). Furthermore, they showed good signal-to-noise ratio (mean F=23.0) and showed little measurement error (I^2GX=0.956) which may otherwise have diluted regression analysis. Furthermore, instruments relationships appear to predominately act on wellbeing through the exposure insomnia, rather than confounding or pleiotropic pathways through other variables (assumptions 2 and 3). Effect estimates initially showed some heterogeneity among instruments (Q=124, *P*=.001), so MR Egger, mode and median estimators were used to identify signs of directional pleiotropy. The MR Egger intercept (α=0.001) did not significantly differ from suggesting a low risk of bias from directional horizontal pleiotropy. Effect estimates produced by different MR Egger, weighted mode and median based estimators (Figure 2.4) found effects with consistent magnitudes (b=-0.28 to -0.42), and two estimators achieved or approached statistical significance (P_{median}=2.0x10^{-04}; P_{Egger}=0.137). Lastly, instruments did not have a substantial association with the outcome in Stieger tests since 96% of the instruments for insomnia passed and better explained the exposure (mean r^2= 4.98x10-05) than outcome (mean r^2= 6.92x10-06, Steiger *P*=.098, n fail=3).


All - MR Egger
All - Inverse variance weighted
-3
-2
-1
0
1
2
MR effect size for
'Sleeplessness / insomnia || id:ukb-b-3957' on 'Subjective well being || id:ieu-a-1009'

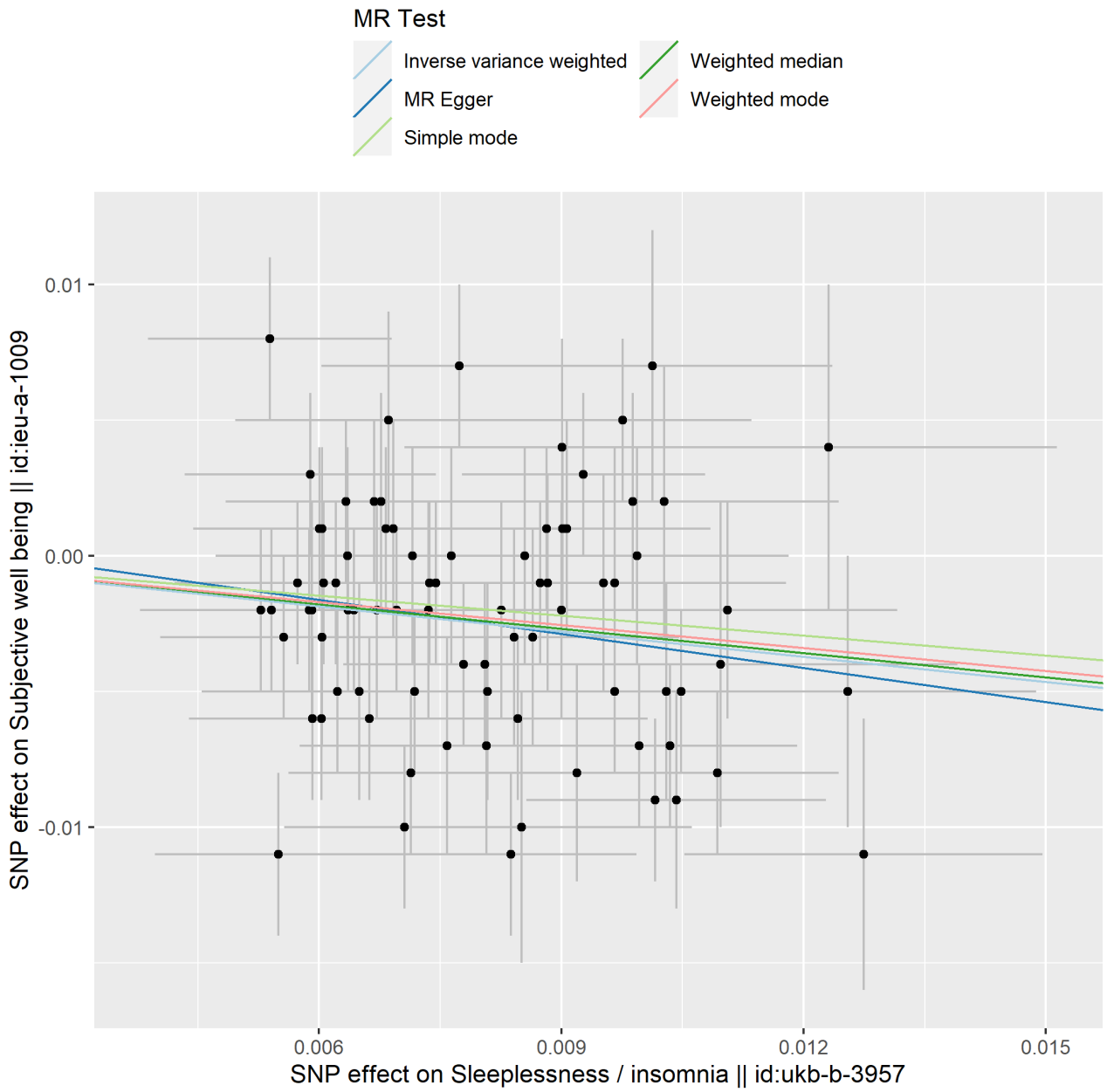


Figure 2.4 Insomnia appears to exert a causal effect reducing wellbeing. Forest (top) and scatter (bottom) plots show the effect estimates for instrumental variables (n=80) used to predict the effect of insomnia on wellbeing. The forest plot of single per-instrument effects indicates that instruments produce a range of effect estimates and IVW meta-analysis, indicated in red, suggests that the average effect is negative and reaches statistical significance (b=-0.3, P=4.45x10^{-07}). The scatter plot indicates that different estimators (Egger, median, mode) generally agree on the direction and magnitude of the effect. Confidence intervals indicate the 95% confidence that the true effect lies within the ranges of values bound by the lines indicating the intervals around each point.

2.3.2 Effect of wellbeing on insomnia

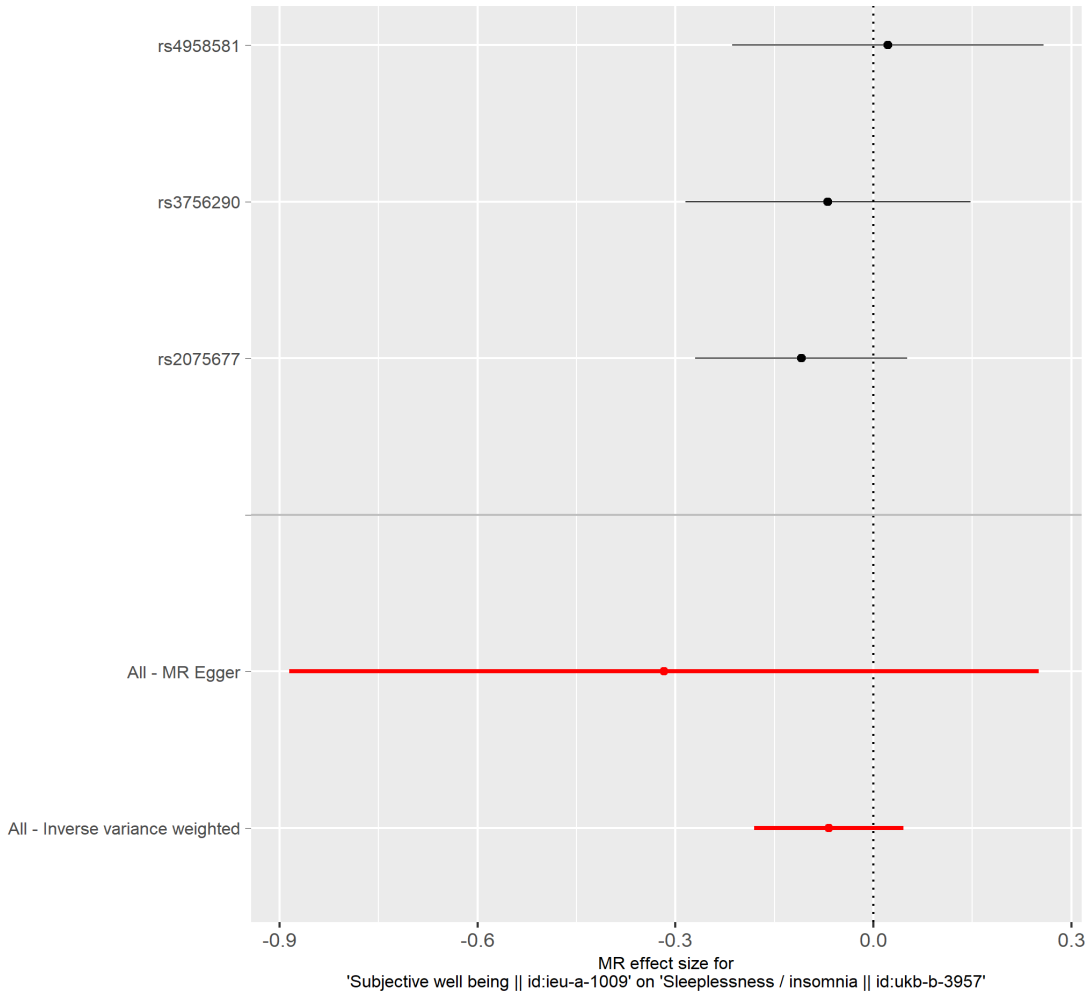
There was some indication that poor wellbeing causes insomnia. IVW estimation (Figure 2.5) revealed a small effect where good wellbeing reduces the risk of insomnia (b=-0.068) although there

was little certainty, reflected in a large confidence interval (b=-0.165 - 0.0293), and the test did not approach statistical significance (*P*=.239).

Sensitivity testing indicates that the combined strength of wellbeing instruments may not have been sufficient to detect small effects (assumption 1). On the one hand, in my study wellbeing instruments (n=3) explained a small proportion of the variance in wellbeing (r^2 = 0.02%). On the other hand, the mean instrument showed good signal-to-noise ratio (mean F=20.7) and did not present substantial dilution to IVW and Egger regression by way of measurement error (I^2GX=0.959). However, on balance, a good signal-to-noise ratio is not a guarantee that instruments are sufficiently strong (Burgess & Thompson, 2011), and MR estimation relies on sufficiently strong instruments to detect small effects (Pierce et al., 2011). There was little evidence of pleiotropy (assumption 2) since instruments produced homogenous effect estimates (Q=0.8, *P*=.667), no substantial evidence of directional pleiotropy (MR Egger intercept = 0.013, *P*=0.471), and follow-up MR Egger, mode and median estimations produced compatible, albeit not statistically significant, effects of wellbeing reducing insomnia with a magnitude between b=-0.08 and b=-0.32 (Figure 2.5). However, sensitivity tests rely on heterogenous effects to identify pleiotropy, and the restrictive number of SNPs will have reduced the power and ability to detect signs of pleiotropy (Burgess et al., 2017). Lastly, instruments better explained the exposure (mean r^2= $6.96x10^{-05}$) than outcome (mean r^2=$1.58x10^{-06}$, Steiger *P*=.003, n failed = 0). Overall, there are some indications that instruments for wellbeing were weak, and the impact this may have had will be discussed along with methods of improving instrument strength.

It is worth recounting that tests such as the MR Egger require variance in instrument-exposure associations in order to calculate a dose-response relationship between the exposure and outcome My instruments for wellbeing meet this requirement in one respect since each of the three variants have different instrument-exposure associations (Burgess & Thompson, 2017) and the consistent and linear relationship between instrument-exposure and instrument-outcome associations

supports that these were valid instruments which met assumptions for MR (Burgess et al., 2017). However, in another respect the MR Egger is considered most robust when using more than 10 genetic instruments (Howell et al., 2020). On balance, obtaining more instruments for wellbeing would improve the power of sensitivity testing though it is unlikely that I missed a marginal effect in this case since the statistical significance of this effect did not approach my threshold for significance (P=.471). I will continue discussing the impact of weak instruments in the next section.



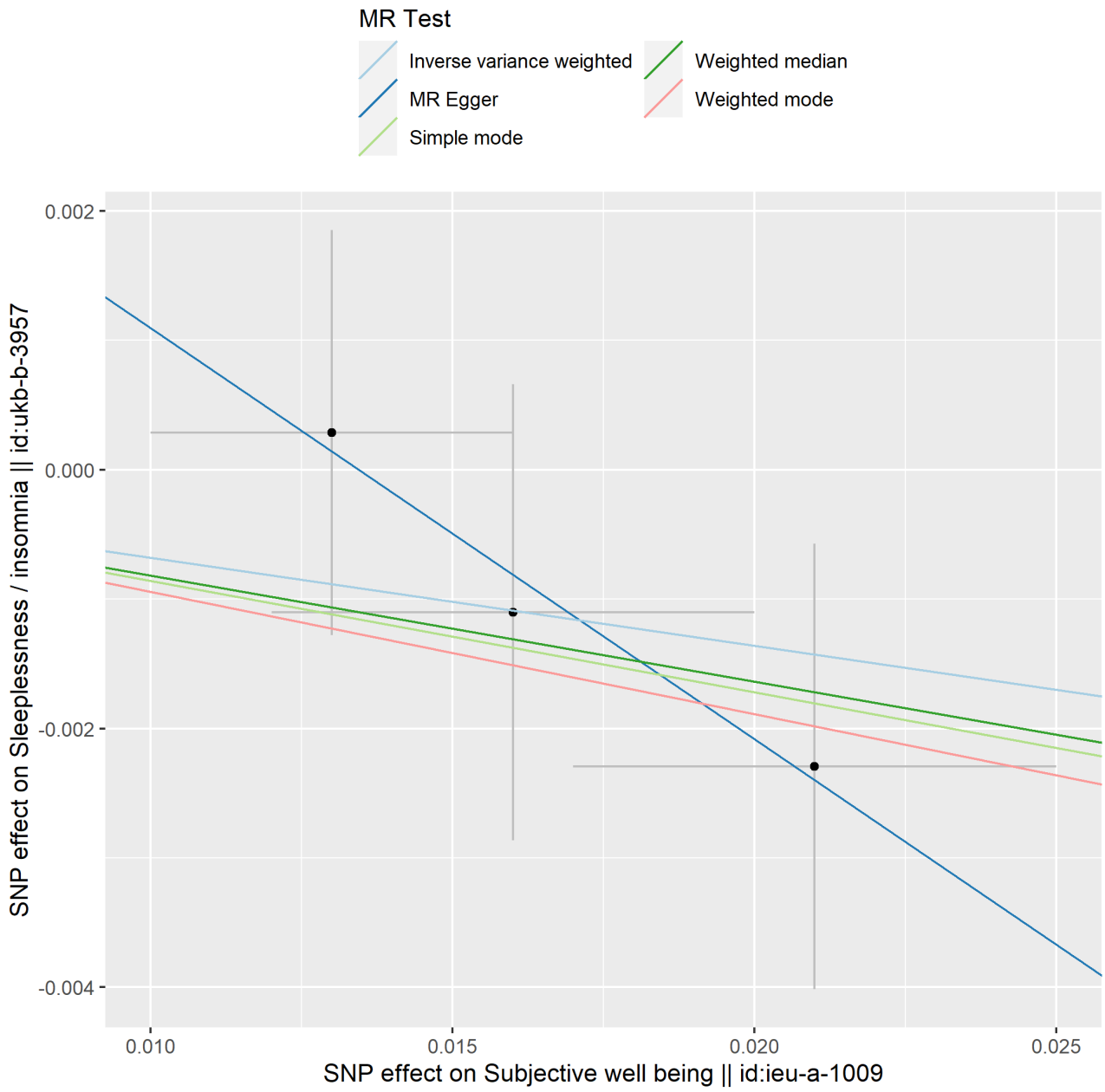


Figure 2.5 Wellbeing appears to exert a small effect causally reducing insomnia, but this did not reach statistical significance. Forest (top) and scatter (bottom) plots show the effect estimates for each instrumental variable (n=3) used to predict the effect of wellbeing on insomnia. The forest plot indicates that the few instruments used for wellbeing produce a range of estimates but IVW meta-analysis (indicated in red/blue) suggests that the average effect is negative but not statistically significant (b=-0.06, *P*=0.23). The scatter plot indicates that different estimators (Egger, median, mode) generally agree on the direction and magnitude of the effect, although Egger shows a larger gradient. However, the instruments used for wellbeing were few and appear weak which may have reduced the power of IVW effect estimation and sensitivity analysis. Confidence intervals are used as in the previous figures and represent 95% confidence.

2.4 Discussion

I conducted an MR study estimating the causal pathway between wellbeing and insomnia by conducting bi-directional MR. This discussion will offer an interpretation of these results with respect to four hypotheses in the background of previous observational research (Kyle et al., 2010; Ong, 2010), experiments (Boggiss et al., 2020; Haack & Mullington, 2005; Krystal, 2007), and MR studies (Jansen et al., 2019; Zhou et al., 2021) of insomnia and wellbeing.

The first hypothesis I investigated was that insomnia causes poor wellbeing. Evidence was found to support this since insomnia exerted a strong effect causally reducing wellbeing. This effect (b=-0.31) is almost identical to a previous estimate from a one-sample MR analysis(n=88, b=-0.31, P=9.42x10^{-14})(Jansen et al., 2019), as well as corroborating a negative causal effect found using a different measure of insomnia (Zhou et al., 2021)(n=53, b=-0.07, P=1.3x10^{-07}). My estimate is also in-line with experimental findings that sleep deprivation reduces wellbeing (Haack & Mullington, 2005) and insomnia medications improve wellbeing (Krystal, 2007). Sensitivity tests support the main effect estimates in that my instruments for insomnia explained a good proportion of variance, were not directly associated with wellbeing, and no indications of substantial horizontal pleiotropy were found. Furthermore, although my MR Egger estimate did not reach statistical significance, a previous study observed a similar finding (Zhou et al., 2021) and this is likely due to the test's lower statistical power to detect marginal effects (Burgess et al., 2017). In the background of previous estimates, the present study therefore adds to evidence that insomnia causally reduces wellbeing.

The second hypothesis I investigated was that poor wellbeing causes insomnia. My main estimate for a wellbeing on insomnia does not provide strong evidence although it numerically trends towards a protective effect where good wellbeing improves insomnia (b=-0.07). Previous research arrives at a similar conclusions, but reach statistical significance, and even produce similar effect sizes (n=29, b=-0.09, P=1x10^{-05})(Jansen et al., 2019)(n=13, b=−1.01, P=4.9x10^{-09})(Zhou et al., 2021). A protective effect of wellbeing on insomnia is also supported by previous experimental research

demonstrating that wellbeing interventions reduce symptoms of insomnia (White et al., 2019). However, though my instruments for wellbeing were stronger than has been used in previous research (Appendix 2.4), they were not numerous enough and this reduced my power to detect marginal effects. This is reflected in numerous non-significant *P* values across the various estimators (IVW, MR Egger, median, mode). Overall, this study supports previous research indicating an effect in the reverse direction, for wellbeing on insomnia, is possible, but stronger instruments are required to make further inference about the magnitude of this effect.

The third hypothesis investigated the possibility of a bi-directional effect. The present study finds some evidence that wellbeing and insomnia exert causal effects on each other, however, I was not able to obtain strong evidence for an effect in the reverse direction. Since assessing a bi-directional effects relies on similar power in both directions I am not able to make a strong conclusion for this hypothesis (Zheng et al., 2017). . A bi-directional effect is suggested by observational research (Kyle et al., 2010) and quasi-experimental studies showing that prior changes in wellbeing and insomnia predicts future changes in each other (Ong, 2010). It is supported by other previous MR research as well which use more instruments for wellbeing and find significant bi-directional effects (Jansen et al., 2019; Zhou et al., 2021) though I will explain in a following section that their methods for obtaining more instruments did not necessarily improved their strength of evidence over-all. Therefore, although previous research suggests a bi-directional effect exists between insomnia and wellbeing I was not able to find strong evidence.

The fourth hypothesis states that no causal pathway exists between wellbeing and insomnia. The present study found evidence that a causal pathway does exist and so I found evidence to reject this hypothesis. It remains possible that this hypothesis is true since a false positive relationship could be caused by discrepancies between objective and subjective measures. For example, objective measures of sleep duration (Jean-Louis et al., 2000) and sleep quality (Driscoll et al., 2008; Kyle et al., 2010) often find no association with wellbeing. I will end my discussion by suggesting several key

areas for research to continue investigating this hypothesis, including determining why objective and subjective measures disagree.

In summary, I found evidence that insomnia has a causal effect that reduces wellbeing. In the following sections of the discussion, I will draw out the implications of this finding, my strengths and limitations and suggest areas for further research.

2.4.1 Implications

The finding that insomnia may causally reduce wellbeing has important implications for society. Insomnia is likely to affect up to 40% of the population that experiences insomnia symptomology (Fernandez-Mendoza & Vgontzas, 2013), and this proportion may be growing (Albrecht et al., 2019; c.f.: Youngstedt et al., 2016). Insomnia can have profound impairments including memory problems (Alhola & Polo-Kantola, 2007) and accidents on the road and in the workplace (Wade, 2010). Insomnia is therefore a prevalent issue which has severe effects on society and warrants intervention. The implication that an intervention on insomnia could additionally improve wellbeing gives further incentives to develop such an intervention. It could follow that improving insomnia could also benefit society in ways including improving productivity (Marks & Shah, 2004) and protecting against disease (Garbarino et al., 2016a). Therefore, my findings emphasise the importance of treating insomnia as a prevalent issue affecting society.

The issue of whether insomnia reduces wellbeing is also highly relevant to Individuals suffering depression. Depression is known to cause insomnia (World Health Organization, 1993), where unipolar depression reduces sleep quality and bipolar depression results in manic wakeful phases (Garbarino et al., 2016a). Persistent low mood or unhappiness are also core symptoms (World Health Organization, 1993) which implies that treatments for insomnia may effectively treat two symptoms of depression. This is supported by evidence that medication for insomnia can improve wellbeing as well as depressive symptoms although it is not clear that these improvements are directly caused by improving sleep (Krystal, 2007; Kyle et al., 2010; Perach et al., 2019). Future

research should therefore continue to investigate the value of sleep treatments for depression and expand this investigation to other mental illnesses which may impact wellbeing.

2.4.2 Strengths and limitations

The instruments I selected for insomnia performed well but the instruments for wellbeing did not. This reduced the strength of evidence for a causal effect of wellbeing on insomnia. I will highlight some advantages of instruments used in the present study, and review several approaches future researchers can take to obtain stronger instruments, including a method of combining many weak instruments (Jansen et al., 2019; Wootton et al., 2018), investigating a wellbeing spectrum of multiple related phenotypes (Baselmans et al., 2019), and using information about genetic correlations to increase statistical power in GWAS (Turley et al., 2018).

The GWAS used for insomnia and wellbeing in the present study have some strengths. The GWAS for insomnia (Jansen et al., 2019) remains the largest collection of insomnia related genetic variants. The authors achieved this by combining two large cohorts, UKBiobank and 23AndMe, to achieve high power in GWAS (n=1,331,010). It is therefore more likely that the GWAS sample and discovered variants are representative of the general population. The GWAS for wellbeing (Okbay et al., 2016b) used a similar approach, combining 49 smaller cohorts into a single large sample (n=298,420), and was among the first GWAS to identify variants robustly associated with wellbeing. The authors also perform a detailed examination of these three variants, and their likely path of biological action is plausible, on areas of the body and brain associated with stress (pancreas) and emotion (limbic system). However, there has been some doubt about these variants since two are associated with depression (Okbay et al., 2016b), only one is associated with wellbeing in an independent sample (UKBiobank)(Wootton et al., 2018), only one was replicated in a GWAS of wellbeing related phenotypes (Baselmans et al., 2019), and none were replicated in a subsequent wellbeing GWAS (Turley et al., 2018). Taken together, it is not clear to what degree the three instruments I used have reliable and specific effects on wellbeing. Future research should consider using other instruments

for wellbeing. I present some alternative approaches to obtaining instruments for wellbeing in Table 2.3 and will review them now.

Table 2.3 Previous studies have identified and selected instruments for wellbeing varying in number and strengths. Compared to other instruments, the dataset identifying three SNPs used in the present analysis (Okbay et al., 2016b) appear relatively weak but are specific to wellbeing and less related to other variables which is important for MR analyses which assume instruments do not act through pleiotropic pathways.

Study	GWAS size	SNPs	R^2	Notes
Okbay et al., 2016	298,420	3	0.9%	Data: SSGAC
Wootton et al., 2018	197,174	84		Lower threshold ($P<5x10^{-05}$), Data: SSGAC
Baselmans et al., 2019	2,370,390	304	0.92-1.06%	Described a "wellbeing spectrum", Data: SSGAC, UKBiobank, British Household Panel Survey (Brice et al., 1993)
Turley et al., 2018	388,542	49	1.57%	Data: SSGAC, UKBiobank, 23andMe
Jansen et al., 2019		29		Lower threshold ($P<1\times10^{-5}$), Data: UKBiobank, 23andMe,

Note: Mean R^2 as reported in discovery GWAS. Blank cells indicate data not reported.

The first method which may increase instrument strength is reducing the threshold for statistically significant association since this can allow researchers to select a greater number of variants. Combining many instruments can help explain a greater proportion of variance so can improve instrument strength (Pierce et al., 2011). This approach was taken in a previous study (Wootton et al., 2018) where a lower threshold for association ($P<5\times10^{-5}$) increased the number of variants selected from a GWAS (Okbay et al., 2016b) from 3 to 84. The authors demonstrate that the increase in power from combining multiple instruments helped achieve better power to predict happiness in an independent cohort, and to perform MR with increased *P* value confidence. One further study provides few details (Jansen et al., 2019) but performed a similar method ($P<5\times10^{-5}$) to obtain 29 variants used as instruments for wellbeing. However, this must be performed carefully to avoid selecting many weak instruments, less robustly associated with the exposure, and to avoid including invalid instruments, given the increased likelihood of including instruments which act through

pleiotropic pathways (Burgess et al., 2013). For example, in the previous analysis (Wootton et al., 2018) the greater number of instruments showed a greater degree of measurement error which significantly diluted regression analyses (I^2GX = 0.37, n=84 vs I^2GX 0.96, n=3). This demonstrates that selecting more instruments is not necessarily better, and there is often a trade-off between achieving statistical power and reducing the risk of not meeting instrumental variable assumptions (Pierce et al., 2011).

A method of identifying more variants without compromising strength of association is to analyse GWAS data using higher powered multi-trait methods of statistical analysis. Multi-trait GWAS methods exploit the high degree of correlations between related phenotypes to increase statistical power. This method is particularly relevant for variables such as wellbeing which has high genetic correlations with variables such as mental illness (Okbay et al., 2016b) and insomnia (Jansen et al., 2019). GWAS estimates for different phenotypes are often correlated and this information can be incorporated into more sensitive multi-trait analysis which accounts for different sources of correlation. Using samples from the SSGAC, UKBiobank and 23andMe (n=388,542), a multi-trait analysis of GWAS (MTAG)(Turley et al., 2018) was compared to a traditional GWAS. Analysis increased the number of variants robustly associated with wellbeing from 13 in GWAS to 49 using MTAG (277% increase). The authors argue this is representative of increasing the discovery sample size by 55% (Turley et al., 2018) and multi-trait methods have been used before (Baselmans et al., 2019) to increase statistical power and achieve an increase of 57% in estimated association strength between variants and wellbeing. The implications for MR are that multi-trait methods can be used to obtain more instruments for wellbeing. For example, a previous MR study finding a bi-directional effect for insomnia and wellbeing (Zhou et al., 2021) leveraged the MTAG dataset (Turley et al., 2018) to obtain a greater number of instruments (n=39) for wellbeing and accordingly achieved greater statistical power in IVW estimation to detect an effect of wellbeing on insomnia. Although a subsequent adjusted analysis dropped many of these instruments (n=13), their initial analysis with more instruments had far greater confidence returning a *P* value multiple orders of magnitude times

smaller (P=4.0x10^{-17}, n=39 vs P=4.9x10^{-09}, n=13)(Zhou et al., 2021). The disadvantage of multi-trait methods in MR though is the risk of including instruments which act through pleiotropic pathways, through related traits, and this was the basis for the previous MR study (Zhou et al., 2021) to drop MTAG instruments from their analysis due to other phenotype correlations. Researchers should therefore consider multi-trait analysis results when selecting instruments sufficiently strong to investigate the effects of wellbeing.

An alternative method of obtaining an even greater number of instruments for wellbeing is to combine related traits and investigate a "wellbeing spectrum". Researchers (Baselmans et al., 2019) combined many variables from a range of datasets, including the SSGAC and UKBiobank, into a single analysis to achieve a sample size much larger than previous studies (n=2,370,390). Samples were combined from variables related to wellbeing including life satisfaction (N=80,852), positive affect (N=410,603), neuroticism (N=582,989), and depression (N=1,295,946). The result was identifying the greatest number of variants associated with wellbeing to-date (n=304). However, these variants are not entirely specific to wellbeing, a greater proportion of the sample size comes from depression (55%) rather than wellbeing (21%)(Turley et al., 2018). Researchers may therefore use these SNPs in the hope to achieve greater instrument strength where the gain in power may be worth the loss in specificity.

2.4.3 Future directions

A key future direction is the identification of stronger instruments for psychological variables, as well as conducting more research using objective measures and comparing the results with self-report measures.

Given the importance of strong instruments in MR, attention should focus on identifying more variants robustly associated with wellbeing. This is especially important in psychology given that many variables are complex traits with smaller genetic influences. For example, authors of wellbeing

MR studies comment that the lack of quality instruments for wellbeing hampered their investigation either reducing power (Zhou et al., 2021), preventing analyses (Jansen et al., 2019), or producing unreliable instruments not associated with wellbeing in independent samples (Wootton et al., 2018). This is not specific to wellbeing and a restrictive number of variants are found for other psychological variables of interest including depression (Okbay et al., 2016b), personality (van den Berg et al., 2016), and for behaviours including exercising (Doherty et al., 2018) and caffeine consumption (M. C. Cornelis et al., 2015). Taken together this demonstrates that there is a need for stronger psychological instruments. It has been demonstrated that large sample sizes (Okbay et al., 2016b; Turley et al., 2018) and multi-trait methods (Baselmans et al., 2019; Turley et al., 2018) identify more variants. There is therefore a clear case for a novel and more highly powered GWAS with a large sample size designed to identify variants robustly associated with a range of phenotypes for use in psychological MR studies.

While one approach to improving instrument strength is to identify more variants, another approach is to improve the quality of instruments by using better measurements. An issue is that researchers define wellbeing (Diener, Lucas, et al., 2018) and sleep disorders (Garbarino et al., 2016a) in different ways, so the first step is agreeing on clear definitions for variables of interest. This is especially important in psychology since many phenotypes overlap, such as sleep symptoms and poor wellbeing in depression (World Health Organization, 1993). A similar issue is that researchers use a range of measures for wellbeing (Diener, 1984) and sleep variables (Ziporyn et al., 2017), so the second step is to agree on what measures are most accurate and practical for use in large-sale studies. Often it is more practical to ask respondents to complete a single questionnaire which includes a range of items designed to measure multiple variables simultaneously (e.g., quality of life surveys)(World Health Organisation, 2012), and in these cases it is important to report sub-scores so that items specific to a single variable can be extracted, but this is not always performed (Lins & Carvalho, 2016). This is relevant since researchers (Okbay et al., 2016b) often combine similar items in different measures across many cohorts to increase sample size. Improving the specificity with

which variables such as wellbeing are measured may increase instrument-exposure associations and reduce measurement error, thus improving the signal-to-noise ratio and increase instrument strength overall. Using measures which precisely measure a clearly defined construct, and are compatible with other measures, should improve the quality and strength of instruments available for future psychological MR studies.

Lastly, a gap currently exists in evidence from MR using objective measures of sleep. Human inaccuracies in sleep duration report (Lauderdale et al., 2008) may contribute noise to instrument-exposure associations when used in MR. Objective accelerometery data, by contrast, is more accurate (Jean-Louis et al., 2000) and may produce higher quality instruments. For example, a pre-print article (Salzmann et al., 2021) used MR and found greater effect sizes for an effect of sleep duration on cognitive abilities when using objective measures of sleep compared with self-report measures, which suggests that more accurate measurement reduces noise in MR. Similarly, polysomnography could be used to measure sleep quality (Driscoll et al., 2008) and produce higher quality instruments for sleep quality. Furthermore, research using objective measures would be able to investigate the degree to which existing MR estimates for sleep and wellbeing are biased by self-report measurements. However, objective measures often limit sample sizes, for example UKBiobank has fewer observations of objective sleep duration (n=85,499) compared with self-report (n=446,118)(data from: http://sleepdisordergenetics.org). This is because they are more expensive and require specialist equipment, and it would be useful to develop objective measures which are more practical to deploy at scale.

2.4.4 Conclusion

This study investigated four possible hypotheses about the causal pathway between wellbeing and insomnia. Two-sample MR results provide some evidence to support previous findings that insomnia and wellbeing exert bi-directional causal effects on each other. Evidence was stronger that insomnia causally reduces wellbeing but the strength of evidence for an effect of wellbeing on insomnia was

hampered by weaker instruments. These findings have implications for insomnia treatments which may additionally improve wellbeing. Future research seeking to interrogate this causal pathway may benefit from stronger instruments. This study focussed on the effect of sleep and wellbeing in isolation of other related variables, and the next chapter will build on these findings using an advanced network method of MR to reveal information about a wider network of related variables.

3 Exploring the wider network of causal variables for insomnia and wellbeing

3.1 Introduction

The causal pathway between insomnia and wellbeing does not exist in isolation but instead as part of a wider network of variables and relationships. These relationships exemplify the type of complexity present in the network of variables related to human health. In subsequent chapters I will investigate methods of understanding this complexity through visual (4), simulation and game mediums (5, 6). The present chapter outlines the steps used to obtain the materials needed for these later studies, including a network dataset describing the relationships between 16 health variables and demonstrating a method of using MR estimates to calculate indirect effects.

3.1.1 A network of human health

Many physical and mental health variables are highly inter-related and this complicates our understanding of them. For example, there is a high degree of network complexity in the relationships between variables like wellbeing (Diener et al., 2017) and insomnia (Hom et al., 2020; Kyle et al., 2010). This complexity can be modelled by mapping the relationships between variables across the human phenome.

The "human phenome" includes every phenotype, disease or characteristic, that a human can possess (Freimer & Sabatti, 2003). Building a "map" of the human phenome refers to collecting data to understand the relationships between many or all phenotypes. Some genomics datasets contain information on a wide range of phenotypes and these are valuable sources of information for researchers wishing to map the human phenome. For example, genomics data has been used to map how different variables across the phenome share associations with similar genetic variants (Evans et al., 2013).

Causal maps of the human phenome can be produced by applying MR to make causal inferences in a hypothesis-free manner (Evans & Davey Smith, 2015). Recent advances provide frameworks and methods for performing hypothesis-free MR on large scales (Brown & Knowles, 2020; Hemani, 2022) and have made it more practical by collating GWAS summaries in databases like MR Base (Hemani, Zheng, et al., 2018b). MR researchers are now applying mediation analyses (A. R. Carter et al., 2021) in network frameworks (Burgess et al., 2015b) to estimate the causal relationships between hundreds (Brown & Knowles, 2020) and thousands of variables (Hemani, Bowden, et al., 2017).

3.1.2 Mediation analyses in MR

Mediation analyses (A. R. Carter et al., 2021) allow researchers to decompose MR effect estimates into direct and indirect components. They work by incorporating information from additional variables to determine whether, and to what degree, effect estimates are influenced by mediating variables (Figure 3.1). For example, an effect of insomnia on wellbeing may be partly explained by insomnia causing depression which in turn reduces wellbeing.

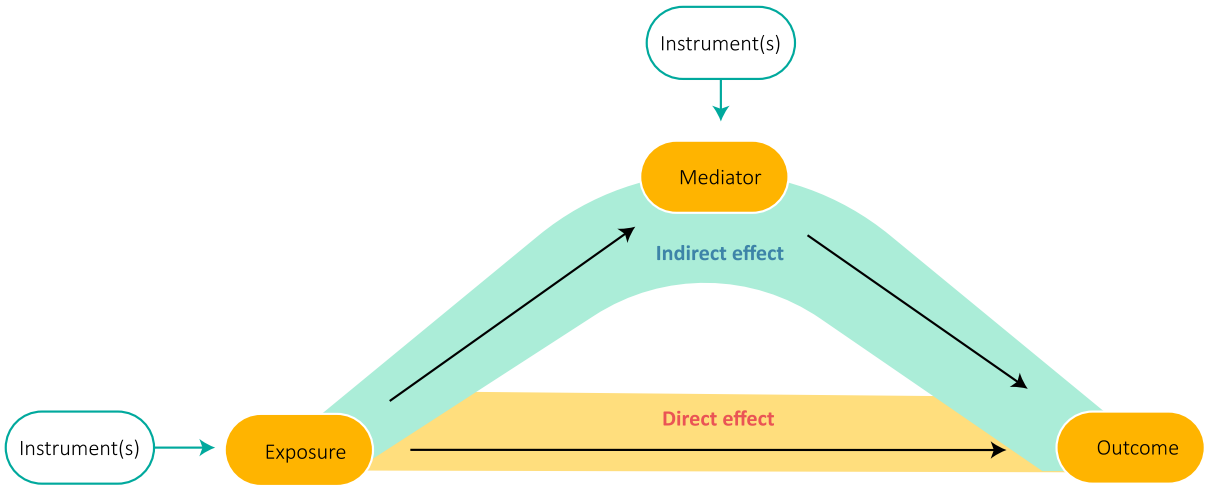


Figure 3.1 Mediation analysis in MR allows researchers to decompose an effect estimate into a direct component and an indirect component which acts through a mediator. Note that although they are represented similarly in causal diagrams, a mediator is not a confounder since it represents a valid causal pathway.

The effect of the exposure on the outcome is considered the "total" effect estimate. The total effect is then adjusted for the "indirect" effect through the mediator to arrive at an estimate which better

describes the "direct" effect of the exposure on the outcome. The precise calculations depend on the type of mediation analysis; the main types are "Multivariable MR" (Sanderson et al., 2019) and "Product-Of-Coefficient" (Relton & Davey Smith, 2012). Both methods are similar in that genetic instruments are obtained for each variable in analysis but the instruments in Multivariable MR can be related to multiple exposures, confounded, whereas instruments in Product-Of-Coefficient methods cannot. In this chapter I opt to use a Product-Of-Coefficient method because there is a framework for extending it in network analysis (Burgess et al., 2015b).

3.1.3 Network MR

"Network MR" (Burgess et al., 2015b) is an emerging framework which extends mediation analyses to large networks of causal effects. Conducting network MR involves three stages of variable selection, discovery and analysis and in this section I will expand on these stages, drawing out the opportunities and challenges they present in this chapter.

The first step in network MR is variable selection. The aim of this stage is to define the scope of network analysis by selecting which variables will be included in analysis. Variable selection could be guided by natural constraints that researchers have little control over. For example, a map of the human phenome might include all variables for which there is sufficient data and only exclude variables they cannot study (Brown & Knowles, 2020). Variable selection might otherwise be guided by criteria imposed by researchers to guide variable selection and obtain a curated dataset. For example, one MR study restricted analysis to psychiatric traits in order to focus on and understand causal pathways between mental health disorders (Gao et al., 2019). This first stage guides the focus of network MR research and determines the variables entered into a subsequent stage of network discovery. In the present study, I will implement criteria to guide this stage of variable selection and ensure I obtain a valid network dataset for further study.

The second step is network discovery (Figure 3.2) and this has the aim of estimating causal effects between the variables in analysis. The relationships between variables for analysis are investigated

using “hypothesis-free MR”. A hypothesis-free manner of investigation in MR refers to investigating all pairwise relationships within large datasets comprising many variables (Evans & Davey Smith, 2015). The output is a series of bi-directional causal effect estimates for the effects of each variable on every other. These estimates are considered suggestive and require further hypothesis-testing to form strong conclusions. Furthermore, a correction to the threshold for statistical significance is typically made to account for having conducted so many tests. Additionally, performing Two-Sample MR (Lawlor, 2016), where exposure and outcome information is compared across two separate GWAS, allows researchers to compare a wider range of variables with exposure and outcome information across different GWAS summary datasets (Hemani, Bowden, et al., 2017). The result of this stage is a “network dataset” which refers to a collection of causal effect estimates between multiple related variables. These datasets exemplify network complexity because studying the causes and effects of diseases reveals multiple inter-dependent variables.

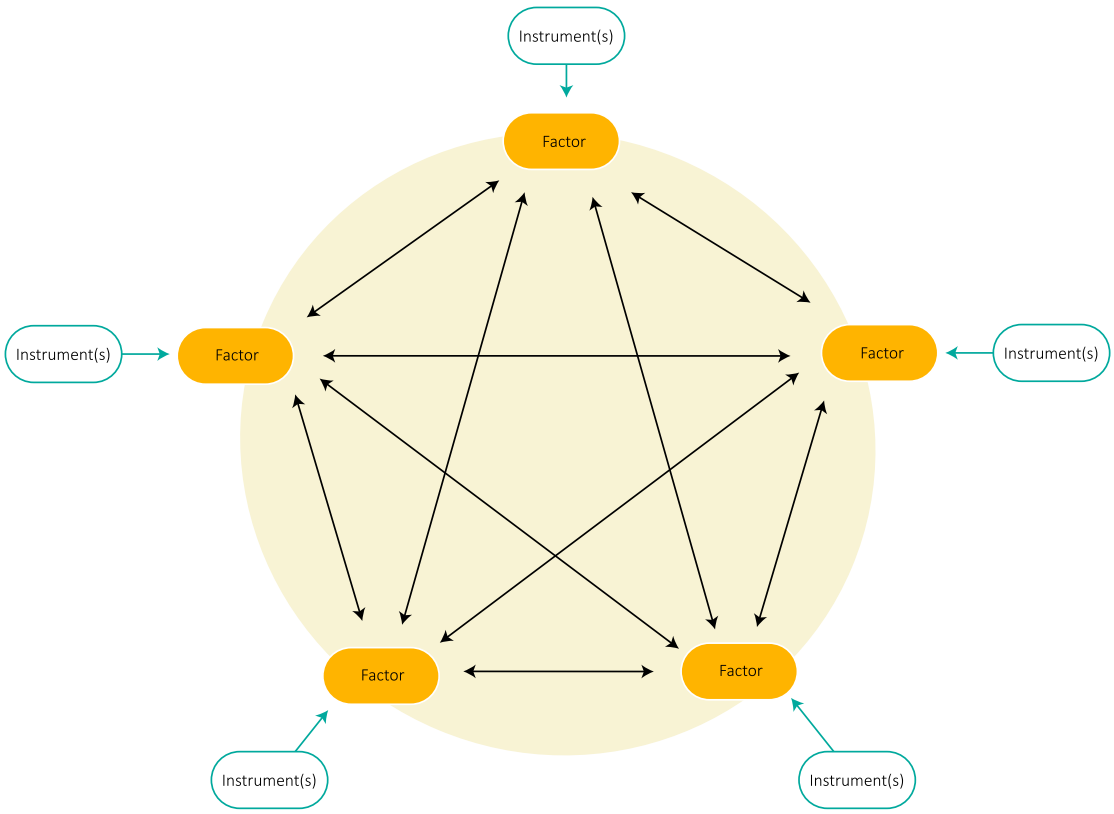


Figure 3.2 Network MR discovery involves using instruments for many variables in a network to obtain estimates for the effects of every variable on every other.

The third step in network MR is decomposing total effect estimates into direct and indirect effect estimates (Figure 3.3). Mediation analysis (Burgess et al., 2015b) is used to decompose the causal effects in the network dataset. In the present chapter I will demonstrate how this method can be applied in a limited manner, using the network dataset to identify and test potential mediators related to select variables of interest (as is done in Brown & Knowles, 2020). In chapter 5 I will apply this method on a larger scale, estimating indirect effects between all variables (as is done in Hemani, 2022), as part of developing a public health intervention simulation.

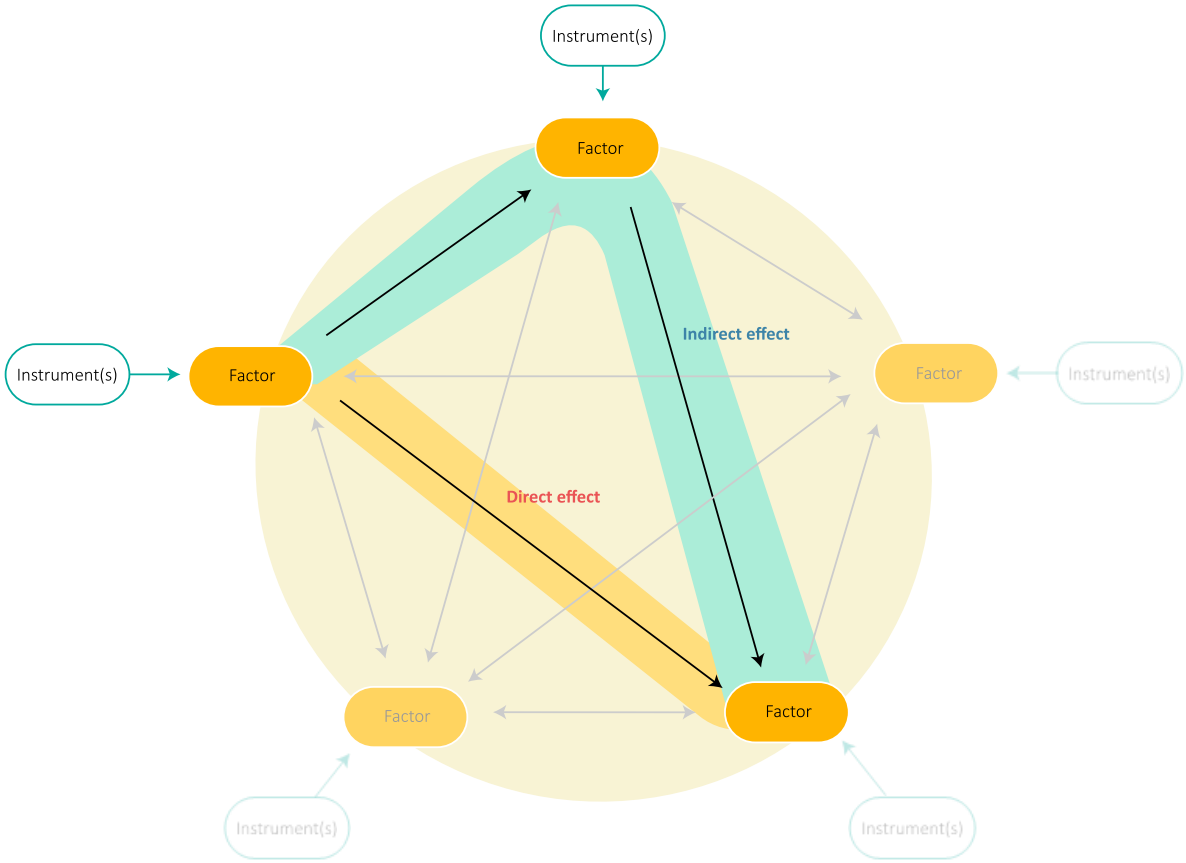


Figure 3.3 Network MR involves identifying potential mediators for an exposure-outcome relationship and performing a mediation analysis to decompose total effects into indirect and direct components. This process can be repeated for each possible pairing of variables in a network.

Network MR is therefore a powerful method which allows researchers to estimate the causal relationships between many variables, obtain a network dataset, and decompose estimates to investigate indirect effects. One advantage of this method is that it extends the well-established mediation MR method, so it is grounded in a valid framework, and the methods may already be familiar and understood by researchers. However, one disadvantage is that the method has stricter assumptions for valid causal inference (Burgess et al., 2015b). In particular, causal effects are more strongly assumed to be linear and fixed, meaning that they do not vary with the levels of the exposure. In the discussion I will come back to the limitations of this method and discuss the implications on obtaining valid network datasets.

3.1.4 Applying network MR to insomnia, wellbeing and related variables

In this chapter I will use network MR to develop materials representing the type of network complexity present in public health. I will later use these to develop a simulation of public health interventions which will be used by a lay audience to understand relationships between health variables (chapters 5, 6). In this section I will explain how I will use network MR to obtain a network dataset in three stages of variable selection, network discovery, and mediation analysis.

First, I will iterate on previous work by implementing criteria into a stage of variable selection. This will help ensure my network dataset overcomes challenges faced by previous research and is interpretable by a lay audience. A previous study (Hemani, Bowden, et al., 2017) obtained a network dataset called the MR of Everything versus Everything ("MR EvE"). Every measure in the MR Base GWAS summary catalogue (Hemani, Zheng, et al., 2018b) was entered as variables in hypothesis-free MR. The aim of MR EvE therefore was to collect a series of causal effect estimates and explore the network of effects between 2407 variables from across the human phenome. MR EvE is a seminal work in network MR that demonstrates how to apply network MR at-scale, overcomes various challenges, and highlights areas for future research. However, the result is highly complex and difficult to understand even at a small scale (Figure 3.4). I will now explain how I will use variable selection to address specific issues and improve the interpretability of my network dataset.

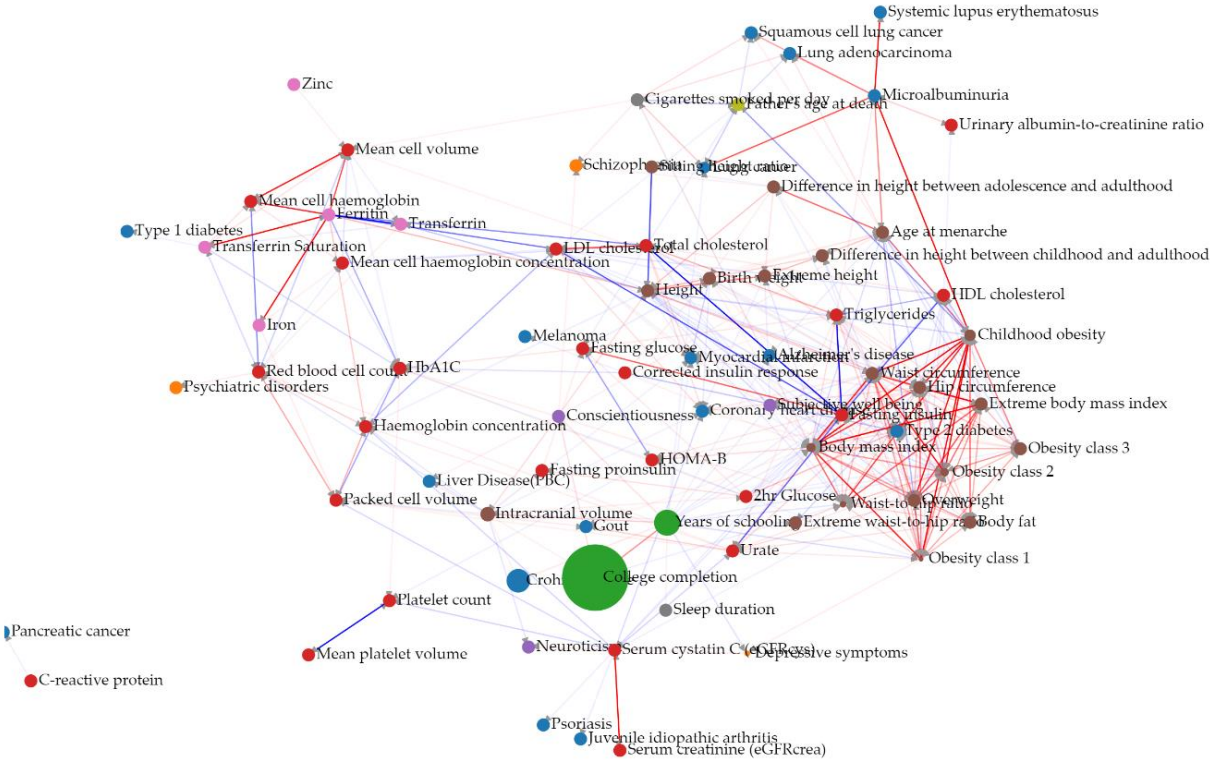


Figure 3.4 A sub-network from MR EvE containing almost 5% of the total variables in the network and 2.5% of the relationships (Hemani, Bowden, et al., 2017). It demonstrates that the causal pathways between health variables are numerous and complex to understand. However, this network dataset is difficult to understand for a number of reasons that are not related to understanding causal pathways, including presenting duplicate and analogous measures.

Three issues make the MR EvE dataset difficult to understand (Free Ice Cream & Davis, 2018). One issue is that the MR EvE network contains 5660 causal effects, and understanding a dataset of this scale would be very difficult and take an impractically long time for my future participants to learn. Another issue is that 2258 (94%) variables are highly specific biological variables, such as blood plasma levels and metabolites. This is problematic because variables like “urate levels”, “transferrin” and “HbA1C” are not accessible to a lay audience, and this frustrates understanding in a way which is not related to network complexity. Furthermore, when multiple GWAS summaries in MR Base use the same measurement, for example measuring body weight, these are each recorded as separate records and are not grouped with identical measurements. Consequently, MR EvE contains many variables representing identical constructs, for example there are 30 measures of body mass index. Similarly, many measures are highly similar, highly correlated, analogues of one-another which are

unlikely to provide different information. For example, 90% of the variance in body mass index and waist circumference are explained by overlapping genetic variants and so would provide redundant information in MR (https://ukbb-rg.hail.is/rg_browser/). Therefore, the issues that make MR EvE difficult to interpret are the large scale of the dataset, the inclusion of obscure variables, and multiple duplicate or analogous measures for similar variables.

A key contribution in the present chapter will be introducing selection criteria to curate a highly interpretable network MR dataset. This will allow me to model network complexity and present this to lay audiences (chapter 5, 6). I will manually select variables for inclusion based on inclusion and exclusion criteria. This will help me select a smaller number of variables, exclude variables which are not accessible to a lay audience, and ensure that each variable in the network provides distinct information. I will use my previous MR investigation as the starting point and select variables which are related to wellbeing, insomnia, or both.

Second, I will obtain causal effect estimates between selected variables by performing hypothesis-free MR. My methods will be based on previous network MR research (Burgess et al., 2015b; Hemani, Bowden, et al., 2017). The result will be a network dataset that describes the causal pathways between each variable in analysis. Although network MR estimates are considered suggestive, requiring further validation to ensure they describe true causal effects, this dataset will be valuable as it will exemplify the types of network structures which exist in the relationships between health variables (Brown & Knowles, 2020).

Third, I will demonstrate the mediation analysis method. I build on this in chapter 5 where I use it to simulate how public health interventions can have many, indirect, spreading side-effects. Mediation analysis provides a method of estimating indirect effects and in the present chapter I will demonstrate how mediation analysis can be applied to network MR. I will continue from the previous chapter, using the example of wellbeing and insomnia, and identify whether any of the variables in my network mediate the causal pathway between them.

The aim of this chapter is therefore to create a network dataset. I will use this as the basis for further investigations in later chapters to develop methods of understanding network complexity.

3.2 Methods

Using insomnia and wellbeing as my key variables, I selected related variables to include in my network dataset, and performed network MR. I sourced information from GWAS summary statistics in MR Base (Hemani, Zheng, et al., 2018b) and used the TwoSampleMR package for *R* to perform MR (Hemani, Zheng, et al., 2018a). This is the typical method for collecting information and performing network MR because conducting summarisation and data preparation manually would be prohibitively time consuming and make it impractical to make the large numbers of required comparisons (Brown & Knowles, 2020; Hemani, Bowden, et al., 2017). All code and data used in this chapter are available online (https://osf.io/sy3ne/).

3.2.1 Variable selection

The MR EvE network dataset (Hemani, Bowden, et al., 2017) is unnecessarily complex for the current demonstration, so I used the criteria below to curate variable selection. These ensured variables were relevant, interpretable by a lay audience, and did not contain duplicates and analogues:

1) Variable is generally intuitive to a lay audience
    a. Complex traits (e.g., intelligence), diseases or biological intermediates (e.g., body mass index) were included
    b. Metabolites (e.g., urate), plasma serum proteins (e.g., erythrocyte volume), or indicators (e.g., HbA1C) were excluded
2) Variable has summary GWAS information present in MR Base (Hemani, Zheng, et al., 2018b)
3) Variable has been observationally or clinically associated with either wellbeing or insomnia
    a. I discovered evidence as part of my literature review on wellbeing and insomnia conducted in the previous chapter (2, search terms in Appendix 2.1)

4) Variable is not a duplicate or close analogue of another variable in the network

    a. A close analogue of a variable was defined as a variable with a genetic correlation above 80% (r>.8)

This resulted in 16 variables (see Table 3.1) to include in this demonstration network dataset.

Table 3.1 Variables selected for inclusion

		Evidence relating variables to wellbeing and/or insomnia	
Variable	**MR Base ID**	Wellbeing	Insomnia
Alcohol consumption	ukb-b-5779	(Parackal & Parackal, 2017)	(E. O. Johnson et al., 1998)
Body mass index	ukb-b-19953	(Diener et al., 2017; R. T. Howell et al., 2007)	(Garbarino et al., 2016b; Knutson et al., 2007)
Caffeine consumption	ukb-b-5237	(Fibiger & Phillips, 1988) (Rogers, 2007)	(Snel & Lorist, 2011)
Coronary heart disease	ieu-a-7	(Diener et al., 2017; R. T. Howell et al., 2007)	(Garbarino et al., 2016b)
Depression	ieu-a-1187	(Das et al., 2020)	(Wulff et al., 2010)
Diabetes (type 2)	ieu-a-24	(Kinmonth et al., 1998)	(Knutson et al., 2007)
Education	ieu-a-1239	(Eisenberg et al., 2009)	(Alhola & Polo-Kantola, 2007)
Intelligence	ukb-b-5238	(Eisenberg et al., 2009)	
Exercise	ukb-b-4710	(Lane et al., 2014)	(Driver & Taylor, 2000)
Loneliness	ukb-b-8476	(Diener, 1984; Lane et al., 2014)	(Hom et al., 2020)
Socialisation	ukb-b-5076		(Hom et al., 2020)
Neuroticism	ieu-a-118	(Das et al., 2020)	(Lai, 2018)
Smoking	ieu-a-961	(Robson, 2010)	(Jaehne et al., 2012)
Worries	ukb-b-6519	(Lai, 2018)	(O'Kearney & Pech, 2014)

* Education & intelligence, and loneliness & socialisation were not considered analogues because they had genetic correlations lower than 80% (75% and 41% respectively; https://ukbb-rg.hail.is/rg_browser/)

3.2.2 Obtaining measurement information

My network dataset is intended to describe the causal relationships between wellbeing, insomnia and potentially causal related variables. Having selected variables for inclusion, I sought to use MR in a hypothesis-free manner to estimate the causal effects between each variable on every other (though, strictly speaking, this was hypothesis-light since I had prior evidence that variables were related). This involves conducting a series of bi-directional MR estimates treating each variable in the network as exposure and outcome exhaustively until each possible relationship had been tested.

This involves collecting measurement information for each variable as well as the associated genetic variants.

Information on insomnia and wellbeing from the previous chapter were joined with information from GWAS summaries for fourteen additional variables (Table 3.2). UKBiobank (Elsworth et al., 2019) was used for eight variables: body mass index (BMI), worries, intelligence, socialisation, loneliness, exercise, alcohol and caffeine consumption. Additional consortia were selected for information on education (Lee et al., 2018), smoking (Tobacco and Genetics Consortium, 2010), depression (Wray et al., 2018), neuroticism (Genetics of Personality Consortium et al., 2015), diabetes (Morris et al., 2012) and coronary heart disease (CHD)(Nikpay et al., 2015).

Ideally each variable would have been sourced from independent GWAS samples but many of the largest, or only, available GWAS were conducted with overlapping samples. This can present a risk of bias from overlapping samples which typically bias two-sample MR effect estimates towards the observational estimate. To estimate the risk of this I calculated the overlap between samples (Table 3.3). There is no way to determine this precisely given the publicly available data, so I estimated this by collecting information on the 190 cohorts used in 8 GWAS consortia (for cohorts see https://osf.io/sy3ne/), and then calculating conservative estimates of sample overlap assuming the maximum possible overlap. The results of this indicate most consortia have no substantial overlap but some combinations have large overlap (e.g., CARDIoGRAM and TAG, SSGAC and UKB). Overlap between certain consortia risks biasing their effect estimates towards the observational estimate (Lawlor, 2016). One method of managing this bias is using strong instruments with high signal-noise ratios (LeBlanc et al., 2018). Fully accounting for this effect is still an open question in network MR that previous research has not addressed (Hemani, Bowden, et al., 2017), though one network study (Brown & Knowles, 2020) took advantage of the negligible risk of bias when applying two-sample MR to compare measurements within UKBiobank (Minelli et al., 2021), and the present study benefits in a similar way since most measures are sourced from UKBiobank.

Table 3.2 Contributing GWAS summary dataset information

First author (year)	Consortium	Variants*	Population	Gender	Variable(s)	Sample size
(Elsworth et al., 2019)	UKBiobank	9851867	European	Mixed	Loneliness	455364
					Worry	450765
					Alcohol	462346
					Intelligence	149051
					Coffee intake	428860
					Socialising	461369
					Exercise	440266
					Insomnia	462341
					BMI	461460
(Tobacco and Genetics Consortium, 2010)	TAG	2459119	European	Mixed	Smoking	68028
(Nikpay et al., 2015)	CARDIoGRAM plus C4D	9455779	Mixed (77% European)	Mixed	CHD	184305
(Morris et al., 2012)	DIAGRAM plus Metabochip	127904	Mixed (95% European)	Mixed	Diabetes	149821
(Wray et al., 2018)	PGC	10000	European	Mixed	Depression	480359
(Genetics of Personality Consortium et al., 2015)	GPC	6949615	European	Mixed	Neuroticism	160958
(Okbay et al., 2016a)	SSGAC	2264177	European	Mixed	Wellbeing	197174
(Lee et al., 2018)	SSGAC	10101242	European	Mixed	Education	766345

* The number of variants is the total number of single nucleotide variants which were tested in each GWAS sample. Small numbers of tested variants will contribute to some exposure-outcome pairs not being testable in the present study since variants used as instruments, with information in the exposure GWAS, did not necessarily have information available in the outcome GWAS. Additionally, larger numbers of tested variants increase the likelihood of finding more variants associated with a given variable. Similarly, effect estimates are more valid in two-sample MR when comparing information across GWAS with similar populations, such as having similar ancestry and genders.

Table 3.3 Maximum possible overlap between samples

	Sample overlap with other GWAS						
GWAS	CARDIoGRAM	DIAGRAM	GPC	PGC	SSGAC	TAG	UKB
CARDioGRAM		12%	1%	0%	2%	19%	0%
DIAGRAM	16%		1%	8%	5%	14%	0%
GPC	0%	0%		1%	2%	0%	0%
PGC	0%	18%	6%		0%	0%	47%
SSGAC	3%	4%	5%	0%		3%	68%
TAG	51%	28%	0%	0%	8%		0%
UKB*	0%	0%	0%	6%	24%	0%	

Total sample size	*195,813*	*145,599*	*480,395*	*63,036*	*164,408*	*74,019*	*462,346*

Note: Samples sizes represent the largest sample used in the present study.

Measurement information (Table 3.4) was extracted from the GWAS summaries for each variable. Variables were measured in a variety of ways including objective measurements (e.g., BMI, intelligence), medical records (e.g., CHD) and self-report questionnaires (e.g., loneliness, socialisation). Socialisation was reverse-scored so I will refer to this as “not socialising”.

Variables were measured using a range of units. Most variables were measured using a z-score which represented participants’ responses as the relative difference to the mean score. Standard deviation (SD) units were used to express this difference where a score of 0 indicated a mean response, +1 indicated a response one SD above the average, and -1 indicated a response one SD below the average. This was used for some ordinal scales by converting responses on Likert scales, such as “rarely”, “sometimes” and “often”, into integer scores, such as 1, 2 and 3. This allowed researchers to express responses on categorical measures in terms of difference from the mean response. For example, if the mean response was 2.5 this would indicate most respondents indicated either “sometimes” or “often”, and an individual responding 1 (“rarely”) might score 1 SD below the mean. While this process makes interpretating individual effects more difficult, it was a product of the pipeline process summarising UKBiobank variables for use in MR Base.

Table 3.4. Measurement details for all variables in analysis

Variable	**Measurement details**	**Units***
Alcohol consumption	Response to: "About how often do you drink alcohol?" (units per week)	SD
BMI	Participants weight and height was measured in a testing centre and used to calculate BMI (mass/height2)	SD
Coffee consumption	Response to: "How many cups of coffee do you drink each DAY? (Include decaffeinated coffee)"	SD
CHD	Response to: "Has a doctor ever told you that you have had any of the conditions below?" "CHD (Chronic Heart Disease)" was one of the options listed.	Odds (%)

Depression	Medical history of Major Depressive Disorder	Odds (%)
Diabetes	Response to: "Has a doctor ever told you that you have diabetes?"	Odds (%)
Education	Meta-analysis of questionnaire items asking participants their education history. Measured as years of schooling.	Years
Exercise	Response to: "In a typical WEEK, on how many days did you do 10 minutes or more of moderate physical activities like carrying light loads, cycling at normal pace? (Do not include walking)"	SD
Insomnia	Response to: "Do you have trouble falling asleep at night or do you wake up in the middle of the night?" (Never/rarely, sometimes, usually)	SD
Intelligence	Correct answers given to 13 fluid intelligence questions (https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=100231) within a 2 minute time limit.	SD
Loneliness	Response to: "Do you often feel lonely?" (yes/no)	SD
Neuroticism	Meta-analysis of responses on various personality inventories (including the International Personality Item Pool, Eysenck inventories, the Temperament and Character Inventory, and the Multidimensional Personality Questionnaire)(van den Berg et al., 2014)	Score
Smoking	Self-reported number of cigarettes smoked per day	Cigs/day
Not socialising	Response to: "Which of the following [social/leisure activities] do you attend once a week or more often? (You can select more than one)". (selected at least one category, selected no categories)	SD
Wellbeing	Meta-analysis of questionnaire items which measure positive affect	SD
Worry	Response to: "Are you a worrier?" (yes/no)	SD

3.2.3 Selecting instruments

Genetic instruments for each variable were selected from GWAS summary statistics (Table 3.5). A standard approach was followed for selecting genetic variants robustly associated with each outcome of interest (at genome-wide significance, $P<5x10^{-8}$). Similar to the previous chapter, related variants which occur close to each-other were clumped and represented by one lead variant (up to 10000kb distance). Where information on the desired variants was not available in the outcome GWAS, highly similar proxies were obtained for that analysis ($R^2=.8$). Instruments were also

harmonised prior to analysis, where minor alleles with a low affect allele frequency were used to orient the direction of ambiguous affect allele coding (MAF=0.3) and ambiguous palindromes were excluded. The parameters for clumping, selecting proxies, and harmonisation are the default values in the TwoSampleMR package for *R* (Hemani, Zheng, et al., 2018a), so are considered valid albeit conservative. Full details for each instrument is available on the OSF (https://osf.io/sy3ne/).

Table 3.5 Number of instruments selected for each variable in analysis

Variable	Number of variants selected as instruments
Alcohol	99
BMI	458
CHD	41
Coffee intake	40
Depression	36
Diabetes	39
Education	317
Exercise	18
Intelligence	79
Loneliness	16
Neuroticism	1
Not socialising	10
Sleeplessness	80
Smoking	1
Wellbeing	3
Worry	67

3.2.4 Hypothesis-free network discovery

Hypothesis-free network discovery was conducted to estimate the causal effect of each variable in the network on every other. Each variable was treated as exposure and outcome in turn against every other variable in the network (n=120).

Two-sample MR (Lawlor, 2016) was used to compare instrument-exposure information from one GWAS with instrument-outcome information in another GWAS. Bi-directional MR (Richmond & Davey Smith, 2019) was used to estimate causal effects in both directions, treating the first variable in the pair as the exposure and then as the outcome. Each pairing was therefore investigated for effects in both directions, investigating a total of 240 possible effects.

I estimated causal effects using either the Wald ratio or Inverse-Variance Weighted methods (Smith & Hemani, 2014) depending on the number of genetic instruments available for each exposure. The Wald ratio was used where only one instrument was available to calculate the effect estimate from this single instrument. The Inverse-Variance Weighted approach was used where multiple genetic instruments were available to calculate the average effect of many instruments in meta-analysis.

I investigated whether my instrumental variables met the MR instrumental variable assumptions by inspecting relevant statistics and performing sensitivity analyses. I assessed whether instruments were robustly associated with their exposures by inspecting instrument-exposure associations (R^2 association strength) and signal-to-noise ratios (F statistics over 10)(Burgess & Thompson, 2011). Where analyses were conducted with more than two genetic instruments, making sensitivity testing possible, I assessed whether instruments share no common cause with the outcome, and have no effect on the outcome other than through the exposure: I used Cochrane's Q as an indicator of heterogeneity between effect estimates from different genetic instruments (Burgess et al., 2017). Heterogeneity is the first sign that an effect estimate might be biased by horizontal pleiotropy. This was followed up with comparing estimates from different estimators which make different assumptions about horizontal pleiotropy including the MR Egger (Bowden et al., 2015), weighted median (Bowden, Davey Smith, et al., 2016) and weighted mode (Hartwig, F. P., Davey Smith, G. & Bowden, 2017).

3.2.5 Network analysis

Mediation analysis followed the formula for Product-Of-Coefficients mediation outlined by Burgess and colleagues (Burgess et al., 2015b). I performed this using custom code in R to extract, transform and load MR estimates from network discovery into a matrix of effects containing the effect size of each variable in analysis with each-other. Relationships that did not exceed a value of statistical significance adjusted for multiple testing (Bonferroni: P/n_{tests}), were excluded from this matrix to leave only network-wide significant effects. The information in this matrix is used to identify

variables that might mediate an exposure-outcome relationship in the network, by being implicated both as an effect of the exposure and a cause of the outcome.

The magnitude of a mediation effect is then by decomposing the total effect estimate into an indirect effect (mediated) and a direct effect (the component of the total estimate that is not mediated). This is achieved by multiplying the exposure-mediator and mediator-outcome effects to estimate an indirect effect. The total, direct, exposure-outcome effect is then compared with the mediating, indirect, exposure-mediator-outcome effect. For example, take the following example where exposure X has an indirect effect on outcome Y through mediator Z:

$$x \rightarrow z \rightarrow y$$

The indirect effect ($\Longrightarrow$) of X on Y ($\hat{\beta}x \Longrightarrow y$) is calculated by multiplying together the effect estimates ($\hat{\beta}$) for each step in the mediation analysis ($x \rightarrow z$ and $z \rightarrow y$):

$$\hat{\beta}x \Longrightarrow y = \hat{\beta}x \rightarrow z * \hat{\beta}z \rightarrow y$$

The direct effect of X on Y ($\hat{\beta}x^1 \rightarrow y^1$) is then estimated by subtracting the indirect effect from the original total effect estimate ($\hat{\beta}x \rightarrow y$):

$$\hat{\beta}x^1 \rightarrow y^1 = \hat{\beta}x \rightarrow y - \hat{\beta}x \Longrightarrow \mathrm{y}$$

3.3 Results

3.3.1 Hypothesis-free discovery

Of the 240 possible causal effects between the 16 variables in the network, I was able to obtain effect estimates for 233 (94%) of these. Most of these effects were estimated using multiple genetic instruments (n=203), and some with one instrument (n=30), though analysis was not possible in seven cases where outcome information was not available for the exposure's instruments and no proxies were available.

65 effects reached network-wide significance (Figure 3.5) at a Bonferroni-corrected value ($P<2.15x10^{-04}$). All 16 variables were implicated in a network of inter-related effects with some variables exerting a greater number of effects than others (mean=4, min=1, max=16). Education had the most outgoing effects in the network (n=12) whereas smoking, wellbeing and neuroticism had the fewest (n=0). The reason why some variables had few effects is likely due to weak instruments, which I will discuss later.

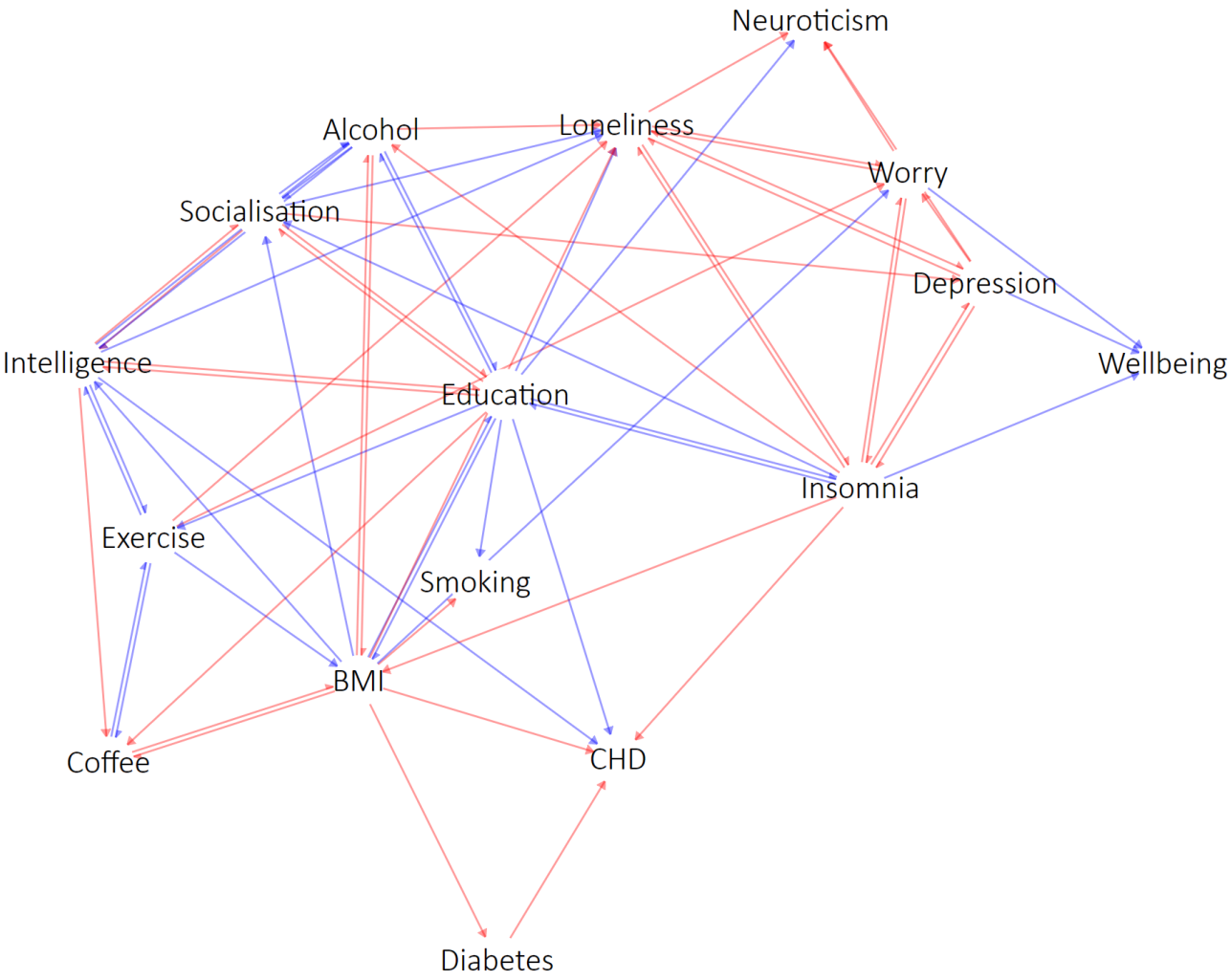


Figure 3.5 Network-wide significant effects between variables in my network. Relationships are colour coded indicating if they increase (red) or decrease (blue) the target variable. BMI = Body mass index, CHD = Coronary heart disease.

Sensitivity tests were calculated for each of the effects in the network to ensure that a valid network of estimates had been obtained for the subsequent mediation analysis. Meeting the three

instrumental variable assumptions for MR would increase confidence that my effect estimates represent true causal pathways. In this section I will outline the results of sensitivity testing.

First, instrument strength was assessed to ensure instruments were robustly associated with their exposures. Although some issues with weak instruments for some psychological variables persisted, the instruments for most variables were strong (Table 3.6). The average instrument explained a modest 1.17% of the variance in the exposure (R^2, min=0.02%, max=6.32%, SD=1.65%) but some instruments for wellbeing, neuroticism and loneliness explained a small proportion of variance (0.02% - 0.13%). Instruments had a good signal-noise ratio, contributing an average F-statistic of 51:1 signal to noise (min=21, max=153, SD=32) and were measured with little error indicated by regression dilution coefficients well above 90% (I^2GX, mean=98%, min=96%, max=99%). Therefore, the instruments for wellbeing, neuroticism and loneliness appear weak but overall instruments showed good instrument-exposure associations, signal-to-noise and low measurement error.

Table 3.6 Instrument strength statistics per exposure

Variable	Number of instruments	Variance explained (r2)(%)	Mean signal-noise ratio (F)	Mean measurement error (I2GX)
Alcohol	99	1.13	83.7	0.98
BMI	458	6.32	54.6	0.98
CHD	41	1.35	60.6	0.98
Coffee intake	40	0.68	54.5	0.99
Depression	36	0.28	36.6	0.97
Diabetes	39	3.15	62.6	0.99
Education	317	2.03	46.5	0.98
Exercise	18	0.15	35.7	0.97
Insomnia	80	0.40	22.6	0.96
Intelligence	79	2.14	39.1	0.98
Loneliness	16	0.13	37.0	0.97
Neuroticism	1	0.02	33.1	N/A
Not socialising	10	0.08	35.4	0.97
Smoking	1	0.22	152.8	N/A
Wellbeing	3	0.02	21.1	0.96
Worry	67	0.62	38.5	0.98

Note: Variance explained describes the proportion of variance in the exposure explained by all genetic variants combined as a single instrument. Instruments which explain less variance in the exposure can be considered weaker and less reliable instruments. This calculation requires information on allele affect frequencies (EAF) but this was not available for depression, diabetes and

neuroticism so I imputed values (EAF=0.5) though these calculations will be less accurate. Mean signal-noise ratios refer to the ability for the average genetic variant to contribute more signal than noise to an analysis, where an F ratio of 10 and above is generally used as a rule-of-thumb that this is acceptable (Burgess & Thompson, 2011). Mean measurement error represents the amount of error with which the average instrument-exposure associations are measured; I2GX values above 90% indicate that measurement error does not present a significant risk of dilution regression estimates towards the null (Bowden, del Greco, et al., 2016). Note that I2GX estimation is not applicable for variables with a single instrument.

Second, the possibility that instruments might act through pleiotropic pathways was investigated. Sensitivity tests were conducted for the 203 effect estimates conducted with more than one genetic instrument. 182 (78%) estimates showed substantial heterogeneity (Q P<.05) and this is a first indication that estimates are at risk from horizontal pleiotropy. One method of assessing the risk of bias in hypothesis-free MR is to compare the results from the main Inverse-Variance Weighted estimation with estimates obtained from sensitivity tests (Hemani, Bowden, et al., 2017). Comparisons revealed that my estimates predicted effects in the same direction as Egger, weighted median and mode estimates, demonstrating that they agreed on whether a change in the exposure resulted in an increase or decrease in the outcome. 132 (65%) of my estimates agreed with the Egger estimate, 156 (77%) agreed with weighted median estimate, and 169 (83%) agreed with weighted mode estimate. Furthermore, the network-wide significant effects which will form my network dataset showed even more agreement with the MR Egger (83%), weighted median (100%) and mode (93%). Therefore, my estimates agreed with estimates from sensitivity tests at a rate higher than expected by chance (50%), and this agreement indicates that a substantial influence of directional pleiotropy is less likely (Lawlor, 2016). Although, some effect estimates were more likely to show disagreement than others, particularly where alcohol consumption or BMI acted as exposures, and this indicates that some of the instruments for these variables act through pleiotropic pathways. However, with so many different comparisons it is difficult to accurately assess the influence of pleiotropy beyond stating that results which reached network-wide significance showed fewer indications of pleiotropy compared with those that did not reach network-wide significance. Therefore, it is more likely that instruments did not act through invalid pleiotropic

pathways though further research should interrogate the potential sources of pleiotropy, particularly for variables including alcohol and BMI.

Third, I tested whether instruments acted directly on the outcome. Instrument-exposure associations were 90% stronger (mean $R^2_{exposure}=1.05x10^{-02}$, min=$6.64x10^{-05}$, max=$6.25x10^{-02}$) than instrument-outcome associations ($R^2_{outcome}=1.10x10^{-03}$, min=$2.8x10^{-08}$, max=$2.09x10^{-02}$) and Steiger testing confirmed that 97.4% of analyses used valid instruments which were significantly better predictors of the exposure than the outcome ($P<.05$, $n_{pass}=227$, $n_{fail}=6$). My network dataset did not include any estimates that failed Steiger testing, but if it did they would have been excluded (see Appendix 3.2 for full results). Therefore, the direction of effects in my network dataset were correctly inferred.

Taken together, sensitivity testing indicates that most instruments appeared strong and valid, but a minority of effects showed indications of directional pleiotropy. To identify whether the source of this was pleiotropy I conducted a follow-up analysis investigating instrument overlap. The analysis (Appendix 3.3) indicated that as many as one in three instruments in the present analysis might be associated with other variables. This suggests that at least some of the indications of horizontal pleiotropy can be attributed to instruments which are at risk of having substantial effects on the outcome through paths other than the exposure, such as directly or through a third variable.

Overall, sensitivity testing indicates the estimates in my network dataset were obtained using valid genetic instruments. However, there were some indications of horizontal pleiotropy, so I will discuss the implications of this in the discussion.

3.3.2 Mediation analysis

Mediation analysis focussed on identifying potential mediating pathways for the causal pathway discovered in the previous chapter between insomnia and wellbeing. Three variables were identified as potential mediators (Figure 3.6). Depression, worry and education were implicated as mediators since they were significantly affected by insomnia and exerted significant effects on wellbeing. Since

wellbeing produced no significant effects, mediation analysis focussed on the effect of insomnia on wellbeing.

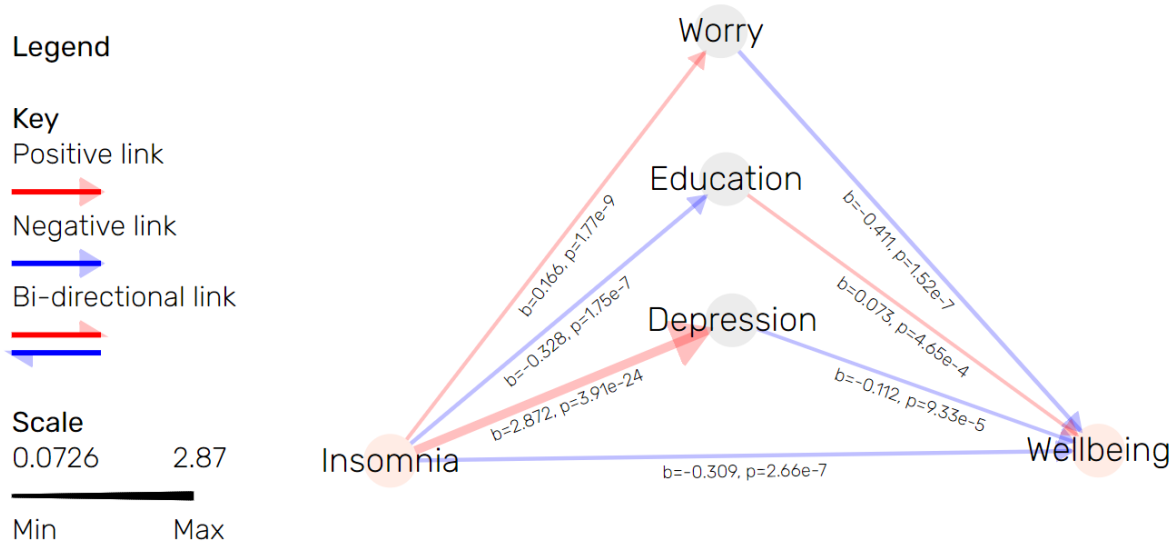


Figure 3.6. Mediation analysis indicates that the main effect of insomnia on wellbeing (grey) is mediated by worry, education and depression (red). Includes sub-network-wide significant effects (P<1.8x10^{-3}).

The total effect of insomnia on wellbeing was decomposed into direct and indirect components. For example, the indirect effect of insomnia on wellbeing through depression as a mediator was calculated by multiplying the effect size for the exposure-mediator effect (b=2.872) by the mediator-outcome effect (b=-0.112) to give an indirect effect estimate (b=-0.32).

Overall, this analysis indicates insomnia acts on wellbeing partly through a large indirect effect mediated by depression (b=-0.32) and smaller indirect effects through worry (b=-0.07) and education (b=-0.02). Accounting for the sum of these indirect effects (b=-0.41), the total effect of insomnia on wellbeing (b=-0.31) can be decomposed into a negligible direct effect (b=0.10). For a full working of these calculations see Appendix 3.4.

3.4 Discussion

In this chapter I obtained an example network dataset for further study. Network analysis consisted of three stages; variables were selected for inclusion, network discovery obtained estimates between 16 variables, and mediation analysis can be used to decompose total effect estimates into

direct and indirect components. The result was a network dataset describing 65 relationships between 16 variables related to wellbeing, physical and mental health. Network structures can be analysed at the micro, relationship and macro levels (Lima, 2013) and I will examine the validity of my network dataset at these three levels. I will draw on previous research (Brown & Knowles, 2020; Hemani et al., 2017; Jansen et al., 2019; Wootton et al., 2018) to demonstrate that my network dataset contains relevant variables (micro-level), has reliably estimated relationships (relationship-level), and a general pattern of network complexity which is similar to previous network datasets (macro-level). I will close this discussion by drawing out the implications of my findings, discuss the strengths and limitations of network MR as a method of discovering network complexity, and highlight some opportunities to make automated hypothesis-free estimation more robust.

First, at the micro-level, I selected relevant variables to include in my network dataset. I selected 14 variables based on previous research related to insomnia and wellbeing. Basing variable selection on previous research ensured that there was evidence that variables were relevant to one-another since variables were already plausibly jointly related to insomnia and wellbeing. This is important because it increased my chances of obtaining a network dataset which consisted of plausible relationships and an overall pattern of effects. Therefore, selecting relevant variables helped ensure my network dataset would be valid for my purposes and demonstrate the type of network complexity present in public health.

Second, at the relationship-level, my network appears to contain reliable causal effect estimates since many have been reported in the literature before. I obtained 64 effect estimates which reached network-wide significance. The reliability of individual estimates in a network has been investigated before by comparing them with previous research (Hemani, Bowden, et al., 2017). I compared effect estimates in my network to ten previous MR studies investigating insomnia or wellbeing (Gao et al., 2019; Jansen et al., 2019; L. Song et al., 2022; Taylor et al., 2014; Wootton et al., 2018, 2020, 2021; Zhang et al., 2019). This comparison (Table 3.7) indicates that 14 of 21

previously reported effects involving variables in my network reached network-wide significance. Therefore, the individual effects in my network appear to be reliable with respect to previous MR research.

Table 3.7 Previous MR estimates and their status as having been corroborated by my network (green) or not (grey).

Effects of insomnia	Causes of insomnia	Causes of wellbeing	Others
Heart disease[1]	Worries[1]	Depression[5]	Exercise → Worry[8]
Worrying[1]	Intelligence[1,2]	Exercise[5]	Smoking → BMI[9,10]
BMI[1]	Education[1,2]	BMI[6]	Smoking → Worries[9,10]
Education[1,2]	Depression[1,3]		Smoking → Diabetes[9,10]
Depression[3,4]	Neuroticism[1]		Exercise → Depression[8]
Diabetes[1]			Loneliness → Smoking[7]
Intelligence[1]			

Sources: [1] (Jansen et al., 2019), [2] (Song et al., 2022), [3] (Gao et al., 2019), [4] (Zhang et al., 2019), [5] (de Geus, 2021), [6] (Wootton et al., 2018), [7] (Wootton et al., 2021), [8] (Choi et al., 2019) [9] (Taylor et al., 2014) [10] (Wootton et al., 2020).

Third, at the macro-level, my network showed similar overall patterns of relationships to previous network MR research. My network indicates that health variables are highly inter-related and have large numbers of relationships with each-other. I estimated that 203 network-significant relationships exist between the 16 variables in analysis, and the number of relationships per variable in my network (12.3) falls within a range of estimates reported by previous research (2.4 – 19.6)(Hemani et al., 2017; Brown & Knowles, 2020). This indicates that my network dataset contains a similar number of relationships to previous research and supports claims that health variables are highly inter-related. Additionally, my network dataset implicated all variables in a single cohesive network structure, and the MR EvE dataset also contains a similar structure where the relationships between variables result in a large web of indirect effects which eventually connect even distantly related variables (Hemani, Bowden, et al., 2017). Therefore, the number and structure of relationships in my network is similar to previous research and this indicates that the structure of my network dataset more likely represents the true structure of network complexity in public health.

In summary, my network dataset likely represents the type of network complexity that would be informative for public health since my variables were relevant, the relationships between them were reliable, and the overall pattern of complexity matches previous network MR research.

3.4.1 Strengths and limitations

The main strength of my design is that I obtained a more easily interpretable network dataset compared with previous research (Hemani et al., 2017) by implementing four criteria in the selection of variables. I will now review three of these criteria and explain how they helped improve the interpretability of my network dataset. I will close this section by highlighting how weak instruments biased my network dataset to be less likely to identify relationships between certain variables.

The first important criteria for improving interpretability was that there must be observational or experimental evidence that variables are related to insomnia or wellbeing. This helped me focus on a selection of 16 variables in a meaningful way, and as a result I obtained a network of 64 network-significant relationships compared with MR EvE (Hemani et al., 2017) which had 2407 variables and found 5660 network-significant relationships. My network is complex enough to demonstrate the network complexity of public health, without being so complicated that understanding would be very difficult. I will come to review some evidence in chapters 4 and 5 which supports this since undergraduate science students were able to learn the relationships in my network dataset and correctly answer questions about it following learning sessions of 5-30 minutes.

The second criteria was that variables must be generally intuitive to a lay audience. The result of this is that the variables in my network dataset are more easily understandable compared with the variables in MR EvE (Hemani et al., 2017) which were not filtered in this way and included obscure metabolites such as urate. I will come to review some evidence in chapter 5 which supports the success of this criteria as well, since undergraduate science students reported that they were able to understand the variables in the network and did not find this to be a source of complication.

The third criteria was that variables must not be duplicates or close analogue of another variable in the network. This ensured that each variable in the network represented a distinct concept and did not present duplicated or redundant information. Browsing the genetic correlations between variables in UKBiobank (https://ukbb-rg.hail.is/rg_browser/) reveal high genetic correlations between many of the variables. This is important because this means that the genetic instruments for these variables in MR would also be similar and produce effect estimates which are similar and do not provide additional information. For example, despite their high genetic correlations with body mass index, body fat percentage (86%), whole body fat mass (90%), waist circumference (90%) and hip circumference (85%) were included in MR EvE as independent variables. By restricting my analysis to variables with lower than 80% genetic correlation I minimised redundant information in my network analysis.

The main limitation of my design is that the instruments available for some of the variables in my analysis were weak. Variables including wellbeing, alcohol consumption, smoking, and caffeine consumption had few SNPs and weak instruments. A consequence of this is that I can be less certain that the relationships in my network represents true causal effects since, for example, fewer causal effects were detected for weak instruments and these may constitute missed true effects. This phenomena has been observed by MR researchers before where poor instrument strength has hampered researchers' ability to detect effects (M. Cornelis & Munafo, 2018; Wootton et al., 2018) for variables like wellbeing (Wootton et al., 2018), alcohol (Pasman et al., 2020), smoking (Wootton et al., 2021), and caffeine (M. Cornelis & Munafo, 2018). This is potentially problematic for my purposes since my network dataset is likely biased towards including more relationships between variables with strong instruments. This is because effects from variables with weak instruments, like wellbeing, were underpowered to detect small differences in outcomes.

Sensitivity tests also indicated that many (1/3) instruments were inter-related and a minority of estimates showed indications of being biased by horizontal pleiotropy (15%). While this is not

entirely unexpected, MR EvE also found wide-spread indications of horizontal pleiotropy (Hemani, Bowden, et al., 2017), it is worth noting that strong instruments are more robust against being biased by horizontal pleiotropy, and for this reason weak instruments present an additional risk to reliability.

The network complexity in my dataset could therefore be improved by obtaining stronger instruments and repeating this analysis to capture a more complete picture of the network complexity between these variables. However, as I detailed in the previous chapter, obtaining stronger instruments for complex traits like wellbeing can be very difficult. Future research could consider managing the risk of weak instrument bias by using PhenoScanner (Kamat et al., 2019) to identify and exclude instruments related to other variables, as well as adopting stringent protocols for diagnosing and treating bias from invalid instruments (Brown & Knowles, 2020; Hemani, Bowden, et al., 2017).

In summary, I obtained a novel network MR dataset which was more easily interpretable than a previous network dataset (MR EvE). Although future research using stronger instruments will be able to obtain a more comprehensive and robust dataset, the current version is likely complex and robust enough to be used as an exemplar dataset to test the use of visualisations and interactive games for improving understanding of, and engagement with a complex dataset.

3.4.2 Future directions

My aim in future chapters is to develop methods for understanding and communicating the network complexity in my network dataset. In the next chapter, chapter 4, I will develop a visualisation tool to help researchers visualise the variables, relationships, and general pattern of effects in network MR datasets. This will be developed further in chapter 5 into a simulation of public health interventions. The mediation analysis method demonstrated in this chapter will be built on to simulate the direct and indirect effects that interventions might have if they improved various variables in my network dataset. In chapter 6 I will also trial a method of using game features to

increase the engagement of undergraduate science students with this simulation model. The dataset obtained in this chapter is therefore a critical material which will enable future research.

3.4.3 Conclusion

In this chapter I used network MR to obtain a dataset which represents the type of network complexity present in public health. I demonstrated that my network is a reasonably valid representation of the network complexity since I selected relevant variables, obtained reliable causal estimates for the relationships between them, and produced an overall pattern of relationships which is similar to previous network research. I also demonstrated that mediation analysis can be used to estimate indirect causal effects. Future research might, however, seek to obtain more comprehensive network datasets by obtaining stronger instruments for variables like wellbeing. The materials I developed in this chapter will be essential for the coming chapters, 4, 5 and 6, where I will build on them to develop methods for visualising, simulating and gamifying this dataset.

4 MiRANA: A tool for visualising network relationships in MR

4.1 Introduction

The previous chapters show that the results of causal analyses can be complex, particularly when applying cutting-edge approaches such as network MR. To date, there is very little software available to help researchers understand these complex results, so in this chapter I develop new open-source visualisation software to do this, drawing on the experiences of Mendelian randomisation researchers at the MRC Integrative Epidemiology Unit, a world-leading centre for causal analysis in medicine.

Network complexity arises in epidemiology when studying the causes and effects of a disease reveals multiple inter-dependent factors. For example, the contributing factors for insomnia include excessive wakefulness at night and anxiety about sleeping, however these are themselves inter-related since a pattern of wakefulness at night could produce anxiety about not sleeping. As demonstrated in chapter 2, network effects can be difficult to understand and present challenges in inferring causality. Network effects are important to understand in part because a full knowledge of them helps predict the effects of interventions (I explore this further in chapters 4 and 5).

Drawing diagrams is standard practice in epidemiology to communicate and interpret relationships (Greenland et al., 1999), and drawing diagrams of networks helps researchers understand the different levels of network complexity, from the micro level, through the relationship level, to the macro level (Lima, 2013). Visualisation helps facilitate micro-level inference by documenting the factors that are associated with a disease. Relationship-level inference is achieved by clarifying whether and how factors are related, and revealing the general pattern of relationships enables macro-level inference. For example, in Figure 4.1 there are five different factors, which both cause and are caused by factors in the network, indicating that our physical and mental health are highly inter-related.

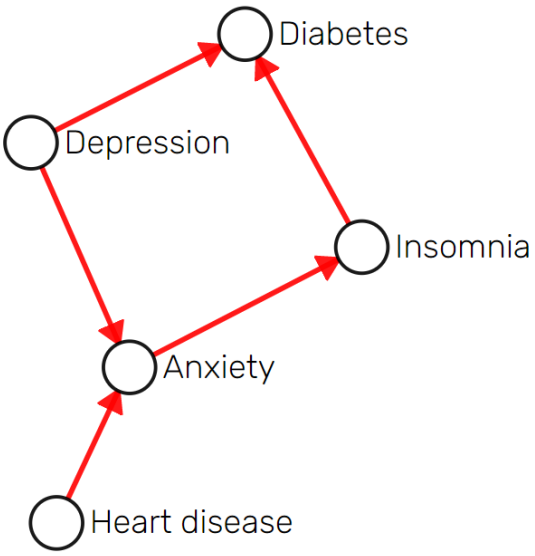


Figure 4.1 Network complexity makes inferring and isolating causal pathways difficult. In this example, the relationship between depression and diabetes is complicated by a second pathway of effect through anxiety, which in turn is affected by heart disease and affects insomnia.

4.1.1 Visualising network outputs from MR

The results of MR studies can often be in the form of complex causal networks. For epidemiologists, it important to understand networks such as Figure 4.1 in terms of the full pathways of effects that diseases, or their treatments, may cause. In MR, this type of network complexity arises when researchers obtain effect estimates between multiple related factors in a single study. To understand the current practice of how researchers use visualisation to communicate and understand networks in MR, I performed a scoping review of 29 MR papers that discuss networks (Appendix 3.1).

A range of methods are used to visualise network relationships (Figure 4.2), but the most common were types of rhizomic "network graphs", such as Figure 4.1. These represent factors with "nodes" (often circles) and the relationships between them with "edges" (lines, or arrows for directional relationships).

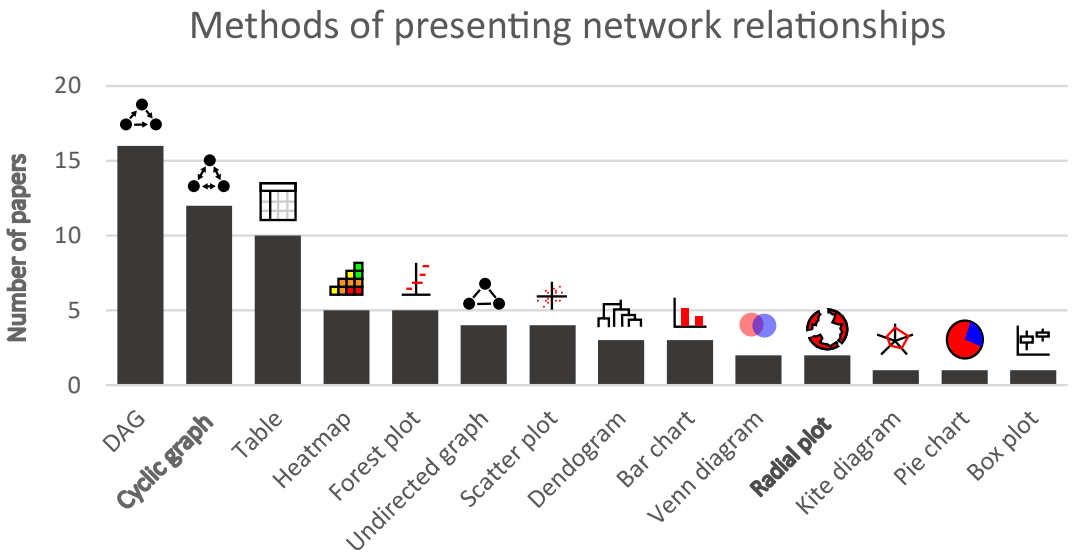
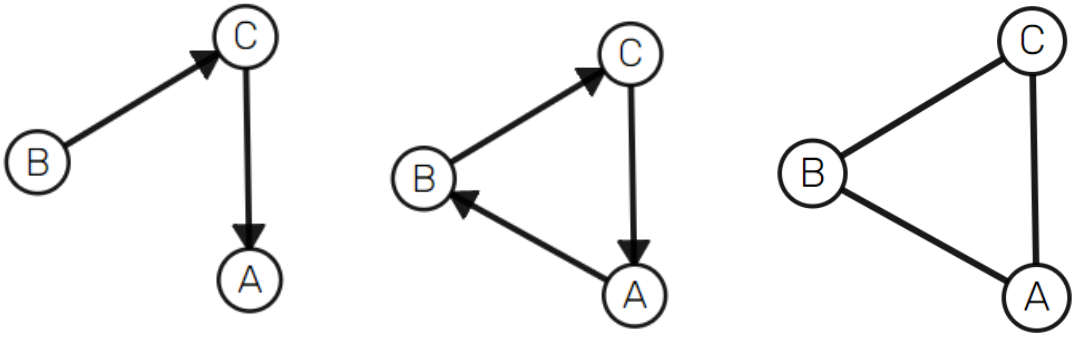


Figure 4.2 A wide range of approaches are used to present network relationships in MR, but the most common are types of network graphs (DAGs, cyclic graphs and undirected graphs).

Three types of network graph ("directed acyclic graphs", "acyclic graphs" and "undirected graphs") have different properties and are used for different purposes (Figure 4.3):



(1) Directed acyclic graph (DAG) (2) Cyclic graph (3) Undirected graph

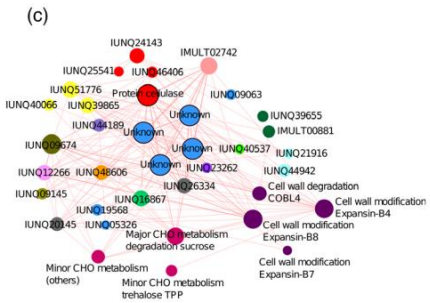
Figure 4.3 The three sub-types of network graph used in the reviewed MR papers: DAGs (1) display causal effects but do not allow cyclic effects, whereas cyclic graphs (2) do, and undirected graphs (3) do not indicate directionality and are used to represent associations where causality is unknown.

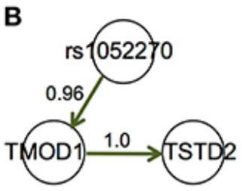
Directed Acyclic Graphs (DAG)s are used to present the design of a study in a way that helps researchers present and test a valid causal hypothesis (Textor et al., 2011). In a DAG, researchers'

assumptions about the relationships that exist between factors are made explicit and open to critical evaluation (Suttorp et al., 2015). Since a cause must always precede its effects, cycles of effects are not regarded as valid causal interpretations and are disallowed in DAGs. The convention for using DAGs to show study designs, rather than showing results, is pervasive and reflects 81% of their use in current practice. Nodes in DAGs are often positioned deliberately using algorithmic or manual arrangement to make the relationships between factors as clear as possible (i.e. a "rigid" layout; see Figure 3.4).

Cyclic graphs are functionally similar to DAGs but can contain cycles. The restriction DAGs impose on cyclic effects is at odds with the reality that many researchers obtain results that contain cycles (41% of papers in this review). Consequently, cyclic graphs are used 30% more often to present results than study designs. Cyclical effects were most common in studies that tested the relationships between many factors, and in large studies factors were often automatically positioned in the graph according to their relationships (for example, using a "force-based" layout; see Figure 4.4).

Undirected graphs are used to present associational (non-directional) data. Unlike DAGs and cyclic graphs, the direction of these relationships is not represented with arrowheads. While MR is a causal (directional) analysis, the data used to inform it is often associational. For example, the co-occurrence of genetic variants, and their associations with phenotypes, can help select and evaluate genetic instruments (these are sometimes laid out in a "circular" arrangement; see Figure 4.4).





(1) Force-directed layout (2) Circular layout (3) Rigid layout

Figure 4.4 Network graphs used in MR papers to present network relationships were arranged using one of three layouts. Rigid layouts, often used for smaller graphs, are deliberately placed to ensure relationships, text and labels are clear (Badsha & Fu, 2019)(Figures 5,7). Force-directed layouts present closely related nodes closer together by simulating a physical system where related nodes attract each other while repelling other nodes from their immediate space (Anacleto et al., 2019)(Figure 5). Circular layouts were used to present many nodes which were related to a single central node (Meng et al., 2018)(Figure 4). For full references see the literature review (Appendix 3.1).

4.1.2 Software for network visualisation in MR

My scoping review of the causal network literature indicates that researchers make particular use of network graphs to represent causal networks, but what software exists for this purpose? From my experience of visualising the network relationships in Chapter 2, I felt that existing software did not meet my requirements, so I performed a second scoping review of software to formally investigate this (Appendix 3.3). From this it was clear that there are both general visualisation tools that can be used to produce network visualisations (e.g., Cytoscape www.cytoscape.com), and specialised tools specifically designed for MR researchers (e.g., DAGitty, www.dagitty.net/). Specialised tools are intended to be more practical and useful for MR researchers, providing a focussed set of core features and an interface which is quicker to learn and use than general tools. However, existing specialised software is primarily intended for presenting study designs rather than presenting results. This leaves a niche for specialised software supporting the visualisation of network relationships in MR results, so I developed the MR Automatic Network Arranger tool (MiRANA) to fill that niche.

4.2 Methods

4.2.1 Establishing requirements

MiRANA was developed to meet eleven requirements (Table 4.1). These include “visualisation requirements” (numbers 1-4) and “practical requirements” (requirements 5-11). I derived

visualisation requirements 1-4 from a summary of features which may be useful based on current practice in the scoping review, and practical requirements 5-8 from the author guidelines for software published in the data science journal Bioinformatics (https://academic.oup.com/bioinformatics). Practical requirement 9 is intended to ensure using MiRANA does not require a substantial time investment. Researchers often need to learn several new software packages to perform complex analyses like MR, so software that is easy to learn is likely to prove popular. Requirement 10 ensures MiRANA can be used without specific prerequisite programming skills since MR is performed using a range of languages (Appendix 3.2) as well as user interfaces such as MR Base (Hemani, Zheng, et al., 2018a), so there is no single standard programming language that MR researchers can be relied upon to know. Requirement 11 ensures my code is open-source, freely available for developers who may want to inspect, modify or borrow from my software.

Table 4.1 Requirements for MiRANA

Visualisation requirements (evidence from review: % of papers which used this feature)
1. Convey properties of nodes (70%). For example, grouping related factors (e.g., by body organ or system), separating nodes representing different types of factors (e.g., proteins and genes), or highlighting important and influential factors.
2. Convey statistical parameters (30%)
3. Support both directional (60%) and non-directional (40%) relationships
4. Support relationships containing cycles (30%)
Practical requirements
5. Should be available for at least three years
6. Should be free to access
7. Should not require users to create an account
8. Should run under nearly all conditions

9. Should not require substantial time investment to access, learn or use
10. Should not require the user to know a specific programming language
11. Should have code published on a freely accessible repository

---

4.3 Implementation

MiRANA is implemented as a website hosted at www.morenostok.io/mirana with a user guide at https://osf.io/tr62v/wiki/home/. It is implemented to meet practical requirements: it will be hosted on this domain for a period of at least three years from launch (requirement 5), it is free to use (requirement 6), and does not require a login (requirement 7). I have tested MiRANA on a wide range of computer systems (requirement 8) including operating systems (Windows, Mac OS), internet browsers (Microsoft Edge, Mozilla Firefox, Google Chrome, Apple Safari and Opera), and screen sizes (800x600px - 2560x1400px).

As is usual with web applications, a Hypertext Markup Language (HTML) document describes the content of the page (text, images), a Cascading Style sheet (CSS) specifies format and design (layout, colours), and JavaScript files provide functionality. The code is published Open Source (requirement 11) on the GitHub repository (https://github.com/CMorenoStokoe/network-mr-vis-tool) which allows others to reuse the code, suggest improvements and contribute to its development. This also opens the possibility of distributing the maintenance of software that is useful to the community among multiple developers. I protected the intellectual property by publishing MiRANA under the Open Source GNU General Public License 3.0 that ensures researchers will always be able to view, reuse and improve upon this software.

4.4 Features

Existing software used by MR researchers offer a range of visualisation capabilities (Figure 4.5), including twelve software packages for network visualisation. Table 4.2 below compares these capabilities to those offered by MiRANA. *Cytoscape* (https://cytoscape.org/) and *visNetwork* (https://github.com/datastorm-open/visNetwork) came closest to matching the set of user requirements I had identified. *visNetwork* met almost all capabilities but required programming expertise in *R* to design network diagrams. *Cytoscape* lacked a default design style which means that users must invest time designing more aspects of the graph. *Cytoscape* also requires a considerable amount of time to download, install, learn and use, although researchers may be willing to invest time in learning a visualisation tool they can use for multiple purposes.

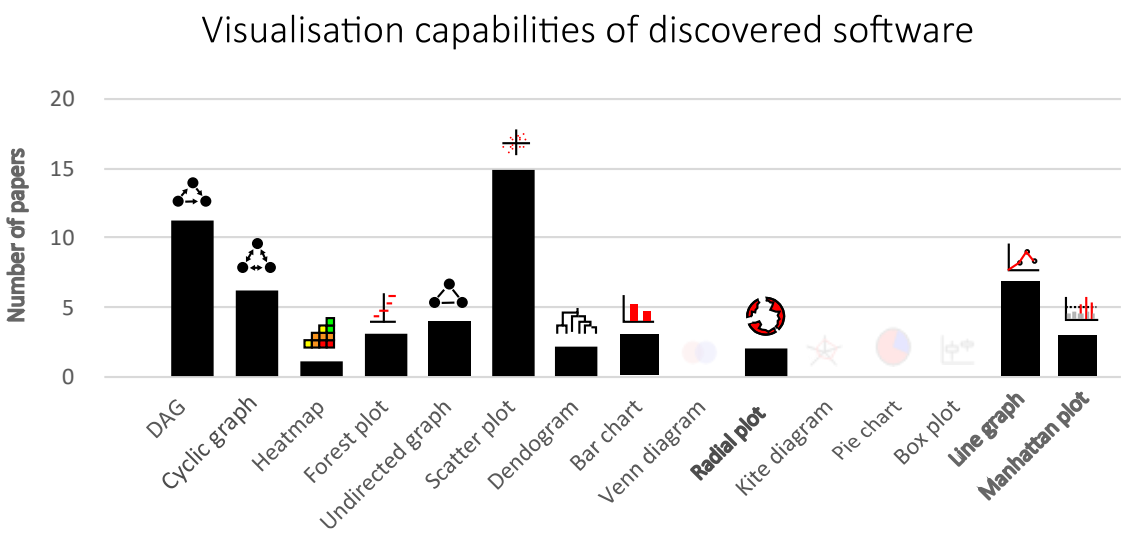


Figure 4.5. The 62 existing software packages identified in the software review (Appendix 3.3) have a range of visualisation capabilities, including twelve packages which can be used to produce network visualisations.

Table 4.2 Overview of features useful for visualising network relationships in MR results. While several software packages exist for network visualisation, these do not support all the features MiRANA was designed to provide (see Appendix 3.3 for included software). Features are indicated as supported (✓), partially supported (~), or not supported (✗) and were selected based on current practice (Appendix 3.1).

Software	Supports non-directional data	Supports cycles	Input data as spread-sheet	Edge weight	Edge labels	Node colours	Re-arrange layout	UI	Default design

MiRANA	✓	✓	✓	✓	✓	✓	✓	✓	✓
MRPC	×	×	×	×	×	~	×	×	✓
DAGitty (web)	✓	~	×	×	×	~	✓	✓	✓
DAGitty (R)	✓	~	×	×	×	~	✓	×	✓
DiagrammeR	×	✓	×	✓	✓	✓	✓	×	✓
visNetwork	✓	✓	✓	✓	✓	✓	✓	×	✓
Cytoscape	✓	✓	✓	✓	✓	✓	✓	✓	×
Epigraph DB	×	✓	×	×	×	✓	✓	✓	✓
D3	✓	✓	×	✓	✓	✓	✓	×	×
Tetrad	✓	~	✓	×	✓	×	✓	✓	✓
ggdag	×	✓	×	×	×	~	×	×	✓
dagR	×	×	×	×	×	~	✓	×	✓
shinyDAG	×	×	×	×	×	×	×	✓	✓
Count of features	6	6	3	4	5	5	9	5	11

4.4.1 Data input

Most existing software packages allow researchers to visualise their own results, although some are intended purely to display a curated database of pre-computed results (e.g., EpigraphDB), or cannot be used to visualise MR effect estimates (e.g., MRPC). MiRANA allows researchers to input their own data as a spreadsheet. In most cases researchers can directly import and visualise their raw results file since the data input format is the same as common MR outputs (e.g., from the popular *R* package *TwoSampleMR:* https://mrcieu.github.io/TwoSampleMR/).

The JavaScript module PapaParse (https://www.papaparse.com/) is used to detect how characters in the selected file are encoded and read its contents (e.g., ANSI or UTF-8 encoding). After parsing, data

are converted into a network graph (node-edge) format. However, this process relies on users inputting properly formatted data, so I took steps to mitigate the risk of ill-formatted data by using identical column names to the *TwoSampleMR* package, providing clear instructions in the user guide, and making the internal model tolerant of slight mis-specifications (e.g., the order of columns). If MiRANA detects ill-formatted data, users are presented with an error screen specifying the columns they need to fix.

4.4.2 Network visualisation

MiRANA can produce three different types of network graph—DAGs, directed cyclic and undirected graphs—which users can download as image files (for example, Figure 4.6). Graphs include a legend showing how direction and magnitude of effects are represented, similar to, for example, DAGitty.

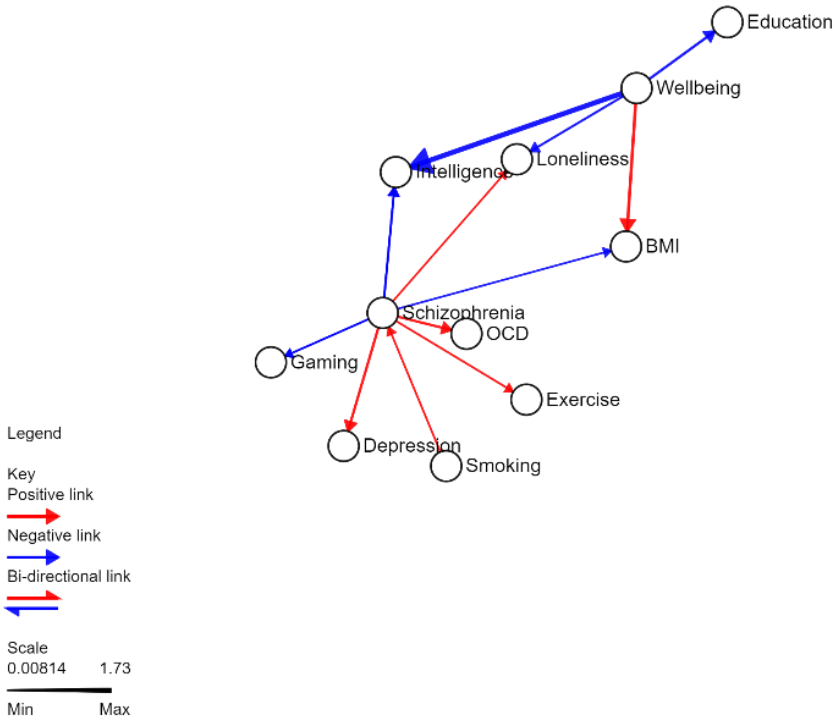


Figure 4.6 MiRANA outputs a downloadable image of a network graph with an accompanying legend. The legend updates as users customise its design. In this example, a min-max scale shows how edge widths scale with effect sizes (discussed later). The example data here indicates that wellbeing causally affects some factors including loneliness, and schizophrenia causally affects other factors, such as exercise and smoking.

I implemented the network graph visualisation using the *D3* JavaScript library (www.d3js.org). *D3* can be used to produce force-directed layout systems, where a physics simulation attracts nodes

together based on the relationships they share, while pushing nearby nodes away to avoid overlap. The output is a graph where closely related factors appear closer together, presented as a Scalable Vector Graphic (SVG) element, which represents shapes as vectors in a coordinate system. To represent relationships, edges are drawn as lines between points on the circumferences of related nodes. The modules *CanvasToBlob* (https://blueimp.github.io/JavaScript-Canvas-to-Blob) and *FileSaver* (https://github.com/eligrey/FileSaver.js) are used to convert the on-screen image into a Portable Network Graphics (PNG) image for users to download.

4.4.3 User interface

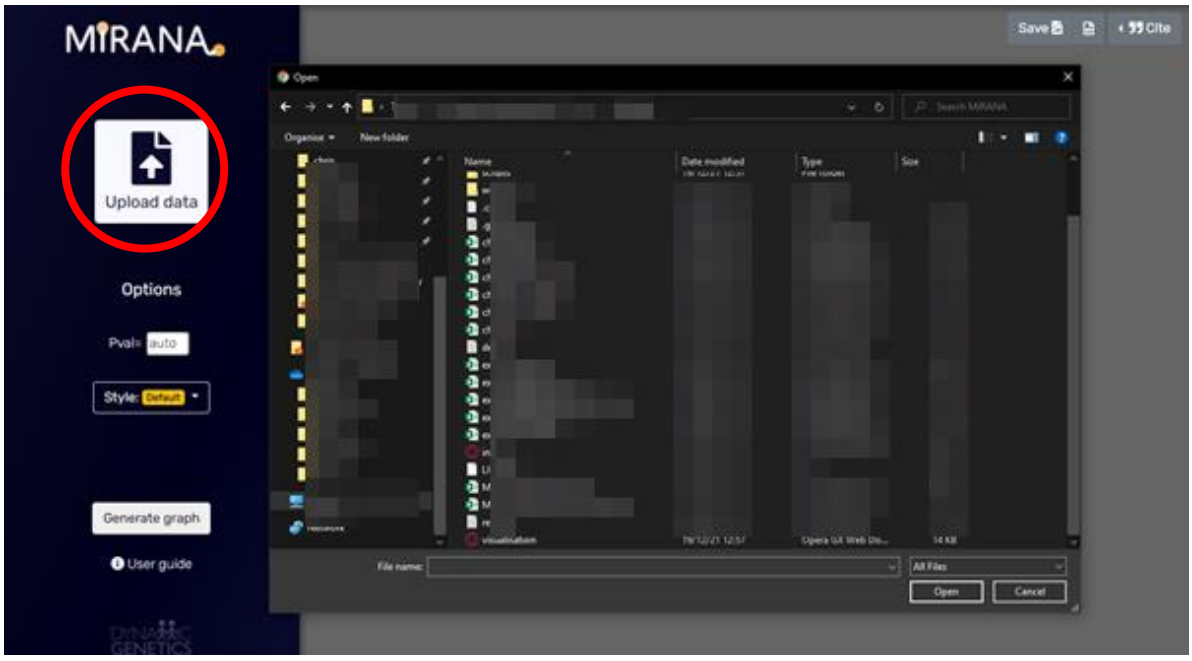
I designed the graphical user interface to allow researchers to visualise their results in four steps without programming (Figure 4.7). This was intended to appeal to researchers both with and without programming experience by being quick to use. This feature was relatively uncommon among the existing software packages I had identified (33%).

The interface includes buttons, toggles, and forms for text input, which change aspects of the graph using the *JQuery* JavaScript module (www.jquery.com). The user does not have to engage with every option presented in the interface since non-essential functions default to pre-set values (e.g., colours, sizing, layout of the graph). The interface also gives users the ability to re-arrange nodes in the graph by dragging them into position. This can be used to achieve circular and rigid layouts (in addition to the force layout generated by default) and is supported by most existing software (75%).

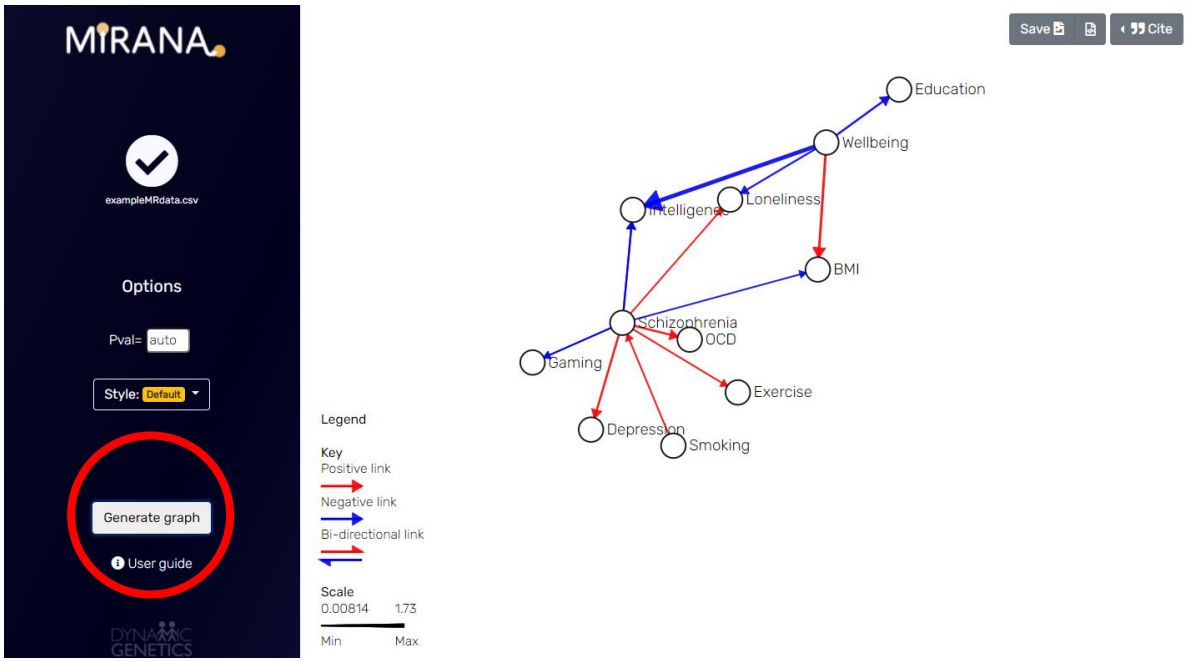
Steps to produce a graph in MiRANA

Step 1: Load website.

Step 2: Press “upload data” and select file.



Step 3: Press “generate graph”



Step 4: Press “save” and download image.

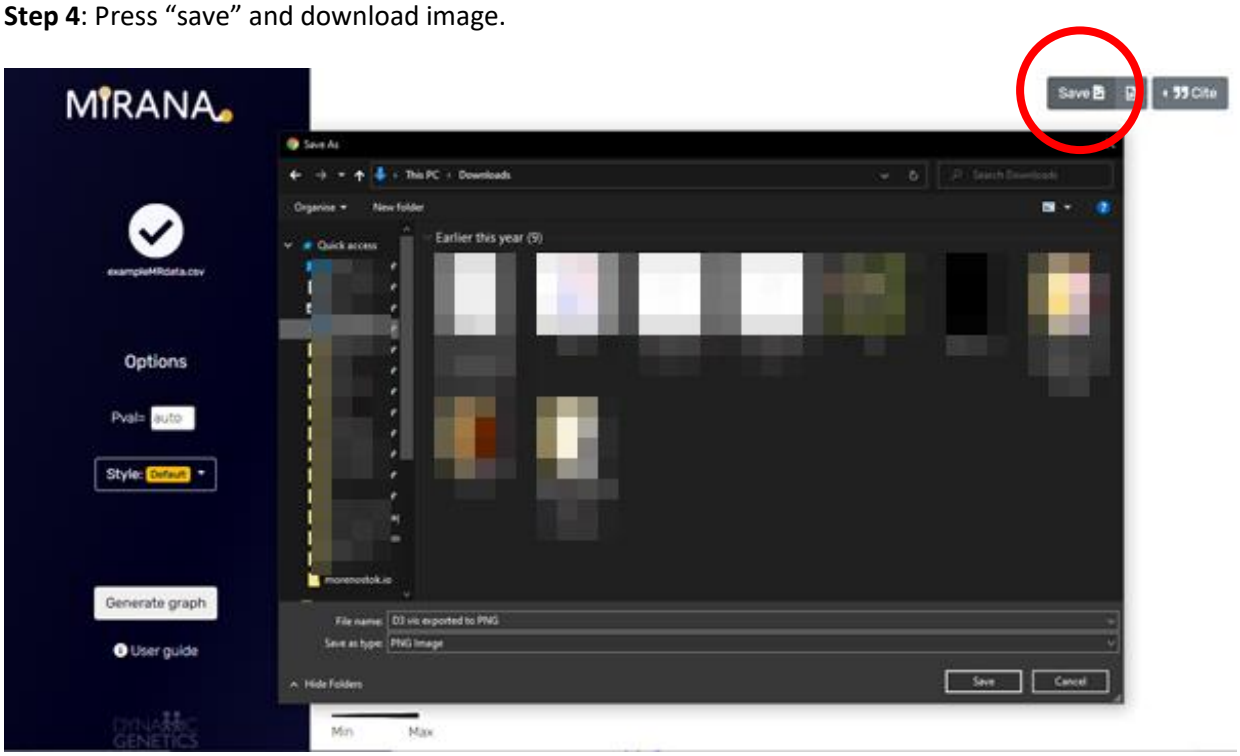


Figure 4.7 Users are able to visualise results quickly in four steps.

4.4.4 Advanced settings

MiRANA has an advanced settings menu (Figure 4.8) with many additional features that allow researchers to customise the design of the graph and visualise different types of data:

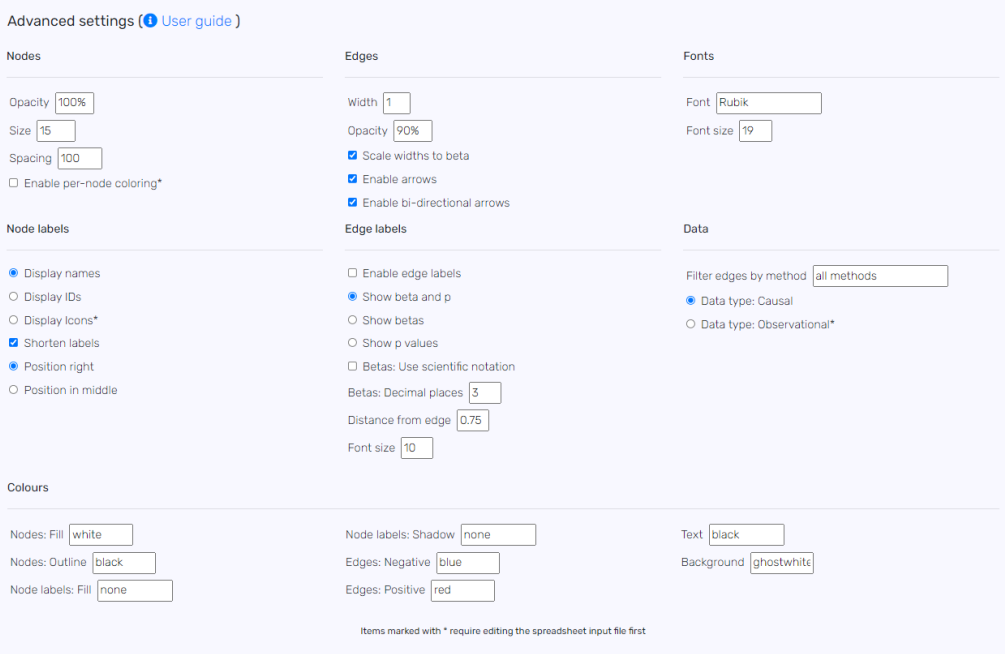


Figure 4.8 The advanced settings menu allows further customisation of the graph's appearance.

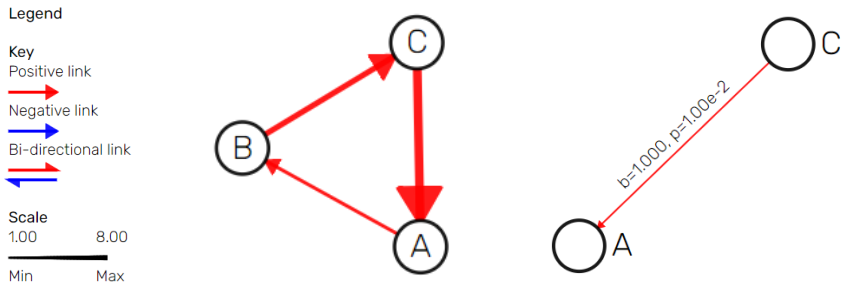
Nodes

Users can design nodes using various options including sizing and colouring. Nodes are identified with labels giving the name of the factor they represent and these can be customised in terms of their size and positioning, and by using methods for shortening long variable names to fit the graph. Furthermore, nodes can be individually coloured to show that different nodes belong to different categories; this feature is restricted by many software packages (42%), which only allow the user to mark nodes as either exposure or outcome, but not any other category.

Edges

Users can customise the edges in the graph so that they more clearly convey information about relationships. The widths and opacity of edges can be configured to help visualise large networks with many intersecting edges. On-the-fly filtering can be used to focus on the most important effect estimates which exceed a particular *P*-value threshold, or were obtained using a particular method (e.g., main IVW effect estimates rather than sensitivity tests). Colours can be used to represent the valence of relationships (i.e., positive or negative) and arrowheads indicate the direction of causal

effects. Bi-directional effects can be represented with either a two-headed arrow (↔) or with two offset arrows (⇆). While all existing software packages supported directional arrowheads, fewer supported removing them for displaying observational data (50%). Last, users can present statistical parameters associated with relationships, such as effect size (beta) and *P* value, either as labels written along the length of edges, or conveyed through the thickness of edges (Figure 4.9). This feature is particularly important for presenting results since these parameters help interpret strength of evidence, although relatively few existing software packages give users the option to label and weight edges (33%).



(1) Edge weights (2) Edge labels

Figure 4.9 MiRANA supports two methods for conveying the effect size and statistical significance of relationships. Edge weights scale the width of each edge proportional to its effect size (b); the example on the left includes edges A→B (b=2), B→C (b=4), C→A (b=8). Edge labels state beta, *P*, or both values as is shown in the example on the right.

Fonts, data and colours

There are additional utilities for specifying fonts and colours, as well as adjusting the strength of the force-based spacing between nodes to either space them further apart or closer together.

4.5 Discussion

MiRANA was designed to fit into the current practice of visualising network relationships in MR research. I set out a number of visualisation requirements based on current practice to convey the factors and relationships in networks. A wide range of input data is supported including cyclical and non-directional data. Labelling and categorising nodes helps researchers document the factors in

networks, and conveying the properties of edges between them helps clarify their relationships. Arranging networks using force-based layouts also helps reveal the general pattern of relationships.

Although MiRANA provides a range of visualisation features beyond those provided by much of the existing software, this was not unique since some software offered similar features (e.g., *Cytoscape* and *VizNetwork*). In these cases, the main differentiating feature of MiRANA is that it is intended to be very easy for researchers to use as part of their MR research process. Specifically, the user interface is quick to operate and designed based on existing software to minimise its learning curve. This is particularly important since some software packages require a considerable time investment to learn to use effectively (e.g., *Cytoscape* and particularly *D3*), and implementing seemingly simple features can be complicated. Furthermore, the MiRANA website interface was designed to reduce barriers to access and so has no requirement for pre-requisite programming skills, is free to use, and does not require users to create an account, which frees them from signing any access or account creation agreements. MiRANA is therefore a functional network visualisation software package which MR researchers can use to present their results and may find more practical than existing alternatives.

4.5.1 Future directions

MiRANA improves upon existing software by providing additional features and aiming to provide an improved user experience. Along with similar projects aimed at making MR data visualisation simpler, such as EpiViz (https://mattlee.shinyapps.io/EpiViz/), MiRANA represents the current culture of iteration and improvement in MR software development. However, in its next phase of development, it will be more important than ever to be clear on the principles driving software development and to understand the users (Redwine & Riddle, 1985; Shaw, 2001). Much of the software for MR are formally documented with information on background theory, software aims and features (e.g., *DAGitty*: http://www.dagitty.net/manual-3.x.pdf), but user studies are not generally conducted or at least not published. While the existence of popular software demonstrates

that user studies are not essential to develop software which researchers will use (e.g., *TwoSampleMR*), they can greatly increase the chances of producing valuable and widely used software. Involving users from the beginning to share their needs and provide feedback has been found to produce better designed software (Zwass, 2010), which is more innovative and better meets the needs of its users (Kristensson et al., 2008).

The next steps for MiRANA will be guided by continuing our dialogue with users, eliciting and responding to feedback, as well as understanding the intended users and how they interact with the software. For example, in the next chapter I will describe interviews and a focus group conducted to understand the processes followed by MR researchers and elicit feedback while playing a game based on MiRANA.

4.5.2 Beyond the rhizomic network

I have made design decisions based on my interpretation of the literature, and these explore only one way of representing network data. Traditional methods for presenting effect estimates, such as scatter plots, are not designed to present networks of effects and it is often cumbersome to combine multiple effect estimates into a single plot (Bowden et al., 2018). I concluded from my scoping review that network graphs are a popular solution to network visualisation amongst researchers. However, current practice does not necessarily represent best practice and it has been argued that network graphs present an over-simplified view of problems in epidemiology (Krieger & Smith, 2016). Other methods for visualising networks exist, including radial network graphs (Figure 4.11), heatmaps (Figure 4.12), and hive plots (Figure 4.14). Each approach has advantages and disadvantages when visualising large networks of effects, identifying influential factors, and producing reliable graphs.

Visualising large networks

When presenting vast amounts of information, researchers face a trade-off between detail and scope (Figure 4.10).

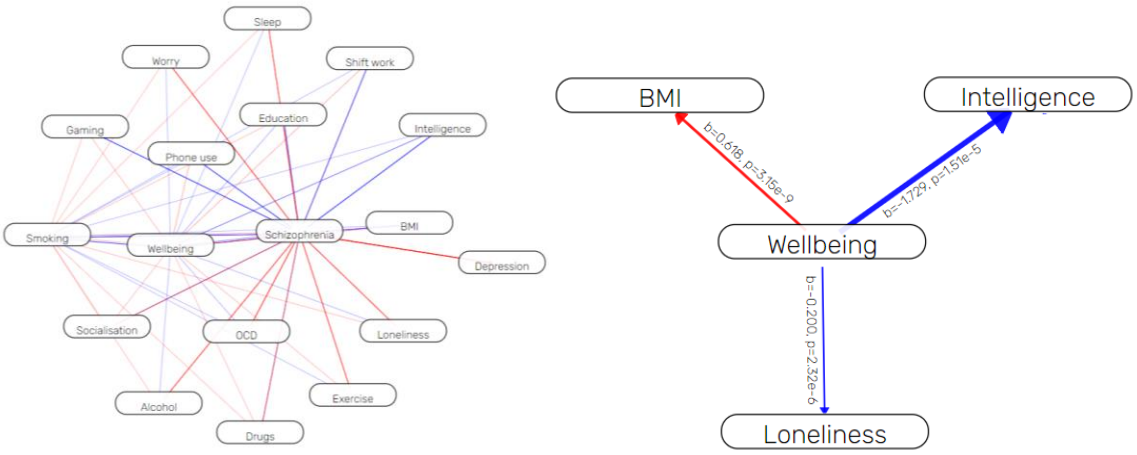


Figure 4.10 Researchers face a trade-off between detail and scope when presenting results using network graphs. While the whole dataset could be presented (left), it is likely easier to present detailed information about a smaller subset of this data (right).

The flexibility of network graphs allows researchers to present small and large datasets since nodes, edges and labels can be resized to fit the entire network in a single image. In current practice they are used to present a range of datasets describing anywhere between a few factors and a few hundred factors. However, when visualising larger networks network graphs face some limitations. First, shrinking elements of a graph to make a large network fit in one image makes it more difficult to distinguish and identify nodes and edges. Second, text labels occupy free space in the graph can only be shrunk as far as they are legible which means that often it becomes less practical to include detailed labels in large graphs. Radial network graphs have been proopsed as a solution which allows researchers to present large datasets within a single graph (Krzywinski et al., 2009). In constrast to the rhizomic network graphs reviewed above, radial network graphs arrange nodes along the circumference of a circle and relationships are drawn as lines between them. In current practice these graphs are already in use by MR researchers and two papers featured these in the literature review. For example, Figure 4.11 shows a type of radial network graph, a Chord diagram, which Bien and Peters use to present hundreds of relationships between genetic variants and various phenotypes in a single graph (Bien & Peters, 2019). Additionally, Chignon and colleagues use one to understand how genes of interest are related to each other which can help determine whether genetic instruments are independent (Chignon et al., 2020).

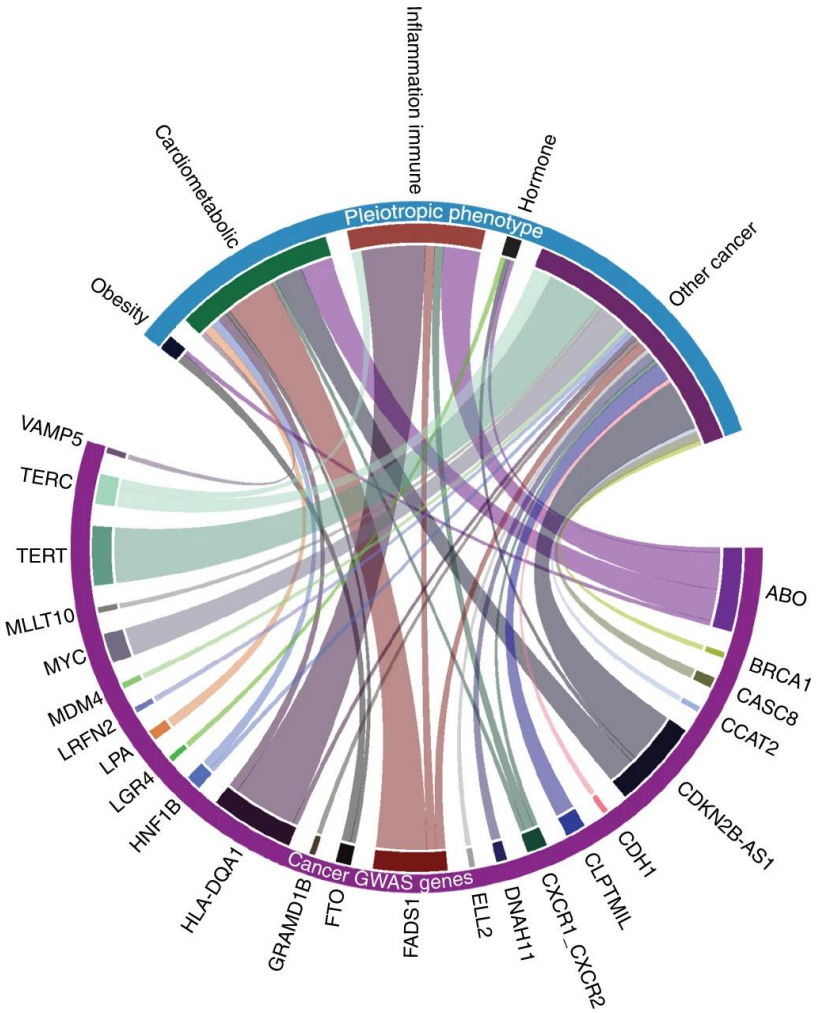


Figure 4.11 Radial network graphs are a solution which can be used to more clearly present large networks within one chart. In this example, Bien and colleagues (Bien & Peters, 2019) present the relationships between genes related to cancer and other phenotypes which may confound MR studies. Genes are and phenotypes are arranged around the circumference of the circle and lines (chords) are drawn between them to indicate relationships. Thicker chords indicate more numerous relationships for different loci within a gene.

Identifying influential nodes and edges

Network graphs may be particularly useful for identifying the most influential nodes in a network. A force-directed layout and edge weighting can be used to highlight strong relationships. As we have already seen, a force-directed network graph arranges nodes by their relationships with other nodes. From this layout nodes naturally arrange themselves into neighbourhoods with nodes they are related to, and separate themselves from nodes with which they share no relationship, or a more distant relationship. It is easy to identify the most central factor in a network graph since it will

often be positioned in the centre of the graph, with strongly related factors arranged nearby. Another method, edge weighting, can be used to distinguish particularly strong relationships by marking them with thicker lines than weaker relationships. However, other graph types can use design elements to emphasise key relationships as well, for example heatmaps, which are particularly effective for highlighting the strongest relationships in a network. In heatmaps, nodes are arranged along the x and y axes of a matrix and the relationships among them are indicated with cells that are coloured according to their values (Figure 4.14). Heatmaps can clearly communicate which are the strongest relationships even among large numbers of relationships, since the matrix structure can shrink or grow to scale with the size of the dataset. Additionally, they have the advantage that they are simple and can be made in word processing and spreadsheet programs such as Microsoft Word and Excel. Heatmaps are in regular use as part of current practice and in the literature review they were the most common type of visualisation behind network graphs (appearing in five papers). Therefore, although network graphs can clearly present the most influential factors in a network, there are other methods such as heatmaps which are easier and may be more effective for this purpose.

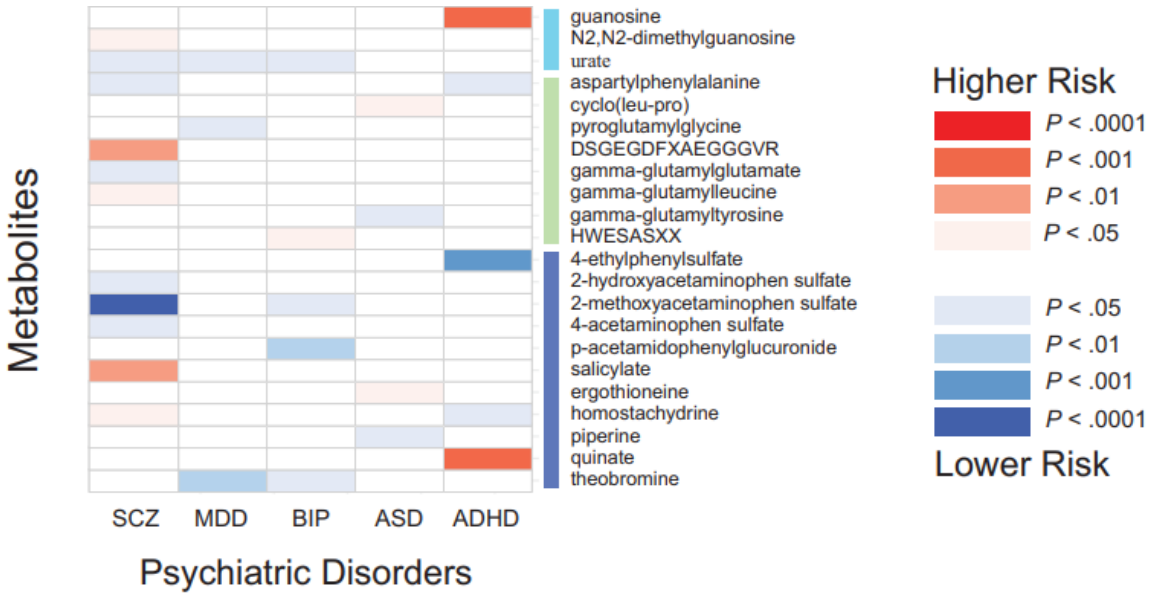


Figure 4.12 Heatmaps clearly highlight the strongest relationships in a network. In this example Yang and colleagues use a colour scale from blue to red to highlight which metabolic factors which present the highest

risk for psychiatric disorders (Yang et al., 2020). SCZ = Schizophrenia, MDD = Major Depressive Disorder, BIP = Bi-Polar Disorder, ASD = Autistic Spectrum Disorder, ADHD = Attention Deficit Hyperactivity Disorder.

Reliability

Network graphs often use force-directed layouts to arrange nodes, using a physics system to calculate a layout solution in real-time. The advantage is that it generates a solution to any dataset it is presented, and reacts to the user manually re-positioning nodes in the graph in interactive graphs (as we will see in Chapter 4). However, it does not guarantee a consistent and reliable layout solution for a given dataset. This is important because the layout of a network graph gives information about the overall pattern of relationships in the data and inconsistency makes it difficult to compare the patterns of force-directed graphs. The Hive plot has been presented as a solution to reliably produce network graphs. In a hive plot, nodes are arranged along perpendicular axes and relationships are drawn as lines between them (Figure 4.13). A layout for the nodes is procedurally generated, so can be reproduced consistently and the patterns of different datasets can be compared reliably. Furthermore, the designers argue that these plots make individual relationships clearer (Krzywinski et al., 2012). Hive plots are not common in MR but have been used to compare different patterns of relationship across multiple networks. In their study investigating longevity, Bou Sleiman and colleagues use hive plots to understand how different sets of genes, used as genetic instruments, are related to each other (Bou Sleiman et al., 2020). The authors compare three different plots to understand how the relationships differ in different tissues (biceps, liver, heart). However, hive plots come with a disadvantage in that they are an unfamiliar way to present data and the reader must first be coached to correctly interpret them. By contrast, network graphs are familiar and intuitive to MR researchers so do not require coaching to interpret.

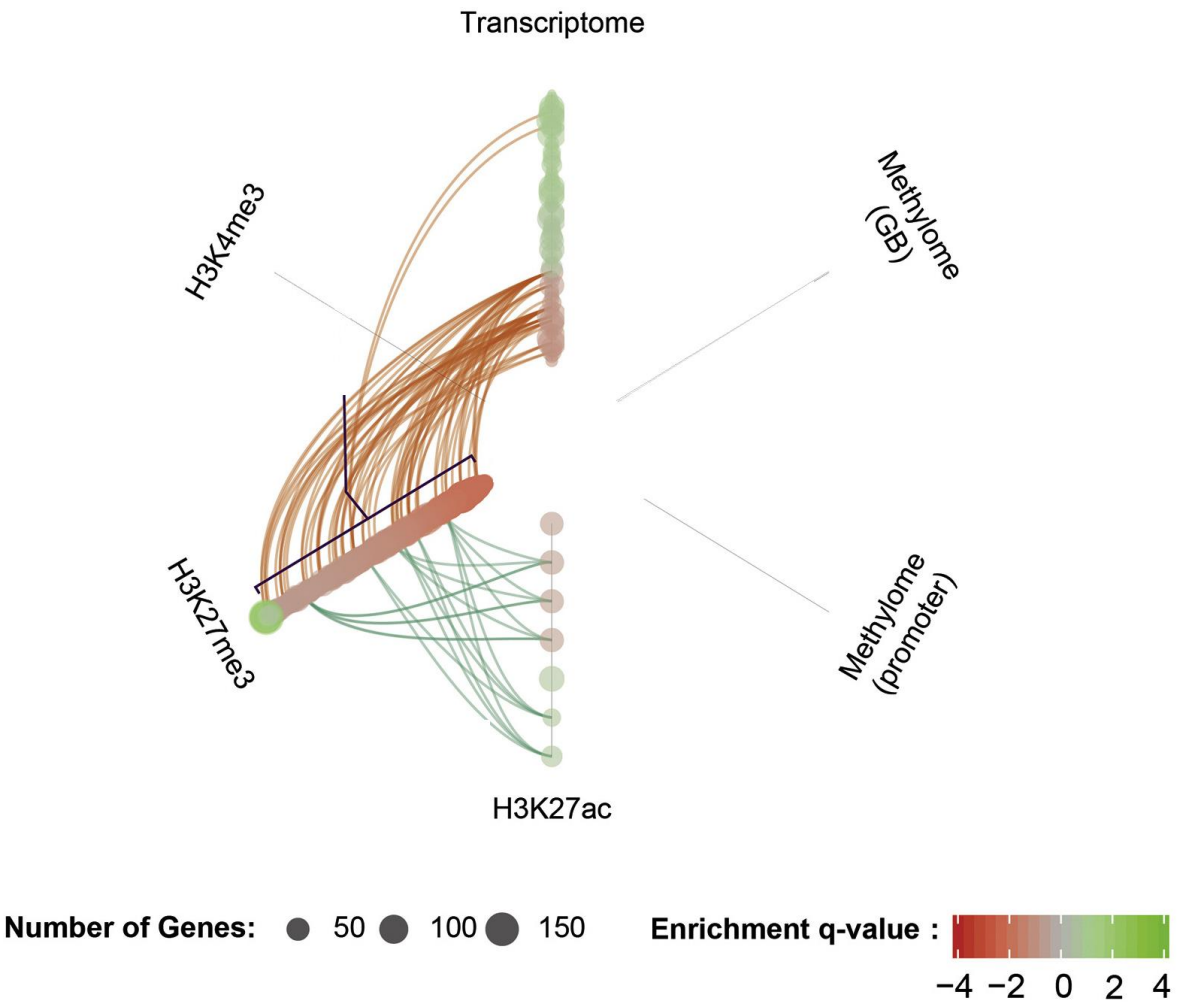


Figure 4.14 Hive plots are an alternative to network graphs and can produce reliable and reproducible outputs, but require additional learning to interpret. In this example (Bou Sleiman et al., 2020) demonstrate overlap between gene sets related to ageing. The circular nodes represent groups of genes. These nodes are arranged by category along perpendicular axes (in this case, whether they belong to gene set H3K27me3 or H3K27ac, or the transcriptome). Two important messages are that one gene set (H3K27me3) contains many more genes and shares many relationships with both the other gene set (H3K27ac) and transcriptome. This is an important graph in the paper because these gene sets were later used as genetic instruments in MR so the overlap between gene sets suggests pleiotropic pathways. However, this may not be immediately clear to the reader without prior understanding of how hive plots work.

4.5.3 Conclusion

MIRANA is a novel tool that allows MR researchers to visualise network relationships in their results. Users can upload MR results, visualise them, and customise the results in terms of layout, labels, sizing, and colouring. I conducted a scoping review to derive requirements for this software and to help ensure it would fit within current practice in terms of capabilities and practicality, accessibility and time effectiveness. This will form the basis of a continuing conversation with MR researchers to

test my assumptions and develop further approaches. But can we do better than simply presenting data for exploration, by motivating and guiding exploration through approaches borrowed from games? In the next chapter I will build on the MiRANA software to explore whether adopting game development practices can take us beyond visualisation in understanding complex causal relationships.

5 Turning a causal health network visualisation into a data game

5.1 Introduction

In the previous chapter I developed a way for researchers to visualise network relationships in their data. While visualising my network MR dataset helped convey the complex pattern of relationships, it was still difficult to intuitively understand how variables would affect one another when intervening in the network. In this chapter I describe the development of a data game to help understand this complexity, which will be tested experimentally in the next chapter. I will introduce concepts underlying data games, describe how I developed a simulation of public health interventions, and explain how I transformed this into a playful experience using game features.

5.1.1 Decomposing games into observable components

I will start this game development chapter by further defining what is meant by a "game". In the introduction I explained that games can be defined as "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome" (Salen & Zimmerman, 2003). However, decomposing this definition into more specific components that can be observed and measured is difficult and there is some disagreement. Researchers often hold one of two different ontological beliefs: that games are defined by material properties, or that they are defined by the subjective experiences of players (Denzin & Lincoln, 1994). These views can be considered materialist or realist views and lead to different definitions of what the constituent ingredients of games are. In this section I will discuss these two different views and explain why I adopt the realist view that games are defined by rules which are experienced as playful by players. This definition is important because it will inform the approach I adopted in this chapter for developing a data game.

Materialists argue that including game-like rules into an activity transforms it into a game. This view is popular with experimental researchers who gamify non-game tasks, such as learning (Qin et al., 2010) or who compare games to non-game equivalents, such as simulations (Joldersma & Geurts,

1998). Objective characteristics make it easier for researchers to demonstrate that one activity is a game and one is not, so theoretically helps establish how games differ from other activities. For example, “gamification” is defined as the addition of game-like rules, such as scoring, to a non-game context (Deterding et al., 2011) . However, this view is problematic because reviewers of gamification research often note that the resulting “games” do not incorporate game features meaningfully, are not experienced as playful by players, and are often not considered games (Looyestyn et al., 2017; Sardi et al., 2017).

Realists argue that a game is characterised by game features that produce a playful experience. This view is popular with game designers (Salen & Zimmerman, 2003) and researchers of game mechanics (Hamari et al., 2014; Malliarakis et al., 2014; Sailer & Homner, 2020), who more closely investigate the psychological actions of gameplay. The concept of game features is used to refer to methods used by players for interacting with the game world (Sicart, 2008), and meaningful game features can be considered those that contribute to players’ ability to win or lose the game (Salen & Zimmerman, 2003). Critically, this definition captures both the execution and result of a ruleset; an activity which includes game-like rules but is not experienced as fun is not considered a game. This realist approach is supported by findings of “player types”. Player types describe that different people have different experiences of the same ruleset. For example, in competitive games, some players strive to beat others, while others are disinterested and frustrated by competition (Bartle, 1996). This is relevant because it demonstrates that there are other aspects, personal characteristics and preferences, which influence how we play within the objective rules of a game . Furthermore, this approach still presents observable components for experimental research, with the stipulation that game features are necessarily tied to play experiences. Game features have been linked to the specific play experiences they evoke, such as competition (Lucero et al., 2013), and are understood to appeal to core psychological needs (Marczewski, 2018). I will adopt this more nuanced definition of games and this will inform the approach I take in designing my data game. I will continue to explain game features and evidence my decision-making process against this

definition and emphasise the importance of meaningful play experiences. Having established what makes a game, I will now specifically explore what makes data games.

5.1.2 Data games

A “data game” is a type of game which uses gameplay to facilitate players’ to exploration and understanding of an underlying dataset (Friberger et al., 2013). Players of data games expect that gameplay will help them understand topics, allowing them to experiment with it in a safe and playful setting (Simonofski et al., 2022). The two main goals of these games are to represent data realistically (Malliarakis et al., 2014) and to transform it into something that can be played with (Friberger & Togelius, 2012). They can be considered a sub-class of serious games, as described in the introduction chapter, which use gameplay for some purpose other than pure entertainment.

An early example of a simple data game is “Bar Chart Ball” (Friberger & Togelius, 2013). The goal in this game is for players to guide a ball from one end of a bar chart to the other by selecting different views of a dataset (Figure 5.1). To succeed at Bar Chart Ball, players need to learn which views of the data will lead to changes in bar height that will send the ball in the required direction. This means that familiarity with the dataset leads to a better score. Games like this exemplify the principles behind data games because data is transformed into a playful, interesting and informative experience. Players are encouraged to explore the underlying dataset in order to find a solution, and in doing so the players learns about the patterns within the data. It serves as a simple example that data can be represented in a game realistically while transforming it into gameplay..

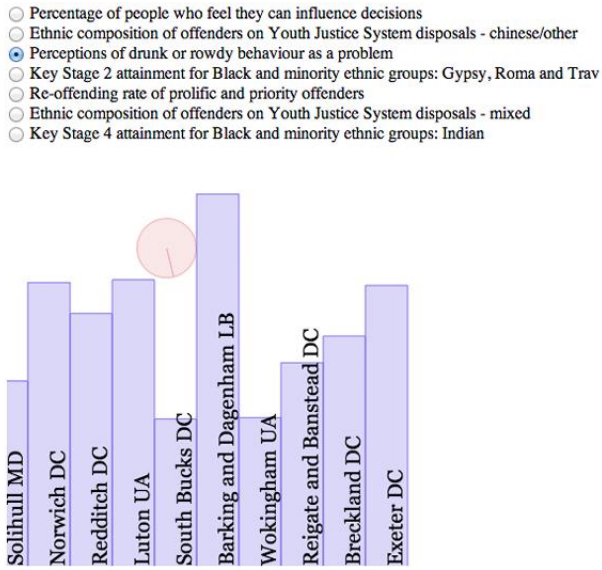


Figure 5.1 In the data game Bar Chart Ball, players use the selection form at the top of the screen to select views of a dataset with different distributions to push a red ball from left to right.

Data games can be described with some specific properties which relate to how they represent a dataset in the game world. While any game can be described as having properties like genres and rules, an important aspect of data games is how they represent data. Two important aspects are the original source of the data and how it is represented in the game (Friberger & Togelius, 2012).

The source data describes the original dataset the game is based on. It can be described in terms of datatype (e.g., numeric) and topic (e.g., politics). For example, the data source in Bar Chart Ball was numeric information on demographics. This information describes the original data, and this is important because it to some degree influences the type of games that can be based on it. For example, one game series generated adventure stories and murder mysteries from descriptions in Wikipedia of people, places and objects (Barros et al., 2015, 2016).

The second aspect is how this source data is represented in the game by making selections and transforming it. Data selection is used to present a subset of data to the player. This does not change the original dataset, but different views of the same dataset can provide different messages or play experiences. For example, the data game Open Source Monopoly generates properties on a Monopoly board from open source data about communities, and different settings can provide different layouts and resulting play experiences (Friberger & Togelius, 2012). Data transformation is

used to alter data before presentation to the player. This transforms data into game features, such as generating the board in monopoly. As I covered in the previous section, creating a game requires a meaningful play experience, and often transforming data is an essential step in making gameplay engaging, which has parallels in more conventional forms of data analysis or visualisation.

Data are often transformed to ensure games are balanced. “Game balance” ensures that the game is neither too easy nor too difficult for players, and this is particularly important with datasets that can introduce highly random or polarised data points. For example, generating realistic maps in the resource-management computer game Civilisation can result in one player commanding all the resources and while this may reflect a real pattern in the data, this can result in unfair and unfun gameplay (Barros & Togelius, 2015). While games do not necessarily need to be symmetrical, problematic distributions and data points can negatively affect the play experience. Achieving game balance requires experimentation and refining the process of data transformation so that it is both balanced and faithful to the source data.

Data can also be transformed in ways that ensure the choices players can make in the data are interesting. This has been noted before as a particularly important aspect of data game design (MacKlin et al., 2009). Exploration and experimentation are large components of gameplay, so it is important to ensure players find their interactions with the data interesting (Salen & Zimmerman, 2003). The grand strategy game “Europa Universalis 4” (Paradox Interactive, 2013) demonstrates how varied gameplay mechanics can keep complex data interesting. Players take the role of a head of state for a country in 1400s Europe. Gameplay takes a free form, players are given a sandbox in which they take a variety of political, diplomatic, military and industrial actions that influence how their country develops over time. Players win by developing their country the most before the 1700s. Different countries have different resources, in-game missions, and events which give players reasons to play each country. This gives players a variety in game mechanics, and incentive to explore the underling historical simulation further. Without this variety, the underlying information

would not be represented in such an interesting way. Data game designers therefore need to strike a balance where data are accurately represented but lead to interesting gameplay.

In summary, data games can transform data into engaging experiences where players explore and experiment with a source dataset. However, the way data is integrated requires careful balancing of data selection and transformation processes to ensure data is represented in interesting yet accurate ways. In the next section I will review some game design frameworks that will help me overcome challenges like these and design an effective data game.

5.1.3 Designing data games

I will use several frameworks to structure my game design process and ensure I develop a game with the desired play experience. I have selected approaches that are likely to help researchers understand games from an academic perspective and to introduce systematic rigor into the game design process (Fullerton, 2018; Hunicke et al., 2004; Salen & Zimmerman, 2003; Werbach & Hunter, 2012; Yardley et al., 2015). They can be grouped into two broad categories, game design frameworks and objective-oriented frameworks, where the former focusses on how to produce engaging gameplay, while the latter focusses on achieving objectives. In this section I will introduce these frameworks and provide a broad overview of their values and processes. I will refer to the frameworks in this section and provide more detail in a later section where I describe my game design process.

“Rules of play” (Salen & Zimmerman, 2003) was the first work to provide an academic framework for studying games. It defines and describes games, explores how they work, and establishes some key ideas in game design. One important idea is that games can be delineated into a set of rules, the play experience that results when players interact with the rules, and the wider context for play which influences players’ expectations and motivations. Another important idea is distinguishing between game design and game development, since the former focusses on designing the actions players will engage with during game play, and the latter focusses on technically implementing this in practice.

The “playcentric approach” described in Game Design Workshop (Fullerton, 2018) details a framework for completing game design. It has rigorous processes and emphases the importance of systematic methods, documentation, and player feedback through playtesting. It is common to start game design with a period of planning, establishing aims and integrating theory as to how they might be achieved.

Objective-oriented frameworks introduce top-down constraints and guide the design process to achieve a pre-decided outcome (Liapis et al., 2019). In design frameworks for serious games the main aim is to achieve a desired play experience theorised to achieve outcomes such as learning (Aleven et al., 2010). The “Six Steps to Gamification” (Werbach & Hunter, 2012) is a particularly influential framework which focusses on taking a non-game task and adding game elements to enhance outcomes. This process of gamification involves selecting discrete game features, integrating them into the non-game task, and testing that the game achieves its objectives. This is important for experimental research where a game is compared to a non-game, and a clear understanding of what differentiates them is essential to constructing an experimental manipulation. However, the focus on creating a gamified app often results in less playful experience as few game features are implemented compared to a full game.

The “person-based approach” (Yardley et al., 2015) is a framework for designing behavioural interventions. This framework is intended to facilitate the development of virtual interventions that modify individuals’ behaviour. Though not originally intended for game design, its stages of user research will prove useful in adding academic rigor to my process. This framework provides systematic methods for collecting and processing user feedback, and the practice of setting clear goals for each stage of design which provide the criteria for assessing progress.

In summary, there are few academic game design frameworks so I will borrow elements from four frameworks. Rules of Play (Salen & Zimmerman, 2003) and the Playcentric Approach (Fullerton, 2018) will help me produce engaging gameplay, while the Six Steps to Gamification (Werbach &

Hunter, 2012) and the Person-Based Approach (Yardley et al., 2015) will help me achieve specific outcomes. I will now explain what type of data game I will make and the outcomes I hope to achieve.

5.1.4 Making a data game about the network of human health

Few games have experimented with representing real-world health data in gameplay. I did not identify any data games based on public health datasets, but I will give three examples of previous health games that have some data game features (Free Ice Cream, 2017; Milstein et al., 2010; Ndemic Creations, 2012). In this section, I will explain how these games have integrated public health data into gameplay, highlight the results, and draw out opportunities for further study.

"HealthBound" (System Dynamics, 2010) is a simulation of public health intended for policy makers. It was developed using data from public health research detailing how various variables impact the effectiveness and affordability of the US healthcare system (Milstein et al., 2011). This data is transformed into a model which simulates how changing various variables, such as the cost of health insurance, might impact the effectiveness of the system. Users experiment with these variables and attempt to identify the optimal set of health policies. This gameplay helped players appreciate the complexity of healthcare (Milstein et al., 2009) and form specific hypotheses about healthcare reform (Milstein et al., 2010). However, though the authors call it a game, it does not meet the definition I have adopted here, since it does not contain game features, such as a progression system, levels, an in-game narrative nor scores. HealthBound therefore demonstrates public health data can be accurately represented in an interactive simulation, though it does not explore the effects of adding game features.

"Plague Inc" (Ndemic Creations, 2012) is a commercial pandemic simulation game intended for a broad audience. It is based on mathematical models describing the variables that contribute to the spread of infectious disease. Although the precise data they are using to inform these models are commercially sensitive, notes from a closed meeting indicate that it uses a standard infection rate model that adjusts the reproductive rate for biological and economic variables (Centers for Disease

Control and Prevention, 2013). These data are transformed in gameplay where players attempt to infect the world by "mutating" a pathogen to be as infectious as possible, such as enhancing the ability to spread by sneezing or through animal vectors. The game world responds to the players by implementing barriers that the player must overcome, such as closing airports to prevent transmission. By overcoming these barriers, players appreciate the variables that contribute to disease spread in the real world. This has been demonstrated by the World Health Organisation who used the COVID-19 expansion as part of an effective awareness campaign (Ndemic Creations & World Health Organisation, 2021). Its style of presenting health information has also been used as a learning aide in medical communication courses (Cheng et al., 2018). However, while the data are based on real-world variables, they are extensively edited and balanced so that the game is enjoyable to play. For example, the rate and effects of mutations are greatly exaggerated and unrealistically represent this aspect of disease for the purpose of making gameplay more enjoyable. Therefore, although Plague Inc uses game features to effectively engage players in real-world data, the data are transformed to a degree where it gives inaccurate insights into the real world phenomenon.

"Playable Data for Human Health" (Free Ice Cream & Davis, 2018) was a prototype data game and precursor to my thesis. The source data are from the MR EvE network dataset that estimates the causal effects among health variables (Hemani, Bowden, et al., 2017). These data were transformed into a simulation of public health, modelling how changes to one variable would affect the others, as a public health intervention might. This gameplay was based on a previous game which engaged attendees at the UN Global Festival of Ideas with a network of the connections among the UN's seventeen sustainable development goals (Free Ice Cream, 2017). Like in that game, the gameplay centred around experimenting and exploring how various policies would achieve different effects in a system of related variables. This style of gameplay has proven effective at engaging citizens to search for information and hypothesise the best solutions and policies to solve problems (Free Ice Cream & Overseas Development Institute, 2017). However, in this prototype the dataset was not

transformed to make gameplay more accurate and interesting. I covered one issue in a previous chapter (3), that the MR EvE network was particularly difficult to understand due to variable selection, including duplicate and analogous variables. Another issue was that the source data were not transformed in ways that made it easier to represent in the game. For example, estimates were not adjusted for units and this resulted in variables with large units, such as mass (kg), having apparently exaggerated effects compared with smaller unit variables, such as blood mineral concentrations (grams/ml). However, Playable Data for Human Health provided a starting point for developing a data game to model the network complexity in public health, and the present chapter will advance this by transforming similar network MR data into an effective gameplay experience.

In this chapter, I will iterate on previous games and develop a data game that represents the network complexity in public health. I will draw on established frameworks to conduct a systematic game design process and, in line with my definition of games, I will develop game features that produce the desired play experience. The benefits that game features provide will be investigated in the next chapter (6) where I will conduct an experimental study comparing game and non-game. I will start by describing how I developed a non-game simulation of public health, and then how I developed this into a game. I will conclude with an evaluation of my game design.

5.2 Developing the simulation

I developed a simulation with the aim of allowing users to explore and experiment with interventions in a network model of public health. The simulation I will describe below is therefore not intended to show an exhaustive and accurate medical model, but rather to model of the type of dynamic and complex network relationships that exist in public health. In this section, I will outline how I built on my previously obtained network dataset to develop an interactive experience where players explore the relationships within by experimenting with virtual interventions.

An intervention simulation estimates how an intervention to improve a health factor such as depression might improve others, such as wellbeing. In my model, I would estimate how changes to

one variable would affect others in my network dataset. I started by importing the network dataset I developed in Chapters 2 and 3 because this provided information on which variables were causally related to one-another. All the variables in my network graph are inter-connected in a single network structure, known as a “connected graph”, and this resulted in a complex pattern of relationships (Thulasiraman, 2011). In this simulation, I will also extend my study of indirect effects from Chapter 3 to include not only first-order interactions, involving a single mediating variable, but Nth-order interactions, involving multiple mediators. In order to capture these complex causal pathways I used a network traversal algorithm. Network traversal algorithms employ a ruleset to traverse a network from node to node and discover pathways in a systematic manner. I employed a “breadth-first search” algorithm that navigated my network one node at a time until all pathways were discovered (Figure 5.2)(Skiena, 2012).

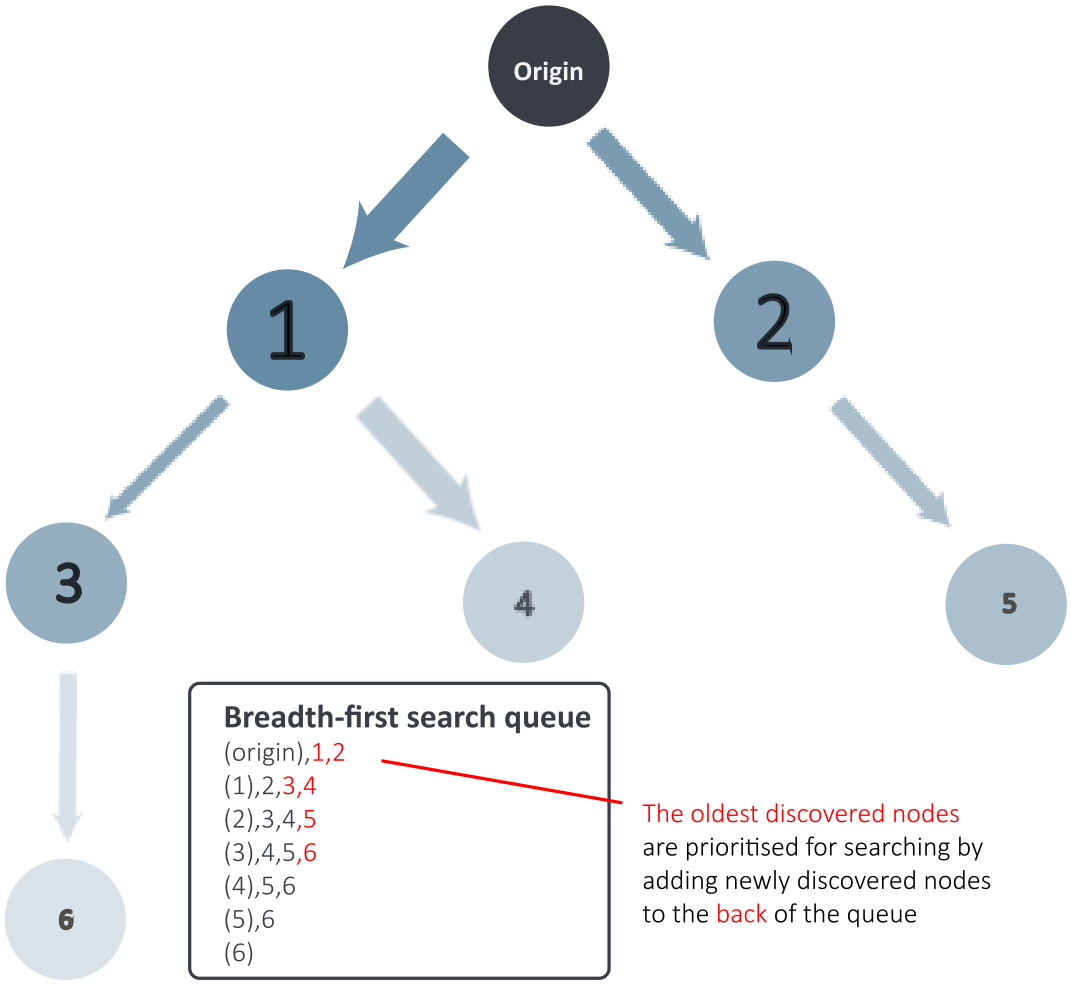


Figure 5.2 A breadth-first search identifies edges between nodes in a network by searching through a network in turn, prioritising searching the oldest encountered nodes first and adding newly encountered nodes to the end of a search queue.

In implementing my breadth-first search, I encountered one problem that I needed to address. My network dataset contained relationships where some variables were both the cause, antecedent to, and effect, proceeding other variables. These are problematic for two reasons. First, they would produce cycles which endlessly trap a simulation of effects (e.g., wellbeing → insomnia → wellbeing etc.). Second, cyclical relationships are not a valid interpretation of my data since, as I detailed in a previous chapter (4), a valid causal effect cannot have an effect which precedes its cause (Suttorp et al., 2015). Therefore, I modified my search to only add nodes to the search queue if they had not been already searched. Effectively this removed causal relationships which would cause a loop, as the final link in the chain. How to manage loops in a simulation is an open issue and I will later discuss how a future model might more completely model cyclical relationships in health.

After identifying how each variable in the network was related, I estimated the effects they might have on one another. To do this, I used a "belief propagation" paradigm where changes to one node spread, or propagate, to all connected nodes (Pearl, 1982). This had the desirable effect of spreading changes to one variable to all variables that are causally connected downstream. Each node was given a default prevalence level, its population mean value in GWAS, and this prevalence level could be changed by players interventions. MR effect estimates were used to calculate how much changes in variable prevalence would affect related variables (Figure 5.3).

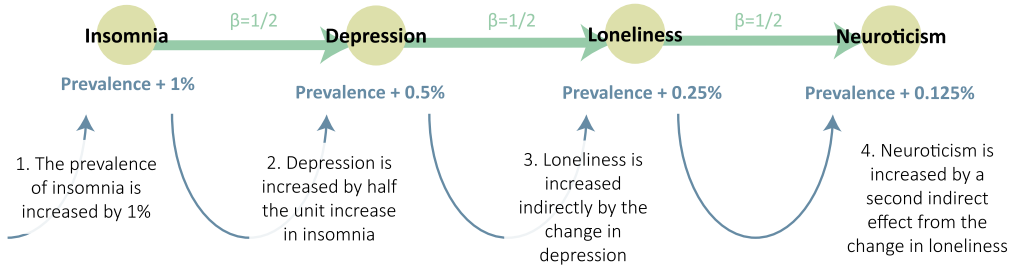


Figure 5.3 The model in my simulation propagates changes in one variable to all related variables in the network

I tested the effectiveness of my model using "Trémaux trees" (Even, 2011). Example trees were constructed to demonstrate all the different types of relationships that my propagation model could encounter in my data. My model returned correct results for network traversal and propagation for a range of network structures (https://www.morenostok.io/mendel/test/propagation.html). For an example see Figure 5.4.

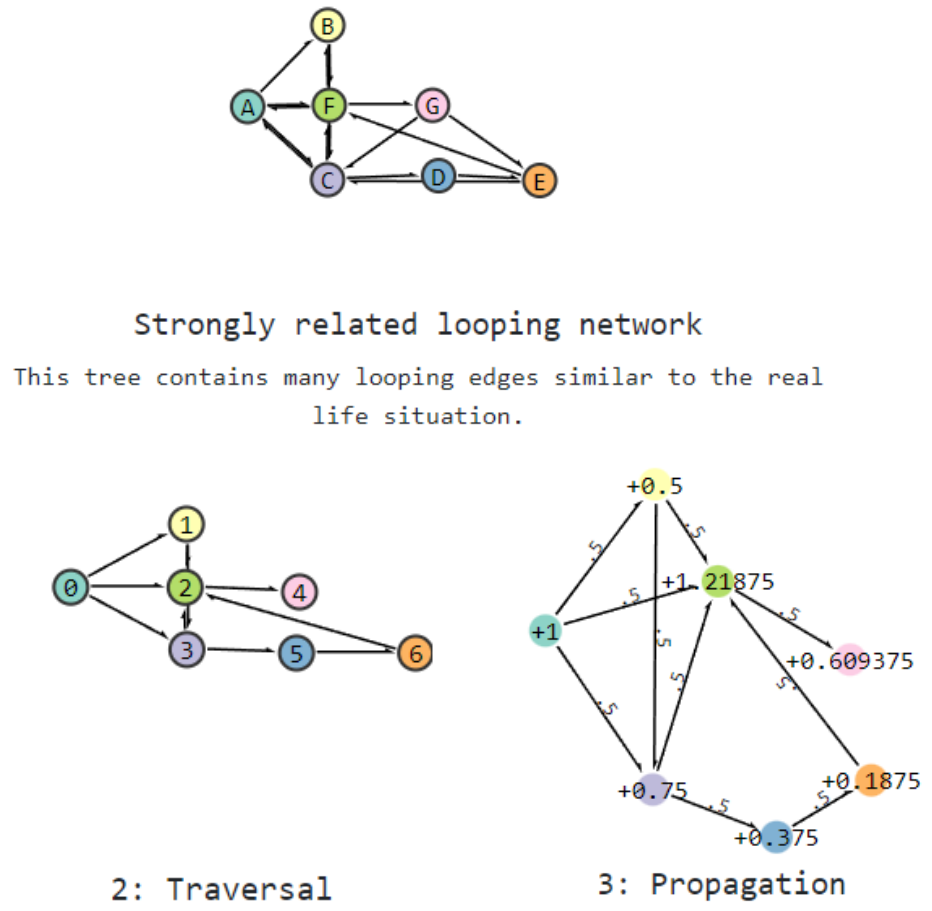


Figure 5.4 Testing revealed that my model effectively traverses and propagates network structures. This example shows the most difficult test including a "strongly related looping network" structure (1st image at the top) designed to test whether my model correctly traverses (2nd) and propagates (3rd) the type of complex network structures in my dataset. Image three shows the results of a change to node A of +1 propagated through a network with beta weights of ½.

I finished the simulation by developing controls for the user to select which factor to intervene on, and a visualisation view of the data (Figure 5.5). Animations and labels were used to convey the complex changes that occurred when interventions were made (Figure 5.6). The final software is available from https://www.morenostok.io/mendel/interactiveVisualisation.html.

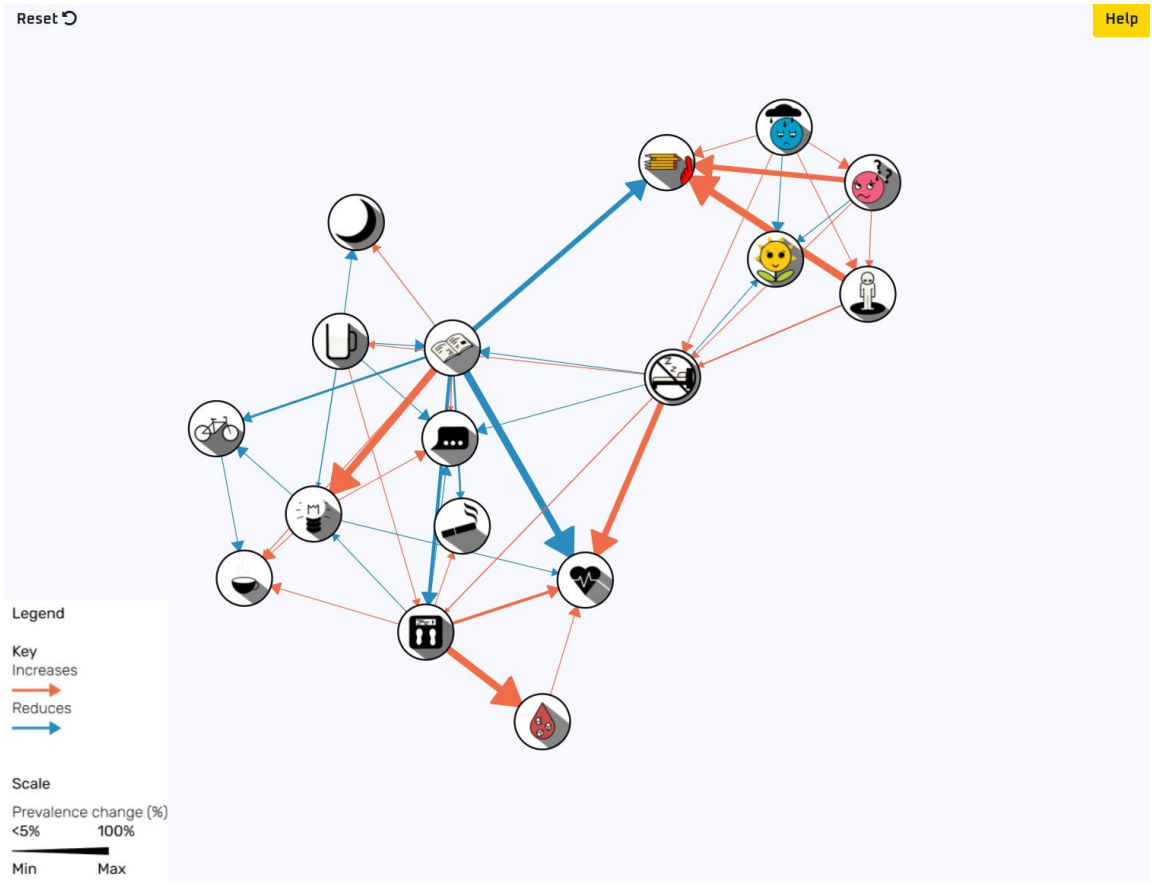


Figure 5.5 The simulation presents a view of the data and allows users to simulate the effects of interventions.

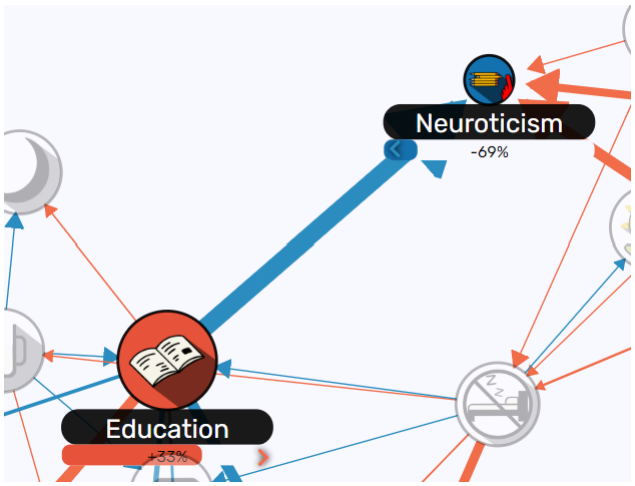


Figure 5.6 Animations and labelling convey changes in the simulation model. In this example, a 33% increase in education has reduced neuroticism by 69%.

5.3 Turning the simulation into a game

After developing the simulation, I turned it into a data game by following a process based on the frameworks I introduced previously (Fullerton, 2018; Salen & Zimmerman, 2003; Werbach & Hunter,

2012; Yardley et al., 2015). In this section I will describe three stages that I used to plan, design and develop the game (Table 5.1). I will refer to some user feedback where the full transcripts and minutes are available in Appendix 6.2.

Table 5.1 Stages in my game design process

	Aims
1) Planning stage	Describe the main outcome(s) Specify the desired player experience to achieve the outcome(s) Describe the intended players and their requirements
2) Design stage	Identify game features to achieve the intended player experience
3) Development stage	Implement game features in a way that achieves the intended player experience

5.3.1 Planning stage

In my first stage, I planned the game by defining my intended outcomes, gameplay characteristics, and audience. These would help guide me during the design process and form criteria against which success would be measured (Yardley et al., 2015).

Outcomes

My game was intended to achieve two main outcomes which would allow me to formally test the differences between game and non-game software in a subsequent chapter (6):

1. Players will explore the relationships in my network dataset
2. Players will contribute datapoints in a data collection exercise

Intended play experience

I hypothesised that these outcomes could be achieved in a game setting by engaging players to explore and experiment with the underlying network MR dataset. I believed that creating a play experience that structures players' engagement with the underlying data, and motivates them to experiment, would help achieve this. This was in line with research indicating game features serve an important motivational role by encouraging players to engage in these structured ways for longer. In

educational games, the actions players make in-game are often deliberately structured to engage them with the learning materials (Aleven et al., 2010). Similarly, gameplay is often used to engage participants in data collection tasks as part of engaging research games (Schrier, 2016). For example, "EvE Project Discovery" embedded an image recognition task into the multiplayer game EvE Online and encouraged players to engage by embedding rewards, scoring and leader board systems (Sullivan et al., 2018). Therefore, my intended play experience would engage players to explore my dataset.

Target audience

I then sought to define and understand my target audience. I originally intended to develop my data game for MR researchers, but, because of logistical issues, this changed part way through development to focus on undergraduate science students. I collected information in two stages, first by interviewing three MR researchers, and second by discussing with twelve student and researcher attendees during a poster session at the 2019 MR Conference (https://www.mendelianrandomization.org.uk/). These interviews were helpful in understanding the requirements that MR students and researchers would have for a data game, though I subsequently widened my intended audience to undergraduate science students for the purpose of enrolling a larger number of participants in the subsequent study. In this section I will detail what I learned, and where I could have spent longer conducting this user research.

I conducted semi-structured interviews with three MR researchers to understand the challenges they face in MR research, and discuss how an MR data game might help explain concepts in MR. From this I identified that a data game could engage lay audiences with complex issues in MR (Appendix 5.2). One statistician argued that a key challenge in MR is understanding and communicating complex methodological issues which affect researchers' ability to obtain valid causal effect estimates. Furthermore, although some problems in MR are complex, they can be broken down into simple representations, and that gameplay could be used to engage people in

ways that does not require statistical knowledge. From this I took away that it would be helpful if my game represented how network complexity affects our interpretations of causal effects in public health. A second researcher who uses MR added to this, arguing that a game could be used to engage non-expert audiences in MR research by giving them opportunities to solve problems in a game that does require expert knowledge. Particularly, they noted that we each have different knowledge bases and this can be helpful in identifying novel solutions to problems. I identified that a non-expert audience could bring individual perspectives and knowledge to offer unique solutions to problems. The third researcher corroborated the engaging nature of games and detailed how they had used gameplay to engage used members of the public with complex issues in MR. Overall I took from these interviews that a game could engage non-expert audiences in the complex topic of MR.

This change of focus fitted well with my existing expertise in presenting public health information to lay audiences in accessible formats. However, it would also have been interesting to fully explore the data collection aspect of these discussions, and produce a data game that could contribute to ongoing MR research.

I presented a poster at a specialist MR conference and held discussions with attendees. My aim was to understand what this audience would expect from an MR data game, as this would help me design an agreeable game that they were more likely to be engaged by (Appendix 5.2). Attendees remarked that they would expect a data game based on MR to present information in a playful way while being accurate and relevant to real-life problems. This fit previous research, for example, players of an open government data game expected data to be accurately represented in the game, and that they would be able to explore and experiment in a safe and playful setting (Simonofski et al., 2022). These discussions were useful in that they confirmed that a data game approach was valid, and that it was a key priority that data be represented accurately.

By the end of this stage, I had planned my data game and collected information on my target audience. My plan was to develop a game that engaged players in the network complexity of public

health in a way that is accurate to an underlying network MR dataset. These values would help guide my decisions during the next stage in designing the game.

5.3.2 Design stage

The design stage is a process of rapid iterations intended to specify how the desired play experience, motivation and structure, might be achieved through the game mechanics (Salen & Zimmerman, 2003). In this section I will explain the playtesting method and how I refined three game prototypes into a final game design.

Playtesting is a key method for gathering player data. A prototype is circulated to a set of playtesters and feedback is collected and reviewed in a cycle until the game achieves the desired player experience (Fullerton, 2018). The person based approach (Yardley et al., 2015) applies some structure for this feedback process by integrating a classic "MOSCOW" prioritisation system (Barker & Clegg, 1994). A "table of changes" is a key method used to record and prioritise feedback, as well as evaluating the effectiveness of changes.

Designing prototype games

I developed three prototype games for playtesting with the view to combine the good aspects of each into a final design (Salen & Zimmerman, 2003). These were constructed from paper (Appendix 5.1) so that I could produce new versions and conduct iterations of playtesting more quickly (Fullerton, 2018). I will outline each prototype below:

1. **Epidemic simulation game**: The player takes the role of a malevolent actor who is attempting to cause as much ill health in a population as possible. Players win by identifying a set of health variables which, once intervened on, would sufficiently reduce health in line with a set goal (e.g., increase loneliness by 10%). This pattern of reverse-psychological play is used in games like Plague Inc, setting a morbid but unique premise and way of engaging with the data. By identifying the variables which most detriment health, players also identify the

variables which would most benefit health. This setting would be unethical in real life, but a game world allows players to explore scenarios safely and this can be an interesting yet safe way of interacting with game worlds.

2. **Narrative adventure game:** Players take the role of a tribe leader, responsible for the health and wellbeing of a group of settlers in a mysterious forest. The player wins by venturing into the forest and collecting magic items which most benefit the tribe, balancing effects on good health with good wellbeing. Each item would have a different effect based on the network dataset, often causing opposing effects which either detriment or benefit health but have the inverse effect on wellbeing. The act of balancing which items the player selects to take back to the tribe would convey the disparate effects of public health interventions.
3. **Fantasy MR game**: This is a multiplayer game where players compete to suggest the best solutions to problems in public health. A round will start by giving all players the same goal, to reduce the prevalence of a disease as much as possible, such as reducing heart disease. Players will then collect three cards representing health variables they can intervene on to achieve the goal, and select the one they think will work best. At the end of the round players are awarded points corresponding to how well their solution would work according to my network dataset. Once a card was played it was removed from the game, so players would have to play cards only at the most appropriate opportunity. The game ends after five rounds and the player holding the greatest number of total points wins. This pattern of play is similar to fantasy-football, where players collect cards representing footballers and score the most points by playing them at opportune moments, when they expect the footballer to do well. A competitive atmosphere encouraged players to find the best solutions, and added an enjoyable social dynamic. Since players could not replay cards and rely on a single reliable intervention that generally has good effects, such as reducing body mass index, players instead were prompted to explore all their options and find new solutions to problems.

Playtesting 1. Paper prototypes

I conducted playtesting with eight playtesters in a mixture of one-on-one and group sessions. My playtesters were selected to be representative of my intended audience, studying science degrees. In one-on-one sessions I prompted them to "think aloud" about the decisions they were making so that I could better understand their aims and perspectives (Jaspers et al., 2004). In group sessions, played games with multiple players concurrently and observed what worked and what did not. In total, I recorded 87 items of feedback in tables of changes and implemented 48 changes which I confirmed and followed up with subsequent playtesting (for tables of changes see Appendix 5.2). An example table of changes is presented in Table 5.2.

Table 5.2 Example table of changes

Date	Issue	Priority	Solution	Did it work?
1st Nov	*"The game ends at level 6, so is quite short"*	1 (High)	Increase the time to complete a level	Yes
1st Nov	*"Understanding what factors are in the network requires a lot of reading"*	2 (Medium)	Represent variables with quick-to-identify icons	No (player still reports issue)
10th Nov	*"It is not always clear what the icons in the game represent"*	2 (Medium)	Add a help section detailing what icons represent	Yes

Note: For criteria by which priority codes were assigned see tables of changes in Appendix 5.2.

Designing the final game

Following playtesting, I described how each prototype game played in treatment documents. Treatment documents formalise the features present in each game (Fullerton, 2018) and allowed me to identify specific game features that worked, and that I would like to take forward into the final game (Sasupilli et al., 2019). Part of documenting the final game was developing "gameplay loops" which describe the actions players take over the course of gameplay (Salen & Zimmerman, 2003). In this section I will describe the game features that would engage players with my final game as part of two "micro" and "macro" gameplay loops.

The “micro” gameplay loop gives players a series of core actions that would structure how they engage with the underlying dataset. Players would be set a goal to achieve, such as improving wellbeing, and be prompted to search the dataset for a solution, such as reducing insomnia. After selecting a solution, they would be scored based on their performance. This feature was effective in my epidemic game prototype which framed data as containing the solutions to public health problems.

The “macro” gameplay loop would encourage players to continue engaging by providing them an AI opponent to compete with. I included competition because playtesters enjoyed this element of the fantasy MR game prototype. I also progressed players to new levels once they collected enough points and this helped contextualise their in-game actions as part of a wider scenario, like in the narrative game prototype.

These micro and macro gameplay loops would serve as a framework to structure the final stage of game development. I will later describe the precise actions players take in the final game (in Results), but this stage is important as it details the precise ingredients, game features, that would be in the final game. These are presented in Figure 5.7.

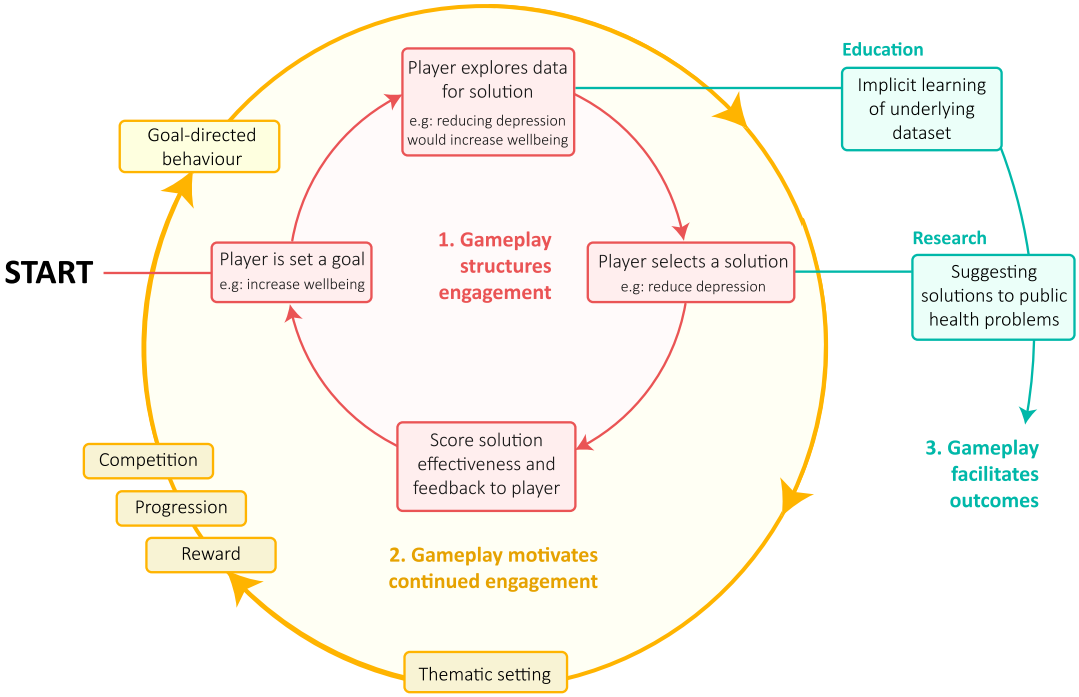


Figure 5.7 A gameplay flowchart (Yusoff et al., 2009) presenting the game features present in my game. Players are motivated (yellow) to engage with a structured play experience (red) involving activities which facilitate education and research outcomes (blue).

5.3.3 Development stage

The aim of the final, development, stage was to implement my game design and achieve my intended player experience. In this section, I will detail two further playtesting sessions that helped me achieve the intended player experience, engaging players with the underlying network dataset.

I developed a prototype computer game that put an interactive user interface to the game design outlined previously (Figure 5.8). Players were given a stated goal, and the ability to enact interventions by clicking on a variable to bring up a menu with details and intervention options.

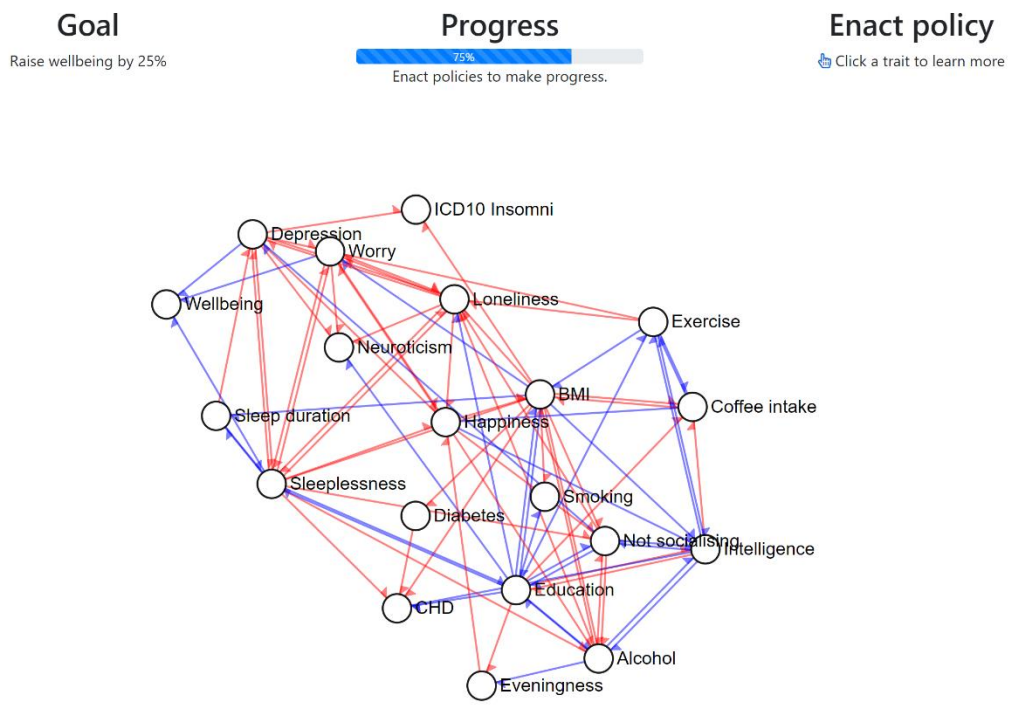


Figure 5.8 A digital prototype of my game featured an interactive visualisation and expandable menus that would appear on the right of the screen when players clicked on variables.

Playtesting 2. Initial digital prototype

Playtesting revealed several issues with the initial digital prototype. I conducted a focus-group-style playtesting session with five researchers from the MRC IEU. I prompted players to think aloud and provide commentary on their intentions and play experience, and to discuss their views on the game with each other (Appendix 5.2).

Players found the game confusing and difficult to play, primarily because they felt the visualisation did not make it clear what the effects of interventions were. They also reported it was not particularly fun, but they provided useful feedback on how this could be improved. Simpler gameplay that was better explained and communicated would be more enjoyable, such as removing the menu system which players felt hid information from them even though it was unclear how it related to their goal in the game. In total, players offered 26 points of feedback, of which 10 received agreement from multiple playtesters.

In response to this feedback, I made two major groups of changes: the first would mitigate concerns that the user interface was overly complicated and involved many pop-up menus, and the second would improve the game features to better motivate play. For example, I implemented an unlock system which allowed players to spend points to use new abilities, such as intervening on two variables simultaneously. At this stage I also developed a space theme for the game with supporting narrative, art, music and sound effects. This decision had implications that I will discuss later.

Playtesting 3. Final game

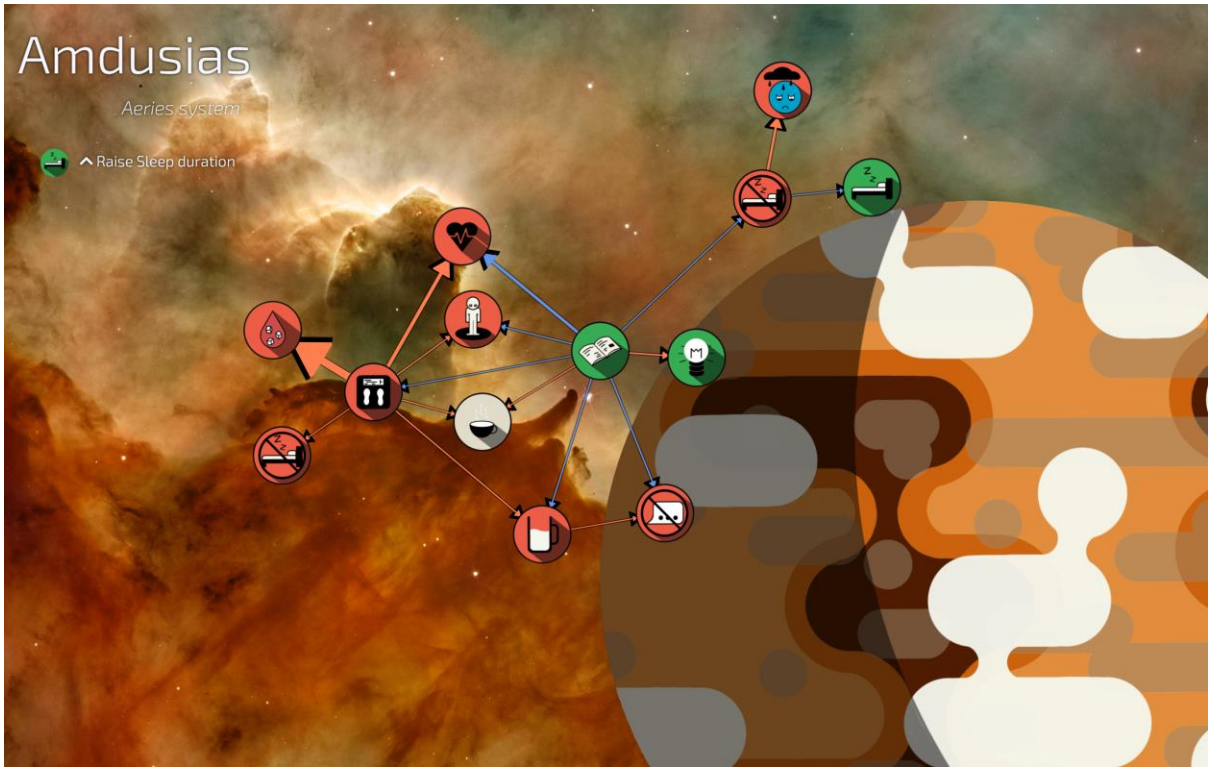
I conducted a final round of playtesting that showed that players responded positively to the improvements I made. During 2020, COVID-19 social distancing had been enforced so I switched to online playtesting, and I will expand on the impact this had in the discussion. I invited 14 playtesters to play the game either offline in their own time (n=10) or over a live video call (n=4). Playtesters were again selected to represent my intended audience and included individuals previously or currently studying science degrees (n=12) or degrees in other areas (n=2).

Players suggested 44 points of feedback, of which five were reported by multiple participants. In general, players felt that the game was challenging, encouraged them to explore the data, and they felt engaged in a competition to find a better solution than the computer did. Additionally, players felt that they were able to learn game mechanics after a period of trial and error but that a tutorial would make things clear from the beginning. Furthermore, players reported several bugs and errors which frustrated their ability to enjoy the game, resulted in unfair scores and impossible levels in which an AI opponent could not be beaten. I recorded feedback and made 35 changes (Appendix 6.2). Taken together, this indicated to me that the game was achieving an engaging effect on players, and that feedback was starting to focus on technical issues than game design. In the next section I will seek to understand whether the game was achieving the intended play experience.

5.4 Results

The final version of the data game situates players in a space-themed world acting as travelling epidemiologists who solve public health problems. It is based on an underlying public health simulation that models how changes to variables in my network MR dataset affect the rest of the network. I transformed this model into a gameplay loop where players are set a goal, search my dataset for a solution, and are scored on their solution according to the effectiveness in my network model. Players win the game by identifying the most effective solutions to each goal in the game. As players suggest effective solutions, they are awarded points. Once players accumulate enough points, they advance up a level in the game. Upon suggesting enough effective solutions to accumulate sufficient points, players progress past the sixth level where they are presented with a win screen. Players will engage with this game to explore and experiment with the consequences of public health interventions, and motivational features such as rewards and progression will encourage players to engage for longer.

The game is presented in Figure 5.9 and is available to play at https://www.morenostok.io/mendel/game.html. The game code is published under a GPL-3.0 open source licence and is available on GitHub (https://github.com/CMorenoStokoe/mr-game-webapp).



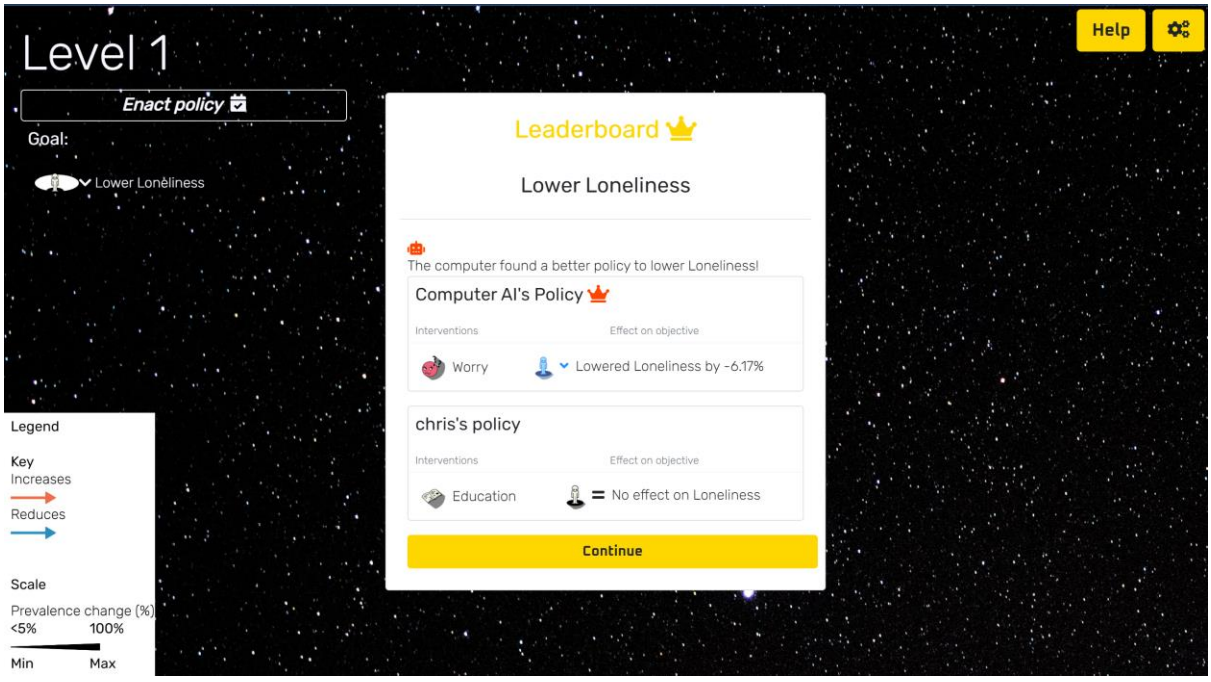


Figure 5.9 The finished game is space-themed and encourages players to engage with an underlying causal network dataset

5.4.1 Play experience

I collected some data describing the play experience. Prior to publishing the final game, I gathered evidence on whether playtesters were engaged to play longer with the game than the non-game

software. Subsequently, I collected evidence from a final study of my game that described the player experience. In this section, I will outline my findings before moving on to a general discussion.

I conducted a pilot testing exercise to test my assertion that the game achieved the intended play experience. Ten of my previous playtesters were asked to help me make final adjustment to make sure that the game was technically ready for launch, but while they did this I timed their interactions to see whether they naturally spent longer on the game than the non-game. Players were not aware they would be timed, and indeed spent substantially longer on the game (mean= 30m) than the non-game simulation (mean= 10m). I will present more detailed statistics of play durations in the following chapter, but I took this as supporting evidence that the game features were motivating.

After launching the game, I gathered free-form responses describing the player experience. In the next chapter I will describe an experimental study comparing the play experience and outcomes for the game to an interactive visualisation control. However, I will present one aspect of this study here instead. I collected free text responses from 90 players who described their experience with the game (for data see Appendix 5.4), and performed a brief thematic analysis to understand the types of themes that could categorise players' experiences (for a figure presenting themes and counts see Appendix 5.4)(Braun & Clarke, 2012). It is likely that a different reviewer would have categorised players feedback with different themes, and this could have been improved by inviting an independent reviewer, but I will share my interpretations below:

First, some players found the data model complicated at first but learned over time. Others were confused by what to do in the game. This indicates the source data, simulation model, gameplay, or all were complicated:

"At first I was quite confused as to what I was doing but after level 2 I understood more…" – Player 72

"I do not quite understand how the percentage change [intervention calculation] is worked out" – Player 2

"I had not a clue what was going on I am not going to lie to you" – Player 11

"I *could find more memorable the effects that I was already aware of like education reduces smoking" – Player 1*

Second, players particularly enjoyed competing with a computer AI rival to find the best solutions:

"…I felt a sense of achievement when I did beat the computer and this did motivate me to do better on the next task." – Player 70

"… the AI would use [interventions] in a different way to me that would bring a greater effect so I tried to follow its strategy" – Player 7

"I especially liked the interaction with the computer that I was competing with something to make a positive effect" – Player 18

Third, players enjoyed the aesthetics of the game in graphic and sound design, with some reporting feelings of immersion, and others that it was relaxing. However, some reported it had negative aspects, such as distracting them:

"… I thought the presentation of the game was quite nice, visual design was pleasing, there was humour in the instructions and the music was calming." – Player 26

"It was a somewhat immersive experience that distracted me for a while." – Player 5

"The soundtrack to the game is very well done." – Player 4

"The space background was both entertaining and distracting/didn't fit the theme of the task" - Player 71

In summary, these findings supported my inferences from playtesting that, although complicated, the game achieved the desired play experience. Though these data collection exercises were brief, pilot testing and participant feedback in a subsequent study indicate my game motivated players to engage with the underlying network dataset.

5.5 Discussion

In this chapter I developed a simulation based on my network MR dataset, and then developed it into a data game that engages players with the underlying data. In the next chapter I will start assessing whether my game achieves its outcomes, for learning and research, but here I will evaluate

the design of my game, note any decisions I might make differently in the future, and highlight the implications and opportunities for future research.

5.5.1 Evaluating game design

Examining the design of games helps build a mechanistic understanding of how games work and how best to design them (Roungas & Dalpiaz, 2016). In this section I will examine some of the key decisions I made and relate them to existing theory, research and player feedback using a structure offered by the "Serious Game Design Assessment framework" (Mitgutsch & Alvarado, 2012). I will review my decisions in the context of theory and evidence with respect to designing gameplay and representing information in ways that help achieve the intended outcome of the game, as well as ensuring a desired player experience is met, and setting an appropriate thematic context for the game.

The first aspect of game design is to consider how the game was designed to achieve its intended outcomes (Mitgutsch & Alvarado, 2012). I defined these early on in the design process so that they would guide development, and this helped me keep the scope of my work focussed, limiting the number of features I added and avoiding "feature creep" (Elliott, 2007), ensuring my game did not become unnecessarily complex (Kanode & Haddad, 2009). I also designed gameplay loops following best practice for how to achieve outcomes for learning (Schrier, 2016), such as identifying an intended learning outcome (Aleven et al., 2010), incorporating learning as a core gameplay action (Winn, 2009), and designing simple gameplay to empower players to contribute to complex topics in research (Cannon et al., 2010). Therefore, I developed targeted gameplay that was designed to achieve my outcomes.

Second, the information presented in a serious game should be valid for achieving the outcome (Mitgutsch & Alvarado, 2012). I identified early on that my playerbase would expect an accurate dataset from which lessons could be relevant for real-world problems in public health. It is well understood that the educational information presented to players must be accurate (de Freitas &

Jarvis, 2006) and I established in a previous chapter (3) that my network dataset is a valid representation of the type of network complexity in public health. Players did note that some relationships were not intuitive to them, for example the lack of a relationship from exercise to body mass index. This is an artefact of the use of an arbitrary value for selecting significant relationships for inclusion in the network since, for example, an effect of exercise on body mass index was found (P=4.54x10^{-03}), it just did not reach network wide-significance (P<2.15x10^{-04}). Arguably, this is a true reflection of my dataset though it illustrates how decisions in data selection can affect the player experience. Two aspects of data transformation could have been improved. First, I removed loops from my network dataset, and as a result my simulation did not represent this component of my network dataset. Presenting a sanitised view of the data without cyclical effects gives a falsely simplified interpretation and precludes the possibility of discussions around why I found them in the first place. Second, interventions were also simplified in that players were not given control over the magnitude of their effects; interventions were always a fixed value and strictly beneficial. While my streamlined controls were more intuitive for players, they removed the possibility of discovering dilemmas where sacrificing one unhealthy variable, such as increasing alcohol consumption, could improve a healthy variable, such as improving socialisation. Therefore, the network dataset was integrated in a way that players reported allowed them to learn about realistic relationships in public health, including relationships they already knew about, though it could have been represented with even more nuance.

Third, it must be clear what the game features are and how they achieve the play experience (Mitgutsch & Alvarado, 2012). My gameplay loops are designed to direct players to search, explore and experiment with my network dataset. Players are motivated to continue engaging through several features which have been demonstrated to improve motivation (Sailer & Homner, 2020). For example, setting goals focusses attention, encourages resilience despite failures, and rewards success (Tondello et al., 2018), while the scoring meets players' psychological needs for feeling competence (Sailer et al., 2017), and competition introduces a social challenge that inspires players

to achieve higher scores (Preist et al., 2014). This is supported by a meta-analysis (Park & Kim, 2018) which found that levels, leader boards, points and competition were strongly associated with players experiencing feelings of challenge (r=.93) and competition (r=.92), while goals and progression contribute to feelings of discovery (r=.93). Furthermore, free-form player feedback indicates players were engaged in the challenge and competition to beat their computer opponent. While my game was single-player with competition against the AI for practical reasons, it is worth considering that the addition of multiplayer features appeal to our need to feel "connectedness" (Marczewski, 2018) and can be particularly engaging (Francisco-Aparicio et al., 2013). Therefore, my gameplay was designed to motivate players to continue engaging with the underlying dataset by appealing to their need to feel competence, though future research should explore and exploit social mechanisms that appeal to the need for connectedness.

Fourth, the context for play should be appropriate for the intended outcomes (Mitgutsch & Alvarado, 2012). The theme of a game plays an important role in setting the context for play, and this includes aesthetics, such as visuals which are aesthetically pleasing stimuli for players (Sardi et al., 2017), and narratives that contextualise players' in-game actions (Kapp, 2012; Winskell et al., 2019). Another aspect is how players interact within the game world. I took advantage of the magic circle by situating players in an other-worldly space setting, where players were empowered to make interventions in a safe fictional scenario. This is important because it overcomes the normal hurdles where players might not feel experienced, competent or qualified enough to contribute to a real-world scenario (Schrier, 2016). Feedback from players demonstrated that they enjoyed this theme, and it appears to have been successful in this respect. However, it also appears to have distracted some players who felt it did not meaningfully tie in with the task. Fantasy settings have been featured in serious games before, but they fit the task, such as a wild-west shoot-out themed reaction (Lumsden et al., 2017). When I considered the issue of theming again in a recent project, I decided instead on a thematic setting more closely related to the real-world application, contextualising COVID-19 transmission models as data from a fictional country, and this helped

achieve both the goal of empowering players in a safe setting while making narrative sense (Moreno-Stokoe et al., 2021).

In summary, my game was designed in four key ways that help ensure it would fulfil learning and research outcomes while achieving a playful experience. The actions players complete in the game direct their interactions with an underlying dataset (1), which was represented accurately in-game (2), and the gameplay engaged players (3) and empowered them to explore (4) the data. These are the strengths of my game, and in the next section I will discuss some of the limitations of my game design process.

5.5.2 Limitations

My design and development processes were limited in some respects. In this section I will outline challenges to my playtesting and user research processes, and how they could have benefited from extensions or revised strategies.

My playtesting processes were limited by the COVID-19 pandemic and the difficulties in prototyping data-based games. COVID-19 social distancing measures imposed restrictions on my playtesting process that resulted in some clear limitations. The timing of the measures in 2020 coincided with my playtesting stages and limited the amount of playtesting I could conduct, as well as reducing the quality of insights I could obtain. My final playtesting session was conducted online, using video calls, so I could not follow players actions and understand their intents and motivations as closely as I could in-person during the earlier stages of playtesting. In the commercial games industry many games were delayed, for example CD Project Red's "Cyberpunk 2077" was delayed due to the transition to remote working and illness (CD Projekt, 2020). There were also indirect consequences that affected researchers since many of my playtesters had less time to give because of disruption or new demands on their own research. This issue was compounded by not having integrated the data model until digital prototyping. This meant I did not spend long playtesting game designs using my data, and consequently the least refined part of the game was how the data model is integrated.

This is a broader issue with data games because while it is possible to rapidly iterate with paper prototypes, it is difficult to incorporate real data without investing some time in programming, which is much slower. Future research should find a way to stay in the paper prototype stage for longer, producing more iterations before selecting a final game design, perhaps by supplementing the need for software models with a simple companion app, like some board games which use them to perform tasks which would be tedious or difficult if conducted manually. For example, "Forgotten Waters" uses a companion phone app to deliver personalised narratives for players based on their in-game decisions (https://order.fwcrossroads.com/). Ultimately, I produced a game which met my play experience goals, but I could have refined it further with additional and higher-quality playtesting.

The different aspects of my user research were not entirely aligned with a common goal and this resulted in insights that could have been more focussed. The research in this chapter helped me formulate a plan for my game, identify key values, formalise my aims, and guide decisions accordingly during game development to produce an effective game. In this respect, my research process was effective since I developed a game that motivated players to engage with an underlying dataset. However, I could have improved my interview process. My interviews with MR researchers could have been an invaluable opportunity to identify and precisely understand the barriers that prevent them from conducting MR research. My interviews uncovered some of these, particularly the need for accessible mediums to convey complex topics in MR, though I muddied my investigation by also trying to elicit player requirements. Player requirements are helpful for understanding how best to design a game for a given purpose, but at that stage of planning I did not have a firm aim for my data game, and probing researchers for barriers further would have helped me identify a more specific issue to address with my data game. Furthermore, requirement insights gathered for researchers became less useful as my target audience shifted to students. As a result, my game was designed to facilitate a conceptual data collection exercise and in the next chapter I will explore the idea that while game features show some promise facilitating data collection, the

exercise itself is not rooted in a real MR problem. In the discussion chapter I will continue exploring this as a direction for future studies, and suggest how a different data game design could aid ongoing MR research.

5.5.3 Future directions

The field of games research requires new frameworks that characterise the game design process as the starting point for experimental research. The key component of experimental studies lies in operationalising the comparison, identifying and delineating the experimental ingredients of the game and comparing them to a control without these ingredients. A similar component of game design frameworks is selecting and forming a good understanding of the game features present in a game, theorising how they might contribute to a play experience, and ultimately achieve an outcome. In this chapter I combined these two types of framework with a focus on designing an evidence-based and considered gameplay loop, but future research should evaluate my approach and investigate whether there is a better way to synthesise scientific research frameworks with game design frameworks.

The field of MR data games could be advanced by resolving some key data science challenges. Network MR effect estimates are not designed for use in simulations that propagate effects through a network, so some issues arise when simulating data in this way. One is that effect estimates are expressed in different units and these must be understood in order to accurately interpret the magnitude of changes in a simulation. This is problematic because units are not routinely recorded in GWAS summary data sets, so researchers need to locate and search through the extensive documentation for each contributing GWAS to find the units, and be sure that the original units were not transformed in any way as part of the GWAS summary pipeline. Another issue is that assigning points and scores based on simulated interventions is difficult, because effects need to be evaluated as good or bad in some way, even though there is no clear consensus for whether some variables are healthy or unhealthy, such as caffeine (M. Cornelis & Munafo, 2018; Snel & Lorist,

2011). Similarly, some variables such as body mass index are unhealthy both in high and low levels and this further complicates interpretation. While these issues are time-consuming to overcome manually at a small scale, solutions need to be developed to systematically process large-scale data for use in data games. Solving these types of problems is the focus of data games research. For example, the key challenge in designing Open Monopoly was developing a systematic process for procedurally generating a monopoly board based on open data (Friberger & Togelius, 2012). Systematic processes generally need to be refined, as was the case generating realistic maps based on open data for the game Civilisation (Barros & Togelius, 2015), and so developing systematic processes will be one challenge and refining them will be another. Therefore, researchers should develop and refine methods for procedurally generating game content so that health data games can be made that explore the challenges in representing open health data.

5.5.4 Conclusion

In this chapter I built on my previous chapters to develop a public health intervention simulation with an underlying network MR model. I then transformed it into a game over the course of three stages, planning the concept and theory for the game, designing gameplay loops, as well as developing and playtesting an implementation as an online computer game. I evaluated the design of this game and concluded that it achieves a play experience that structures and motivates players to engage with the underlying network MR simulation, and that there is evidence and theory to support this. Several factors and design decisions did, however, limit the scope and quality of my playtesting and user research processes, so future researchers should seek to expand on these aspects. Future data games should continue to explore the challenges around transforming and analysing health data as gameplay components. In the next chapter I will conduct an experimental study of this game and assess whether it achieves its intended outcomes for education and research.

6 Evaluating a network data game using an experimental control

6.1 Introduction

In the previous chapters I introduced serious games (1) and obtained a network dataset describing MR estimates of causal influences among health measures that demonstrated one type of complexity in public health (2, 3). I then developed the visualisation, simulation and game software that would allow me to explore a new approach to understanding network complexity through gameplay (4, 5). In the present chapter I will round out these investigations with a formal experimental comparison of the data game to the non-game simulation control. This chapter will evaluate whether the gameplay I implemented engages participants to learn a network dataset and use this information to suggest valid solutions to complex problems in public health. This chapter will also continue to highlight challenges in conducting experimental games research and this will become a key theme for discussion in the final chapter (7).

6.1.1 Education and research outcomes

In the introduction chapter I outlined the state of the serious games literature and noted that previous research has made several claims about the effects of game features. Games are argued to help students achieve better learning outcomes (Johnson et al., 2016; Ricciardi et al., 2014; Sardi et al., 2017) and to contribute more data to research projects (Peplow, 2016; Schrier, 2016). However, the games literature presents a mixed view on whether game features provide advantages over traditional non-game materials. In this introduction, I will offer a critical analysis of this literature, highlight its findings, and draw out some opportunities to build on this evidence. I will focus on two types of games; first, games for education and then games for research outcomes.

Educational games

Educational games are designed to deliver educational materials where the addition of game features is claimed to improve engagement. Recall that in the introduction chapter (1) I explained that one of the main arguments for game features is that they may improve motivation and result in better engagement. I expanded on this in the previous chapter (5), explaining that game features are theorised to appeal to intrinsic psychological motivations. Motivation is hard to observe directly so it is typically indexed with objective measures of play duration, or subjective measures of players' experiences (Hamari et al., 2014). In this section I will review the evidence that game features can improve engagement with education and learning. Two reviews of educational games research, one conducted before 2014 (Ricciardi et al., 2014) and the other after 2014 (Sailer & Homner, 2020) present similar views that game features can improve learning outcomes, though the methods used by games researchers varies from study-to-study, as do the reported effects of game features.

A scoping review (Ricciardi et al., 2014) details the findings of 13 studies evaluating the outcomes of medical education games designed to teach healthcare professionals. Eleven studies (84%) report positive learning outcomes in a range of studies with different designs. At face value, these studies appear to support claims that game features can improve educational outcomes. However, few of these studies used an experimental design, and those that did often compared conditions where the effect of game features was not clear. When taking these factors into account, the studies with strong designs present more mixed evidence for an educational effect. I will expand on these two issues in turn and explain how they reduce strength of evidence.

First, a lack of experimental comparisons prevents games researchers from obtaining strong evidence that game features causally improve educational outcomes. I explained in the introduction chapter that causal research designs involve observing the effects of an exposure as closely as possible, and experimental comparisons allow researchers to do this by comparing outcomes from a game to a non-game control. Despite the strength of experimental designs, relatively few games

researchers use them. In the previous review of medical education games, 7 of 13 studies (54%) used non-experimental designs and were not able to determine whether other factors contributed to differences in learning outcomes. For example, some studies compared medical students' test scores before and after playing an educational game but did not account for the effect of baseline studying in preparation for the test. The proportion of studies using non-experimental designs is consistent with reviews investigating other types of serious games (53-61%), indicating a problem lies with practices in the wider field of games research (M. Brown et al., 2016; Johnson et al., 2016; Kara, 2021). If we focus on just the 6 studies that used experimental designs then the proportion of studies reporting positive effects falls from 84% to 66% (Andreatta et al., 2010; Buttussi et al., 2013; Knight et al. in Diehl et al., 2013; Gentry et al., 2019; Jerin et al. in Mancini et al., 2010; Qin et al., 2010).

Second, some studies are designed in a way that makes it unclear what factors are responsible for differences across game and control conditions. It can be difficult to define and manipulate game features due to individual differences in the nature of play experiences (Salen & Zimmerman, 2003) and disagreement on what game features are (Sailer & Homner, 2020). In games research, often control conditions do not adequately control for confounding factors that could explain the differences between game and non-game conditions. In the previous review, 2 of the 6 experimental studies used inadequate control conditions which did not control for confounding factors that could explain differences between groups (Buttussi et al., 2013; Qin et al., 2010). For example, one study compared outcomes from students provided additional gamified learning materials to a control condition where participants did not receive any learning materials (Qin et al., 2010). If we focus on the 4 studies which used well-designed experimental studies, then the proportion of studies reporting positive effects shrinks further from 66% to 50% (Knight et al. in Diehl et al., 2013; Gentry et al., 2019; Andreatta et al., 2010; Jerin et al. in Mancini et al., 2010). Therefore, the highest quality evidence available indicates mixed results on the effectiveness of designing educational games for medical students.

Another review was conducted into 10 experimental studies of educational games used in university courses (Sailer & Homner, 2020). Their findings agree with the previous review, with weaker experimental approaches (or lack of them) inflating the observed effects of game features. 6 studies (60%) found that game features improved learning outcomes compared to a non-game control. Standardising effect sizes as standard deviations allowed reviewers to compare effect sizes across studies (Cohen's D). On average, gameplay resulted in a substantial improvement of 0.25 standard deviations in learning outcomes (d=0.25, min=- −0.42, max=0.76). Heterogeneity testing was then used to investigate whether studies with different designs reported different effect sizes (Cochrane's Q). Studies that used weaker designs reported substantially larger effects of game features. For example, studies without randomisation reported effects that were four times larger (d_{with}=0.13, n_{with}=6; $d_{without}$=0.51, $n_{without}$=4; Q[1] = 4.67, p<.05), and those without adequate control conditions reported effects two times larger (d_{with}=0.47, n_{with}=8; $d_{without}$=0.81, $n_{without}$=2; Q[1] = 40.92, p<.01). Design elements such as randomisation and adequate controls are effective because they reduce the opportunity for confounding factors to influence the results. Therefore, these findings provide further evidence that the advantages for game features in education may be inflated by weak study designs.

In summary, two reviews of educational games indicate that gameplay can enhance learning outcomes, but the strength of evidence is hampered by less robust study designs. The large differences in findings among the studies in these reviews highlights the need for well-constructed experimental designs. In particular, experimental designs should compare a game that delivers a strong dose of game features, with a control condition that controls for confounding factors, while randomly assigning participants to conditions.

Research games

Research games are intended to facilitate the collection, processing, and interpretation of scientific data. It is argued that the same motivational effects that are theorised to improve engagement with

education can improve participants engagement with research tasks. Less research has been conducted on this topic, but two reviews demonstrate that players can contribute useful data in research games (Schrier, 2016) and that game features can motivate participants (Looyestyn et al., 2017).

A review of serious games (Schrier, 2016) catalogues 8 research games that have successfully obtained large quantities of valuable data from players. Research outcomes are detailed in 15 peer-reviewed academic publications. Players have used image-recognition skills to identify diseased cells (Kwak et al., 2013; Mavandadi et al, 2012a, 2012b, 2012c; Ozcan, 2014; Singh, 2017) and sub-cellular structures like mitochondria (Peplow, 2016; Sullivan et al., 2018). Players have also used pattern-matching skills to identify protein combinations (Eiben et al., 2012), speculate about their functions in diseases (Leppek et al., 2022; Rangan et al., 2021; Wayment-Steele et al., 2021), and align data in DNA sequencing efforts (Kawrykow et al., 2012; Kwak et al., 2013; Singh, 2017). Aside from acknowledging the game-based approach to data collection, few articles comment on the effect of adding game features to the data collection process. One Cancer Research UK study claims that game features reduced the accuracy of players' contributions, and advocate for delineating game and research components (Cancer Research UK, 2015; Coburn, 2014). Conversely, one study investigating game-based identification of malaria cells argued that, although accuracy may at first be poor, players can be trained to give data with accuracies as high as 90%, the same accuracy a subject matter expert could achieve (Mavandadi et al., 2012).

Having established that games can be used to collect accurate and useful data, I will now turn to investigating the specific effects adding game features to the research process may have. There has been no research investigating research games for health data, but previous experimental studies have compared the outcomes of game and non-game research tasks. A review of 4 experimental studies suggests game features can enhance data collection activities by motivating participants (Looyestyn et al., 2017). A sample-weighted meta-analysis in investigated whether participants were

willing to make more contributions to, or spend longer completing, gamified versions of survey and image recognition tasks (Cechanowicz et al., 2013; Guin et al., 2012; Harms et al., 2015; Mekler et al., 2013). Three of four studies (75%) reported a large effect in increasing the number of contributions participants made (d=0.51, $n_{participants}$=2724, $n_{studies}$=4 min=0.08, max=0.80). However, no effect was found to suggest participants would spend longer on the activity, in fact they spent less time (d=-0.30, $n_{participants}$=783, $n_{studies}$=2, min=-0.04, max=-0.36). For example, one study reported that gamifying a market research survey elicited more responses from participants but reduced the time they spent completing it overall (Cechanowicz et al., 2013). These findings are at odds with findings from another series of studies investigating adding game features to cognitive testing materials. One study found that game features encouraged participants to participate for longer (Lumsden et al., 2016a), and increased their likelihood to keep participating across multi-part studies (Lumsden et al., 2017). Taken together, these findings indicate that games can increase participant engagement with research, though there is mixed evidence what the material result for this would be, either the number of contributions or duration of participation.

In summary, reviews of research games highlight many cases where players have contributed to important research activities. There is also some evidence that game features have the potential to motivate participants to engage, and contribute more data, though further research is required to determine whether participants also engage longer periods of time.

6.1.2 Present study

Experimental research investigating the effects of game features suggests they could help improve outcomes for education and research but methodological issues prevented researchers from making strong conclusions. Key issues included weak manipulations of gameplay, constructing control conditions that did not control for confounding factors, and not randomly allocating participants to conditions. There is also generally little research in some areas, such as the outcomes for research.

These issues currently prevent us from having a complete understanding of the effects of game features, and reduce the confidence we have in present findings.

In this study, I will test the effects of game features by constructing an experimental manipulation of game features that addresses some methodological issues. I will compare outcomes from the game and non-game simulation software developed in the previous chapter (5). Both incorporate my network MR dataset (2, 3) and share the same interactive visualisation at their core (4), so the non-game simulation acts as an appropriate control condition that varies from the game only in the absence of game features. Participants were randomly assigned to participate in either the experimental game condition or non-game control condition. Outcomes related to motivation, education and research were compared across conditions.

The aim of the present study is to investigate whether the game features in my data game motivate participants, improve learning, and achieve research outcomes. I will make between-subjects comparisons to assess the effects of game features on motivation and learning outcomes, and investigate the contributions of players to a research exercise. I will investigate outcomes including three hypotheses (for motivation and learning) and one exploratory analysis (for research):

- Motivation

  H_1 Players of the game will use the software for a longer duration than the controls

  H_2 Players of the game will report a more playful experience than the controls

- Learning

  H_3 Players of the game will demonstrate a better understanding of the network relationships in my dataset

- Research

  I will investigate the solutions players in the game contribute in an example data collection exercise solving hypothetical problems using my network dataset

Finally, I will measure the usability of game and non-game software to help ensure differences between groups in hypothesis testing are due to the intended manipulation of game features.

6.2 Methods

6.2.1 Design

I constructed an experimental study where participants were randomly assigned to use either game or non-game software. Participants in both conditions used software which was based on a simulation of public health. This allowed participants to make a virtual intervention by selecting a factor in the network to improve (e.g., lowering depression) and view its simulated effects within my network dataset (e.g., raising wellbeing). Participants in the game condition received a version of the software that contained game features, such as a scoring system, and were asked to interact with the simulation by suggesting solutions to solve various simulated problems in the game (detailed in Materials). Participants were compared on various outcome measures designed to assess effects on motivation, learning and data collection (Table 6.1). Duration of use, playfulness and multiple-choice learning scores were measured using a Qualtrics questionnaire, and in-game actions were recorded using custom database software (described in Materials section 6.2.4 below). Because of the way they were recorded, information from in-game actions cannot be joined with other measures. Ethics approval was obtained from the University of Bristol Psychological Science School Research Ethics Committee (ID: 111083).

Table 6.1 Outcome measures taken for participants in the game and control conditions

Measure		Game condition	Control condition
Motivation	• Duration of use	X	X

	• Playfulness	X	X
Learning	• MCQ learning score	X	X
Research	• In-game actions	X	

Note: Measures marked with X indicate that participants did complete this measure and the absence of an X indicates they did not.

6.2.2 Participants

Participants were drawn from a pool of students studying at the University of Bristol School of Psychological Science. Respondents needed to meet the criteria that they had normal or corrected to normal eyesight for inclusion (all respondents did). Participants were reimbursed for their time with course credit (as is standard in the school). Information on age and sex were not collected as part of this study but given the pool of participants (undergraduate psychology students) we can approximate the likely composition of the sample. Of the students enrolled in the BSc Psychology in the academic year 2021-22, 81% identified as female, and 92% were aged between 18-21 (http://www.bristol.ac.uk/ssio/statistics/). Thus, our sample is likely a predominately female young adult sample. I will discuss the limitations of my sampling in the discussion section.

Participants were randomly assigned to either the game condition (n=90) or control condition (n=85). Note that these counts reflect the final total assignment after some participants were excluded, and I will detail this in the following data preparation section.

I conducted a power analysis calculation to ensure that my sample size was adequately powered to detect small but meaningful differences in outcomes (Appendix 6.1). A minimum of 102 total participants is required to detect the minimum meaningful difference in outcomes at 80% power (assuming a two-sample t-test at alpha=.05). The minimum meaningful difference was defined as a mean difference of one point, a correct answer, on the learning assessment. The sample we obtained gave us around 96% power to detect a single-point difference.

6.2.3 Procedure

Participants were recruited from the experimental hours scheme at the University of Bristol school of Psychological Science via an online advert. Participants who responded were given an information sheet describing the background and procedure for the study, and provided informed consent to continue with testing. Participants were free to withdraw from the study at any time and after the study they could withdraw their data for any reason (up to 3 months after testing). Agreement was required for two further statements. A risk assessment was conducted prior to recruitment and although this study presented no additional risk to participants beyond that encountered in ordinary life, a precaution was taken to ensure that participants did not take my putative network data as medical advice accurate in the real-world. Participants were asked to indicate agreement with a statement declaring their understanding of this fact. Furthermore, due to the technical implementation of my software I required that participants understood and agreed to conduct the study on a computer rather than a mobile device, and to use a web browser other than Internet Explorer (which does not comply with all web standards).

Testing was conducted online during December 2020. The study was administered as an online form, hosted on Qualtrics, which detailed instructions, directed participants to study materials, and collected responses. All participants were told that they would be provided with some software to learn the relationships between health factors in a network dataset. Participants in the game condition were then linked to the game, and controls were linked to the non-game. Links opened websites in a new browser tab so that participants could return to the online form and continue the study after they had finished using the software. Instructions in both conditions were to “spend as long as you like with [the software] until you feel like you have a good enough understanding of the relationships between public health traits, and then return to this form for a learning assessment”.

While using the software various metrics were recorded:

- In both the game and non-game software, whenever a user interacted with the simulation to explore the effects of an intervention, details were recorded including a timestamp and the factor that was selected.
- In the game software, additional information was recorded including the goal or problem that participants were attempting to solve, and the score they received for their solution

When participants returned to the Qualtrics form they were asked whether they took a break while completing the study. Next, participants completed the play experience questionnaire and the learning assessment. The learning assessment was delivered as an open-book test so participants were given a visualisation of the network dataset so that they could refer to this while completing the test. This was not interactive, so it did not give them the answers to questions, but instead helped them remember the factors in the network and how they are related.

After completing all sections of the study, participants were debriefed as to what the aim and hypotheses were for the study. Participants were then awarded their course credits and reminded of their right to withdraw their data from the study if they wished.

6.2.4 Materials

I assessed outcomes relating to motivation and learning, which relate to hypotheses 1, 2 and 3 of this study. In this section I will detail how I measured these outcomes. All measures were collected in a Qualtrics questionnaire, except for research outcomes which were collected using game software that recorded player actions.

Motivation outcomes

My first measure related to motivation was the duration of time participants used the game or non-game software. The Qualtrics form was configured to automatically record the duration participants spent on each page in the form. One page provided participants with a link to access the software

and directed them to not proceed from this page until they were finished using the software and were ready to advance. This time duration was used to measure the duration participants used game and non-game software. I also asked players to report any breaks that they had while the page was open, and this information was subsequently used to adjust durations for self-reported breaks.

My second measure related to motivation was the number of playful experiences participants had using the software. This comprised my measure of "playfulness". Participants were asked to categorise the types of playful experiences, if any, they felt they had. A selection of playful experiences were identified and presented from the PLEX playful experiences framework (Lucero et al., 2013). This is a taxonomy of playful experiences designed to identify the specific types of experiences players have with a game, rather than using broad experiences like "fun" or "enjoyment" which can be harder to interpret and tie to individual game mechanisms. Twelve categories of experience were included in this measure and these are presented in Table 6.2. Participants selected categories by dragging each one into a box declaring either "I experienced this" or "I did not experience this". I recorded responses describing whether each participant experienced each individual experience (e.g., competition), but for hypothesis testing I will be primarily concerned with a composite score I computed for playfulness, and this counts the total number of playful experiences reported.

Table 6.2 Categories of playful experiences that participants could report feeling in the playful experiences framework

Experience	Description
Captivation	Forgetting one's surroundings
Challenge	Testing abilities in a demanding task
Competition	Contest with oneself or an opponent
Completion	Finishing a major task, closure
Discovery	Finding something new or unknown
Exploration	Investigating an object or situation
Fantasy	An imagined experience
Humour	Fun, joy, amusement, jokes, gags
Nurture	Taking care of oneself or others
Progression	Earning momentum and achievement
Relaxation	Relief from bodily or mental work
Sensation	Excitement by stimulating senses such as sights or sounds

Learning outcomes

I developed a multiple-choice questionnaire as a learning assessment. This was based on educational theories that there are different levels of understanding and that these can be assessed by asking learners to demonstrate various abilities. Bloom's Taxonomy of Learning (Anderson et al., 2001; Gogus, 2012) visualises these levels of understanding as a pyramid, with the lowest and easiest levels being the ability to remember the information; interpret, exemplify, and summarise it, and the higher, more difficult levels concerning critical evaluation and ability to use the information. I developed my MCQ on the principle that participants who were able to answer more questions correctly, including those requiring critical reasoning about network complexity, had achieved a better understanding of my network dataset.

I identified seven areas of competency which are relevant to understanding network complexity in public health and these would be assessed in my MCQ (Table 6.3). I used these competencies to formulate 22 questions that assess these competencies at various levels. The first questions in the MCQ ask participants to correctly recall information about the network dataset and later questions ask them to use this information to speculate about the effects of various interventions in the dataset. For example, question eleven asks participants to identify which factor in the network exerts the greatest overall effect on other factors in the network. The full list of questions and correct answers can be found in Appendix 6.3, and I will now outline the scoring.

The MCQ has 22 questions in total worth one point each for a correct answer, giving a maximum score of 22. All questions had a single correct answer and between two and four incorrect answers, except question 1 which required participants to identify multiple correct answers (at least 5 of the 6 correct answers). Final scores were corrected for the score obtainable by random guessing.

This assessment was pilot tested with six individuals who also participated in playtesting the game. Participants had some exposure to my network dataset through previous use of my game software, and this was similar to the experience of participants in my final study who would complete this

assessment following use of the software. Pilot testers found the assessment difficult, scoring an average of 49% (min=16%, max=58%) and taking on average 41.4 minutes to complete it (min=28 minutes, max=82 minutes). Accordingly, I improved the wording of questions, removed questions no participants could answer (the MCQ originally had 25 questions), and added three unscored questions intended to build confidence and reduce frustration. These actions made the assessment easier to complete, and I will discuss the impact of this in the discussion.

Table 6.3 Areas of competency assessed in the MCQ

Questions	Competency
1-4	Ability to read information about nodes and edges in the network visualisation
5-7	Understanding of direct effects: Infer the direct effects of interventions which increase the prevalence of a trait
8-11	Understanding of network properties: Ability to make inferences about the general pattern of effects in the network
12, 13	Understanding of interactions: Critically analyse interaction effects between multiple interventions which increase the prevalence of different traits
14, 15	Ability to negate effects: Infer the direct effects of interventions which decrease the prevalence of a trait
16-22	Understanding of indirect effects: Infer the indirect effects of interventions

Other measures

I collected information on two additional measures in the questionnaire. I used the first measure, descriptions of participants' playful experiences, to assess game design, as described in Chapter 5. Participants were also asked to recall their experience in their own words. They were provided a box to give an answer to "In your own words, please describe your experience with the software. For example, did you have a strategy? Did anything prevent you from achieving what you wanted? Did you find any effects memorable? Did you have any opinions about the presentation?". Answers to this question were reviewed in Chapter 5.

Another measure in my Qualtrics questionnaire was designed to assess usability. Usability describes the degree to which a tool can be used to fulfil a given objective, and is typically measured using self-report (Tullis & Albert, 2008). Participants were asked to indicate how usable they felt the software was. Responses were accepted on a scale from 0-100 in response to the question "Reflecting on your

experience with the interactive visualisation or game software please drag the bar below to indicate how much you found it easy to use", where values near to 0 were marked as "very difficult to use" and values near to 100 were marked as "very easy to use". Usability is relevant to motivation and education-related outcomes because it could impact duration of use and learning measures if participants in one condition found their software difficult to use. As such, this measure was intended to support my inferences about motivation- and learning-related outcomes of game features, by ensuring differences in usability did not impact these outcomes.

Research outcomes

Research outcomes were assessed using participants' responses and interactions with the game. Through the course of game-play, players in the game condition solved hypothetical problems in public health. In-game actions were recorded and used for assessing the research outcomes of the game. I will explain how players' actions were recorded, what information was recorded, and explain why this measure was only collected for participants in the game condition.

The main action taken by users of my game and non-game simulation software was implementing virtual interventions. In the game, these interventions doubled as solutions to public health problems posed as goals for players to achieve. Players were randomly assigned one of 13 public health goals (Table 6.4). Upon making an intervention and suggesting a solution to one of these problems, a record of this action was sent to a custom database. The information I recorded included an anonymous session ID tracking the player, the variable they performed an intervention on, their current goal, and the score they received for this action (scoring detailed below). The technical implementation of this system, tracking sessions not player identities, had consequences meaning that I could not link this data to player data collected in Qualtrics, and players who reloaded their game were tracked with multiple distinct session IDs. 107 sessions were recorded in total, and since this count is 17 higher than the total number of participants in the game condition, this indicates a minority of sessions were reloaded. Since it is not possible to determine which

sessions were made by the same participant, I proceed with analysis under the caveat that some sessions are linked and represent continued play sessions by the same participant.

A scoring system rewarded players for interventions that were valid in the context of their effectiveness achieving goals. Validity was assessed with respect to the mathematical effectiveness of interventions in the underlying network dataset. Interventions were scored from 0-100% where the maximum score was awarded to the most mathematically effective solution players could suggest. All other scores were scored as a percentage of the best score. For example, if the best solution to improve education increased years of schooling by 1 year then another solution which increased it by 0.8 years would be scored with 80%. Interventions were scored in this way to produce a standardised scoring system that was comparable across the 13 goals in the game. The effects of interventions were calculated using the propagation model outlined in the previous chapter.

Data were not recorded for participants in the control condition because they were not assigned goals. Players in the non-game were not directed to solve problems, but instead explored and experimented with data in a free-form manner, and these interactions with the software were recorded. Since the non-game software did not provide goals to achieve, it was only possible to apply this score to participants in the game condition, and I will discuss the implications of this later.

Table 6.4 13 problems players were faced with solving in the game

Lower alcohol consumption
Lower heart disease (CHD)
Lower diabetes
Lower insomnia
Lower loneliness
Lower neuroticism
Lower smoking
Lower worry
Raise education
Raise exercise
Raise intelligence
Raise socialising
Raise wellbeing

6.2.5 Data preparation

Data downloaded from Qualtrics as well as my custom software database were imported into *R* for analysis. Steps were taken prior to analysis to prepare the data. These steps included excluding participants who performed in ways that made them appear inattentive to the task, and adjusting durations of software use for self-reported break-taking. In this section, I will explain steps for excluding participants who were not engaged with the study and transforming variables for analysis.

Participants were excluded following group assignment by identifying outliers. Selected participants were removed from the study entirely (row-wise deletion). 5 participants were excluded on the basis of not having finished the study ($n_{controls}$=5, n_{game}=0). A further 3 participants were excluded on the basis of inattention since they spent excessively long periods of time using the software (>80 minutes, $n_{controls}$=0, n_{game}=1) or on completing the assessment (>60 minutes, $n_{controls}$=2, n_{game}=0). Similarly, 11 participants were excluded for spending excessively short periods of time on the software (<10 seconds, $n_{controls}$=2, n_{game}=0) or assessment (<7 minutes, $n_{controls}$=3, n_{game}=6). Figures presented in Appendix 6.2 support these exclusion criteria by demonstrating that participants spending excessively long on the study were statistical outliers, and that participants who spent less than 7 minutes on the learning assessment achieved scores which are consistent with being inattentive. In total, 19 participants were excluded ($n_{controls}$=12, n_{game}=7). Participants in the control condition therefore appear to have had more difficulties engaging with the study. The final sample size was 175 participants with complete measurement information ($n_{controls}$=85, n_{game}=90). These steps were followed for all outcome measurements, except for research outcomes where software records only permitted exclusion on the basis of time durations, where a comparable number of sessions were excluded (n=7).

Time durations recorded in the Qualtrics questionnaire were adjusted for any self-reported breaks, for example where participants spent 60 minutes using the software but reported a 10-minute break, their final recorded time was 50 minutes. 17 participants reported taking a break in the game

condition (mean duration=16.1min), 9 participants reported a break in the control condition (mean duration=6.7min).

6.2.6 Data analysis

I evaluated my three hypotheses that game features will improve motivational and learning outcomes, as well as exploring research outcomes in the game condition. In this section I explain how my statistical analyses assessed evidence for the outcomes of game features. Selection of statistical tests was guided by the distributions in my data. The distributions of my data are presented in Appendix 6.4 along with formal tests to support assertions that my data meets these assumptions (e.g., Shapiro-Wilks test of normality, and Bartlett's test for equal variances). This revealed that my measures were not normally distributed so I relied on non-parametric testing, which relaxes the strict assumptions of normality imposed by parametric tests such as the T-test comparison of means. In particular, I used the Mann-Whitney U test which, when comparing two similarly shaped distributions, functions as a comparison of medians with relaxed non-parametric assumptions. However, it still makes some assumptions, particularly that there are equal variances across comparison groups. I also used a Generalised Linear Model (GLM) where one measure, duration of use, exhibits a gamma distribution (as expected for this type of data), using an appropriate link function which incorporates information about the expected distribution into the model. The key assumption behind this model is that the link term appropriately describes the distribution of the outcome variable. I interpreted results considering the magnitude and statistical significance of effects (accepted α=.05).

Hypothesis testing

I tested two formal hypotheses about game outcomes related to motivation and learning. Hypotheses 1 and 2 relate to motivational outcomes, and hypothesis 3 relates to learning outcomes.

Outcomes related to motivation were assessed by comparing game and non-game participants on the duration they used software for, as well as how playful they found their experience to be.

Duration of use was distributed with a heavy right-skew (a gamma distribution), and so I investigated differences across groups using a generalised linear model that integrates information about this distribution (a gamma link). Playfulness was distributed with a non-normal distribution, so was tested using the non-parametric Mann-Whitney U test. Taken together, these analyses helped me assess motivational outcomes by evaluating support for hypotheses 1, that game features would extend duration of use, and 2, provide a more playful experience.

Learning outcomes were assessed by comparing MCQ learning scores across conditions. Test scores were distributed with a non-normal distribution, so differences were tested again using the Mann-Whitney U test. Furthermore, in order to assess whether my MCQ learning assessment was a valid assessment of learning, I tested the psychometric properties of the test, including identifying possible sub-scales using exploratory factor analysis, as well as measuring the internal consistency of my questionnaire using Cronbach's Alpha (Tavakol & Dennick, 2011). This helped me assess hypothesis 3 by testing whether game features improved learning of network health relationships.

Usability and research outcomes

Following hypothesis testing, I performed two additional analyses. The first analysis, on usability, informed the interpretation of the analyses described above by assessing whether differences between groups were caused by differences in software usability between the game and non-game conditions. I also investigated the research outcome by inspecting the solutions players offered to health problems during the course of gameplay.

I assessed differences in software usability by comparing usability ratings for the game and non-game conditions. This measure has a slight bimodal trend, so is not normally distributed, and differences were analysed using the Mann-Whitney U test. This helped me understand whether the game and non-game software were similarly usable, and explore a possible source of confounding for the hypothesis tests described above.

I assessed research outcomes by investigating the validity of the solutions players suggested throughout gameplay. Since only participants in the game condition took part in the data collection for research outcomes, no group comparison is available, so I instead describe the performance of players in the game. I report on players' scores as well as describing the volume and types of interventions players made. I estimated whether scores improved as players spent more time in the game, and completed more trials. However, most scores were close to either 0% or 100% and observations from the same participant were not independent of each other. I transformed scores into ordinal categories, representing 0%, 1-99%, 100% scores, and estimated the effects of in-game time and trials on these scores. These thresholds represent whether players were able to identify the optimal solution comprising the "correct answer" (100%), were not able to identify the correct answer but still gave a valid solution (1-99%), or gave an invalid answer (0%). The reason for a qualitative split between answers scoring say, 100% and 99%, is that discovering the correct answer was qualitatively emphasised in the game since it comprised the "win condition" where the player would triumph over their AI opponent who would otherwise identify a better solution than the player. I accounted for the linked nature of my data, caused by players contributing multiple data points, by fitting an ordinal mixed-effects model with participant ID as a random effect (see Appendix 6.4). This allowed me to understand the contributions players are able to make to a data collection exercise during gameplay. I will later discuss how restricting research tasks to game participants limited the inferences I could make about research outcomes.

6.3 Results

6.3.1 Descriptive statistics

Descriptive statistics are presented below for outcome measures of duration of use, playfulness, usability, MCQ learning scores and in-game scores (Table 6.5). The distributions of each measure in analysis are presented in Appendix 6.4, along with formal tests of distribution parameters, including equality of variance between groups and normality.

Table 6.5 Descriptive statistics for outcome measures

	Mean	SD	Min	Max	variance
Duration of use (mins)					
All	10.27	12.29	0.17	62.28	150.98
Control	2.79	3.76	0.17	16.31	14.17
Game	17.34	13.34	0.28	62.28	177.92
Playfulness (0-12)					
All	6.71	2.45	0	12	6.01
Control	5.62	2.28	0	10	5.19
Game	7.74	2.15	1	12	4.64
Usability (%)					
All	56.11	25.76	3.00	100	663.71
Control	56.52	25.17	3.00	100	633.30
Game	55.73	26.45	5.00	100	699.57
MCQ learning score (0-22)					
All	17.17	2.33	10	22	5.41
Control	16.92	2.11	11	21	4.46
Game	17.41	2.50	10	22	6.24
In-game score (%)					
Game	61.03	44.42	0	100	1973.18

Note: In-game scores were recorded across 107 sessions made by 90 participants in the game condition.

6.3.2 Effects of game features

Testing for differences in outcome measurements across the game and non-game conditions revealed substantial effects on motivation-related outcomes (Table 6.6). Participants in the non-game condition did not use the control software for long (mean=2.79 minutes), whereas players used the game for much longer (mean=17.34 minutes). Furthermore, players in the game condition also experienced more playfulness, reporting a greater number of playful experiences (mean=7.74) compared with controls (mean=5.62). Examining specific playful experiences, indicates players of the game particularly experienced feelings of completion, challenge and competition (Figure 6.1). There was some, but not strong, statistical evidence that the game and non-game conditions differed in terms of learning or software usability.

Table 6.6 Main effects for measures of motivation

	Game		**Controls**		**Test**			
	Mean	*SD*	*Mean*	*SD*	df	method	Effect coefficient	*P*

Duration of use	17.34	13.34	2.79	3.76	173	Gamma GLM	Regression coefficient= 1827	$1.84x10^{-09}$
Playfulness	7.74	2.15	5.62	2.28	173	Mann-Whitney	W=7.01	$4.99x10^{-11}$
MCQ learning	17.41	2.5	16.92	2.11	173	Mann-Whitney	W=3260	.089
Usability	55.73	26.45	56.52	25.17	173	Mann-Whitney	W=3848	.096

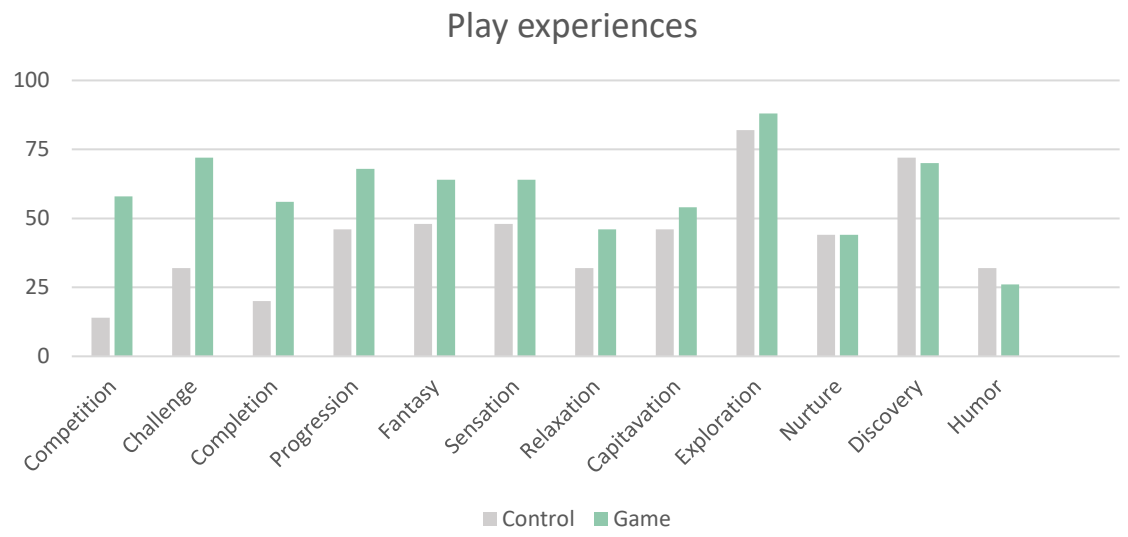


Figure 6.1 Number of participants reporting each play experiences with control and game software

Psychometric properties of the learning MCQ

The MCQ of learning suffered a ceiling effect, compressing the right tail of the distribution, because participants did not find the tasks challenging enough. The assessment also showed poor internal consistency (Cronbach's alpha=.43), so responses on individual questions were not highly predictive of the overall score (Tavakol & Dennick, 2011). This may indicate that different questions in my assessment were not measuring the same construct. However, it is likely that it is at least partly explained by the ceiling effect where most questions were answered correctly by large proportions of participants, reducing the variance along with the potential for co-variance (i.e., producing the results showing poor internal consistency). This is supported by the proportion of correct answers for individual items in the assessment, often over 80%, and a distribution of total scores with the right tail compressed because it is not possible to score more than 100%. I performed exploratory

factor analysis to identify whether instead my questionnaire contained sub-scales, which would indicate it measures multiple constructs in a valid way, but I was not able to find any evidence of sub-scales that sufficiently explain participants scores (maximum r^2<11%). However, this investigation was limited in a similar way to Cronbach's alpha since it relies on variance to detect co-variance. My full investigation is presented in Appendix 6.4, but the consequence of a ceiling effect is that it reduced my ability to assess whether there were any small differences in learning outcomes. I will discuss later how a better outcome measure for learning may have been better able to identify effects of gameplay on learning.

6.3.3 Research outcomes

I analysed the solutions participants in the game condition offered to 13 in-game public health problems. Players offered a total of 2,255 solutions across 107 sessions. Players offered between 1 and 68 solutions during a play session, with an average of 22 solutions per session (SD=17.5).

The game can be considered a controlled system, modelling a public health system that is determined only by the variables in my network dataset. Inspecting players' solutions for each of the problems in the game (Table 6.7) reveals that players solutions tended to be effective, though scores varied across individual problems. These data indicate that if players' solutions were used to crowd-source public health policies in this controlled system, the average suggestion would achieve results within 42-77% of optimal policy (mean=55%, SD=13%). This could be interpreted in two ways; on the one hand, players showed difficulties identifying the most effective policy, but on the other hand players solved problems with some efficiency despite having no formal training in epidemiology. Part of the reason why some problems were more difficult for players to answer, is that some problems had fewer valid solutions that would score players above 0%, and this made it harder for players to identify effective solutions.

Table 6.7 In-game problems and players' solutions

Target variable to improve	Number of valid solutions players could offer*	Players solutions				
		n	Mean score (%)	Min (%)	Max (%)	SD
Intelligence	7	119	76.55	0	100	40.34
Smoking	7	132	73.64	0	100	41.1
Diabetes	7	116	68.57	0	100	33.11
Coronary Heart Disease	8	145	65.21	0	100	39.94
Body Mass Index	6	134	63.47	0	100	42.63
Education	5	142	53.13	0	100	43.49
Worry	1	147	52.38	0	100	50.11
Caffeine consumption	1	136	47.36	0	100	48.47
Socialising	8	124	45.16	0	100	49.97
Wellbeing	4	137	44.59	0	100	38.28
Loneliness	2	160	44.22	0	100	49.7
Insomnia	3	144	42.4	0	100	43.01
Alcohol consumption	4	152	41.63	0	100	45.86

Note: * Valid solutions had at least some effect on the target variable. Some variables were more connected in the underling network dataset, and so are affected by more variables.

Inspecting how players scores changed over the course of gameplay also may indicate practice effects (Figure 6.2). Ordinal mixed effects linear modelling indicates that players scores improve over each minute of gameplay ($z=5.72$, $P=1.06\times10^{-08}$), as well as for each in-game trial ($z=4.60$, $P=4.24\times10^{-06}$). The session ID was entered as the dependent term linking responses from the same players. However, given that some players scores were only tracked within the same session, some players opened multiple sessions and so the mixed effects model only partially accounts for dependence. Consequently, this analysis will be interpreted as contributing suggestive evidence that players learn to better identify solutions over time, but no strong conclusion can be made as to the magnitude or statistical significance of this effect without fully accounting for dependence.

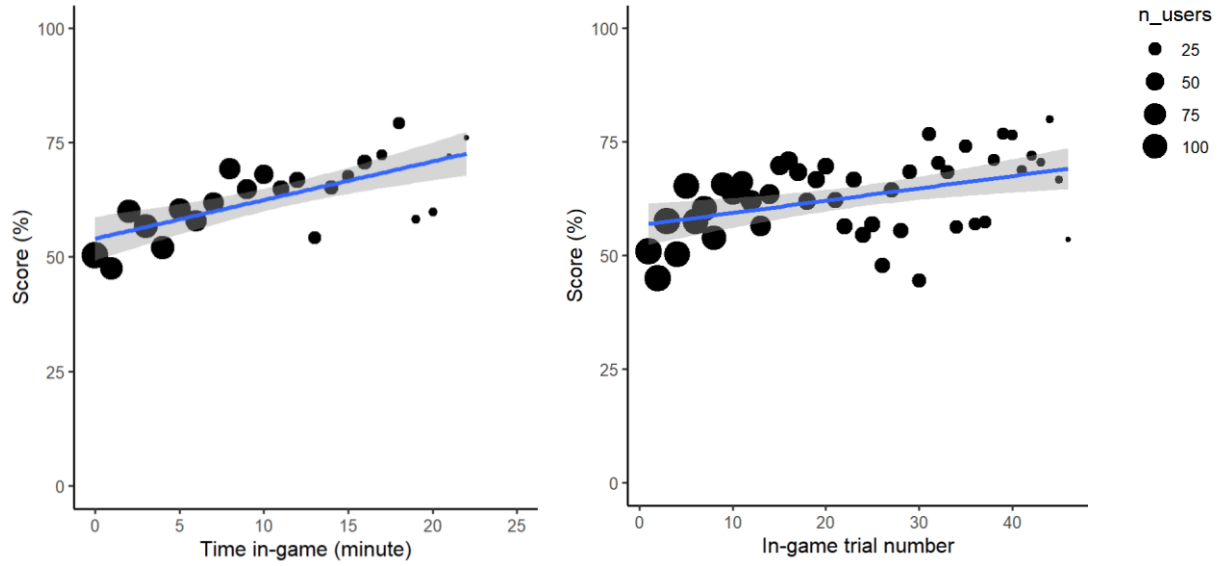


Figure 6.2 Players' ability to solve in-game public health problems appears to improve over the duration they spend playing the game (left), as well as the number of in-game trials they complete (right). A blue line illustrates the trend with standard error intervals. The number of users' scores captured at each data point are indicated with the size of points, with a corresponding scale presented in the legend at the top right ("n_users").

Finally, investigating the types of variables players chose to intervene on revealed the types of solutions players suggested. Interventions on some variables, such as exercise and body mass index (BMI), were less often suggested as the solutions to problems compared to other variables such as education and insomnia.

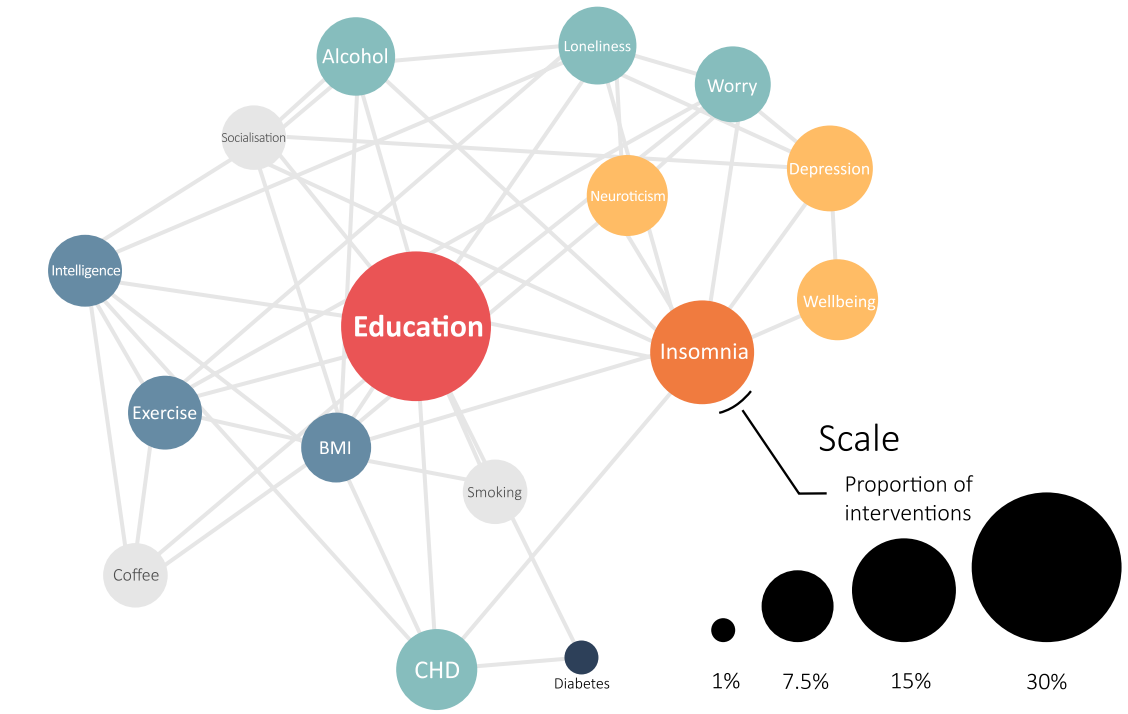


Figure 6.3 Players were more likely to suggest interventions on some variables than others. Nodes are scaled according to frequency (see legend) as were they coloured similarly (scale = blue:<7%, cyan:7-10%, yellow:11-14%, orange: 15-18%, red: >18%).

6.4 Discussion

In this chapter I conducted an experimental study comparing a data game with a non-game control. My aim was to investigate the effects of game features on educational and research outcomes. Here I will interpret my results with respect to three hypotheses: that game features will enhance motivation, education and research outcomes. I will discuss these in the background of previous research into educational and research games (Looyestyn et al., 2017; Ricciardi et al., 2014; Sailer & Homner, 2020) and draw out the implications of my findings. I will close this discussion by highlighting some advantages and limitations in my study design, and suggest how future research could improve the ways in which we conduct experimental game research.

First, I investigated the hypothesis that players would use the software for longer than controls. I found strong evidence to support these hypotheses since participants in the game condition

engaged for more than five times the duration controls did. The motivational effect in this study (*d*=1.45) is substantially larger than previous research which showed a negative effect of game features on the duration participants are willing to spend completing surveys (mean *d*=-0.30)(Looyestyn et al., 2017). This is likely due to my game incorporating more game features and better creating a cohesive and playful experience compared to previous software, such as lightly gamified surveys. This outcome is related to motivation and these findings support the ability for games to motivate players to engage with tasks for longer.

Second, I investigated the hypothesis that players would experience a more playful experience than controls. I found evidence to support this hypothesis as well, since players reported a greater number of playful experiences. In particular, players of the game experienced more feelings of completion, challenge and competition. These experiences are related to psychological needs to feel competence, and in a game they provide players the opportunity to meet these by overcoming challenges, showcasing their skills, and being rewarded for good performance (Deterding, 2015; Marczewski, 2018). The greater playful experience is likely attributable to my game featuring a "play-learn-improve-win" pattern of gameplay, where players were faced with a problem, given the resources to find a solution, and were rewarded for good answers (Malliarakis et al., 2014). This second outcome was also related to motivation and further supports games' ability to engage players with playful experiences.

Third, I investigated whether players of the game demonstrate a better understanding of network complexity. I found limited evidence for this since players scored marginally higher on an MCQ assessment of learning, though this assessment showed a ceiling effect. There is not strong evidence for whether the game features I introduced enhanced learning outcomes, though the effect of learning in my study was smaller (*d*=0.22) compared to positive effects that have been reported before (mean *d*=0.75)(Sailer & Homner, 2020). In the present study, it is likely that the ceiling effect reduced my ability to detect differences in learning outcomes. The ceiling effect was partly produced

by my participants already having a level of public health knowledge and therefore were able to perform at a very high level in the MCQ assessment of learning. I will review in the discussion chapter how future research could improve interpretability of learning outcomes by using pre-testing or by basing data games on obscure datasets.

Reflecting on these three hypothesis tests, it is possible to rule out one source of influence that may have otherwise biased results and confused interpretation. Although there is some evidence of usability differences between the game and non-game software, the ratings were similar (mean of 56% versus 57%), so this likely did not have a large influence on the other outcome measures.

Finally, I investigated research outcomes by inspecting the solutions players suggested for problems presented in-game. Players contributed a substantial number of solutions, and their scores indicated they represented valid solutions in the underlying mathematical simulation they were presented with. Furthermore, there is evidence that players improved over time, and this supports previous claims that players can learn to contribute more accurate data over the duration of a play session (Mavandadi et al., 2012). However, it is unclear whether these observations are due to players simply rote-learning the correct answers to each problem, or a real practice effect where players learned to suggest more effective solutions to problems during the course of gameplay. The manner of recording players sessions, rather than unique individuals, further complicates my interpretation of practice effects because some players reloaded the game and continued playing in new sessions. Overall, research outcomes indicate that players can contribute solutions during play which can be used as valid data points in research projects. Therefore, I found encouraging evidence that players can play a research game, take it seriously, and contribute to data collection exercises.

In summary, I found evidence to support the ability for game features to enhance outcomes related to motivation, as well as demonstrating players' ability to contribute to a game-based data collection exercise. I will now discuss how future research can interpret and build on my findings, to continue investigating the value of game features for education and research purposes.

6.4.1 Implications

The chapter demonstrates that complex topics in health, such as public health interventions, can be modelled in data games with the potential to achieve educational and research outcomes. Players were able to understand the topic and even suggest valid solutions to complex problems. MR in particular is a highly specialist topic and researchers have struggled to explain it to lay audiences due to its complex basis in genetics and epidemiology. My findings imply that public health information, and MR specifically, can be effectively implemented into a game in a way that makes it accessible and understandable to novice audiences. Previous games about MR have been effectively used to explain some concepts, such as odds ratios (http://www.bristol.ac.uk/integrative-epidemiology/engagement/), and my game demonstrates that this concept can be extended using computer modelling to engage players in real MR data. The next step for researchers interested in citizen science would be to transform the wealth of available MR and genomics information into a game which allows lay audiences to contribute to ongoing research. A previous project (Free Ice Cream & Davis, 2018) demonstrates that MR data can be integrated into an interactive simulation and that this concept was well-received in a university setting. Therefore, a key implication of this chapter is demonstrating the concept of the MR data game and showing that it can achieve outcomes for future MR education and research.

6.4.2 Strengths and limitations

I constructed an effective experimental comparison between game and control conditions but weaknesses in my measures prevented me from making stronger conclusions about the effects of game features. In particular, I should have developed a more valid learning assessment, collected demographic information, and assessed research outcomes in the control condition. I will now discuss how future research could improve on these aspects.

A key strength of my study was that I designed an experimental comparison which allowed me to attribute the differences across groups to game features. By constructing a game with many game

features and ensuring it met my desired play experience I exposed participants to a strong experience of game features. This suggests that the small differences in educational outcomes, for example, is not simply due to a weak experimental manipulation. By contrast, reviewers estimate that 37-46% studies use software and activities that contain just one game feature, such as scoring, and this functions as a weak dose of game features which is less likely to produce observable effects (Looyestyn et al., 2017; Sardi et al., 2017). Additionally, by constructing a control condition that was very similar in many details, I controlled the number of factors that could have caused differences in outcomes besides game features. I mentioned in the introduction to this chapter that studies using poorly matched controls tend to find exaggerated effects of game features and this likely indicates bias. In my study I was better able to determine that the effect on motivation was most likely caused by game features and not some other factor. Lastly, by measuring the user experience I was able to understand the types of experiences participants had with the game and non-game software. It is often assumed that games achieve their desired play experience, but by measuring this I was able to estimate that players experienced 38% more playful experiences. This falls within a range of previous findings that players in game conditions report 18% - 67% more agreement to play experience statements like “I enjoyed using the game/non-game” (Papastergiou, 2009; Sward et al., 2008). Therefore, I was able to attribute differences between groups to game features by making a strong manipulation of game features, controlling for other differences, and verifying my manipulation worked as intended. This is especially important since many fields of games research, such as data games, are still emerging, and my study highlights gaps in research while demonstrating some good practices for how they can be investigated.

One limitation in my study was that it was difficult to interpret the results for my learning outcomes. The questions on my MCQ learning assessment showed a ceiling effect, and this made it difficult to understand what the overall scores were measuring. Since most participants answered questions correctly, this reduced the variance and differences between participants responses. Although as much as 90% of games research uses self-developed learning assessments (Sailer & Homner, 2020),

including MCQs (44%)(Kara, 2021), my measure did not work as intended. Adopting a standardised measure would have ensured my measure was proven to be reliable prior to testing. However, no standardised assessments currently exists for assessing knowledge of network complexity. It is worth mentioning that my use of non-parametric testing reduced my power to detect statistically significant differences in many of the measures used in my study though in practice the unreliability of my MCQ materials contributed the largest issue. Therefore, future research should build in time for pilot testing measures to develop questionnaires with favourable psychometric properties prior to the experimental study.

Another limitation is that I did not collect information on the age and sex of my participants. When talking about demographics in gaming it is important to acknowledge that the traditional stereotypes for who is likely to play and enjoy games have been disproven since 47.8% of players are women (ISFE, 2021), who enjoy games as much as men do (Wang & Wang, 2008), and the population of game players is equally distributed in terms of age. Therefore, individuals are equally likely ot enjoy games regardless of age or gender. However, there might be some specific gendered effects. For example, one study reports that women benefit more from an educational game, showing greater learning benefits compared to the men in the study (Garber et al., 2017). However, I was not able to investigate this because I had not collected demographic information. Similarly, while I know the demographics of my population, I cannot be sure that my sample contains more men or young adults than expected. Therefore, while traditional stereotypes about game players are not true, future research should continue collecting demographic information so that they can investigate any possible differences.

One final limitation with my study was that one of my research outcome measurements was only completed by players of the game. Participants in both conditions completed the same learning exercise but the data collection exercise was different in that players of the game were guided to solve problems during gameplay. I collected players' solutions to these problems and I was able to

analyse their quantity and quality, but I was not able to compare the quality of solutions to participants in the control condition because they did not complete the same problem solving research exercise. For example, I could not investigate whether gameplay increased the number of contributions participants made, as claimed by previous research (Looyestyn et al., 2017). This could have been remedied by providing participants in the non-game condition with problems to solve as well. Initially I did not do this because goals are a quintessential component of games (Deterding et al., 2011; Salen & Zimmerman, 2003), and I felt it would risk gamifying the control condition. However, in light of my results that participants reported that the control condition was playful anyway, I believe that adding game features would have improved my ability to infer the effects of game features on research outcomes. Additionally, this would have allowed me to address claims that game features reduce the accuracy of research contributions (Cancer Research UK, 2015). Therefore, ensuring the control and experimental conditions involve identical tasks would ensure participants can be compared on all outcome measurements of interest.

6.4.3 Future directions

Future research should focus on improving the quality of validation approaches in games research. One method of achieving this would be addressing the specific challenges faced by games researchers. Game design and development constitute large tasks for games researchers, but current frameworks are driven by the values of professional game designers (Fullerton, 2018; Salen & Zimmerman, 2003). Games researchers might instead start by developing game design frameworks from a research perspective. In particular, researchers should be guided to develop a non-game experimental control. Inspiration could be taken from gamification frameworks (Werbach & Hunter, 2012) which instruct designers to start from a non-game task, guaranteeing its availability for comparison in a subsequent study. However, these frameworks also guide designers to add limited game features and so new frameworks are required to guide researchers to develop full games from non-game tasks. The robustness of validation approaches in games research is an

important area to discuss and I will continue highlighting possible areas for improvement in the final discussion chapter 7.

6.4.4 Conclusion

In this chapter I experimentally investigated the outcomes for adding game features to a public health simulation. I investigated three main outcomes. First, game features had substantial effects on outcomes related to motivation, including longer play durations and more playful experiences. Second, game features did not substantially enhance learning outcomes in this case. Third, players were able to suggest valid solutions to public health problems during gameplay. A key strength of my study was improving on many common validation approaches in games research by constructing a strong experimental comparison. Future research should build on this by developing effective outcomes measurements and continuing to improve experimental games methodology. In the next chapter I will further discuss the implications of these findings, as well as my previous chapters, in the context of the games literature.

7 Discussion

This thesis addressed an issue that will become increasingly problematic in science. As our methods of data collection and analysis become increasingly sophisticated, and research outputs become more nuanced, we will have to develop new ways of understanding complex data. I sought to understand how we can best make sense of network relationships between health variables. In chapter one, I introduced how this challenges public health researchers who seek to understand the variables that cause ill health. Methods of causal analysis are essential tools, but their large-scale application produces complex webs of causal effects which are difficult to disentangle. In chapter two, I demonstrated one particular method of causal analysis, Mendelian randomisation (MR), and highlighted its utility along with its strict assumptions. In chapter three, I extended the previous analysis in a network MR framework that obtains causal estimates, not just between two variables, but for entire networks of inter-related health variables. The output was a decidedly complex network dataset indicating health variables are highly causally inter-related. In chapter four, I developed a visualisation tool that I have made available for researchers to visualise network relationships in their own research, as well as forming the foundation for subsequent software developments. In chapter five, I developed my dataset and visualisation tool into an interactive simulation of public health interventions. I followed a game design process to add game features and produce a data game that engages players in the underlying simulation. Finally, in chapter six, I combined what I had learned and developed in previous chapters to conduct an experimental study evaluating what impacts adding game features had on outcomes related to motivation, education and research.

In this final chapter I discuss the contributions my thesis makes to the field, and the limitations of my work that present opportunities for future research. In particular, I will consider how a focus on insomnia and wellbeing influenced my studies, and suggest how games researchers can test the effects of game features in more robust ways, as well as drawing on my experience working in a

policy-making environment at Public Health Wales to discuss the practicalities of implementing games in practice. Evidently, this work is also multi-disciplinary in nature and so I will discuss the research impacts of different perspectives and values. The key findings from each chapter are summarised in Table 7.1.

Table 7.1 Main findings for each empirical chapter and the implications for studying data games

Chapter	Main findings	Implications for studying network complexity in public health
2 - Testing the causal relationship between insomnia and wellbeing: A Mendelian randomisation Study	An MR effect estimate indicates that insomnia causally reduces wellbeing.	MR can be used to estimate the causal relationships that may exist between a wide range of variables including psychological variables.
	Genetic instruments for wellbeing are weak and this reduced the reliability of causal effect estimation in the reverse direction (for wellbeing on insomnia).	Weak instruments for psychological variables reduce researchers' ability to detect causal effects reliably.
3 - Exploring the network of relationships among variables related to insomnia and wellbeing	I used hypothesis-free MR to obtain a network dataset describing the relationships between 16 variables.	MR can be extended in a network framework to describe the structure of relationships among many inter-related variables, including indirect effects.
	My network dataset appears to be a valid representation of the type of network complexity in public health.	This adds to a limited body of network MR evidence that demonstrates that public health is a complex system involving many inter-related causal variables.
4 - MiRANA: A tool for visualising network relationships in MR	My scoping review documented available packages for network MR analysis and revealed a niche for network visualisation tools.	The MRC Integrative Epidemiology Unit has used this information to identify the gaps in software provision for MR researchers more accurately.

	I developed software for MR researchers to visualise network relationships in their results.	This tool will help meet a growing need for MR researchers to visualise complex relationships.
5 - Turning a causal health network visualisation into a data game	I developed a public health simulation based on network MR data.	MR has utility beyond estimating simple causal effects and I demonstrated how it can be used to simulate the effects of complex public health interventions.
	I gamified this by adding game features to create a data game.	Data games can be created in an academic design process.
6 - Evaluating a network data game using an experimental control	Game features encouraged players to engage with an underlying network dataset for longer.	The mechanisms behind game features are seldom studied but my results provide strong evidence that game features create a playful experience that motivates players to continue engaging.
	Game features did not substantially improve learning outcomes. Players suggested valid solutions to public health problems.	These findings add to a limited body of evidence experimentally comparing the effects of game features. Future research should continue to identify and refine the optimal research methods.

7.1 Focus on insomnia and wellbeing

My thesis began with selecting the types of health variables I would focus on as examples of network complexity. Owing to my background in Psychology, personal interests, and the expertise of my lab group, I opted to focus on wellbeing and insomnia. However, I did not foresee two key difficulties that selecting these variables would present. First, I discovered that there are numerous challenges to using MR to model the causal effects of psychological variables. Second, selecting well-known

variables limited my ability to assess learning outcomes from my data game. The field of data games primarily exists to identify and solve challenges in transforming data to game play, so these are highly relevant, and I will discuss them in this section.

7.1.1 Opportunities and challenges of using MR in Psychology

Using MR gave me the opportunity to study the causal pathways among a large range of variables, but psychological variables present some challenges that are difficult to overcome. I will summarise some advantages and challenges I introduced in previous chapters (2, 3) before tying this discussion into a recently published article on the subject (Wootton et al., 2022).

I discussed in a previous chapter (2) that the primary advantages of MR include its efficient and ethical study of diverse variables, in part due to advances in methods such as two-sample MR (Lawlor, 2016), as well as the large amounts of information held in publicly accessible genomics datasets. Additionally, formal methods for hypothesis-free (Hemani, Bowden, et al., 2017) and network MR (Burgess et al., 2015b) allowed me to extend MR to perform network analyses. MR therefore performed a critical role in facilitating my studies. However, there are limitations to this approach.

Most psychological variables are complex traits and this makes identifying associated genetic variants difficult. In a previous chapter (2) I explained that complex traits tend to have individually small but additive genetic influences, and researchers often note that this results in multiple weak instruments that cannot be used to obtain reliable causal estimates (Jansen et al., 2019; Wootton et al., 2018; Zhou et al., 2021). I also explained that adequately powered GWAS do not exist to reliably detect marginally associated genetic variants. My network MR study (3) demonstrates that weak instruments reduce the likelihood that effect estimates for psychological variables exceed thresholds for statistical significance. Therefore, it is currently difficult to perform MR reliably with psychological variables.

Psychological variables are also difficult to directly observe, and this results in a reliance on self-report measures that can be vulnerable to several sources of bias. Also, the science of wellbeing is relatively young, so cohorts are unlikely to have access to best-practice measurement scales and instead rely on diverse measures of convenience that make it difficult to aggregate measurements and combine samples into cohorts that are sufficiently large for a powerful GWAS (Okbay et al., 2016b). Furthermore, I noted in a previous chapter (2) that self-report measures may contribute a level of measurement error that reduces the accuracy of measurement for psychological variables (e.g., investigating polysomnography for sleep measurement)(Jean-Louis et al., 2000; Lauderdale et al., 2008), although it seems likely that self-report captures different aspects of human behaviour that are not available to other approaches.

A recent article covers some more of the issues using MR to study psychological variables (Wootton et al., 2022). The paper details four issues. The first two issues concern the difficulties measuring psychological traits that I detailed above, as well as the consequences of violating assumptions like linearity that I covered in a previous chapter (3). The third issue concerns an increased risk of pleiotropy. Since psychological variables are more complex and have more relationships with other variables, this increases the likelihood of actions through pleiotropic pathways. My study corroborates this claim by adding to previous evidence that there are wide-spread indications of pleiotropy and many variables have genetic correlations (Hemani, Bowden, et al., 2017). The article argues that a defence against pleiotropy is conducting robust sensitivity testing. I outlined previously (1) that sensitivity analyses compare effect estimates across multiple genetic instruments. However, many of my sensitivity analyses compared few genetic instruments and this reduced the reliability of these results. I have explained why genetic instruments for psychological variables are scarce, but it is worth noting that an additional consequence of this is that it reduces the ability for sensitivity analyses to detect signs of pleiotropy. Researchers typically conduct sensitivity testing with at least 5 (Wootton et al., 2022) or 10 (Hemani, Bowden, et al., 2017) instrumental variables, however my studies (3) indicate that many causal effect estimates between psychological variables are likely to

have fewer instruments (<3). Therefore, psychology researchers should be mindful that sensitivity analyses are likely to be less reliable where few genetic variants are available.

The fourth issue covered in the article is that many relationships in psychology are plausibly bi-directional (Wootton et al., 2022). This challenges conventional MR analyses which investigate acyclic biological processes which have finite start and end points (Suttorp et al., 2015). Psychological researchers must seek to understand bi-directional effects more closely, and my studies have built methods for understanding complex relationships like these. My visualisation software (4) provides a starting point for communicating and understanding these cyclical relationships and, though my simulation (5) omits cyclical relationships, it builds understanding of complex relationships and could be adapted to convey cyclical relationships too. Therefore, visualisation and simulation methods may have a place in understanding cyclical effects in MR.

In summary, there are several challenges when using MR to study the relationships between psychological variables. Psychological variables tend to have weak instruments, are difficult to measure, and the effects can be complicated and difficult to discern. My studies add to these areas by demonstrating that psychological variables present few effects in a network analysis and by building tools for researchers to continue exploring and understanding their effects.

Consequences of data selection

The variables I selected for analysis had consequences for my ability to assess learning outcomes. In this section I will recap my selection of wellbeing and insomnia as the focal points for my analysis, explain how these variables caused a ceiling effect in a subsequent learning assessment, and outline how this issue could be remedied in future data games.

I selected insomnia and wellbeing as my focal variables and selected variables related to these for a network analysis. I explained that part of my criteria was to include variables that a novice audience could recognise and reason about, such as exercise and education (1,3). In this way I was able to improve the interpretability of effects in a network dataset (Hemani, Bowden, et al., 2017) used in a

precursor game project (Free Ice Cream & Davis, 2018). However, including these variables negatively impacted my ability to assess outcomes when comparing game and non-game modes of learning about the dataset.

Players brought their own intuitions about the relationships between variables. In free-form feedback about the game (5) players noted that relationships in the game often confirmed their intuitions and expressed surprise when they did not. This impacted my subsequent experimental study (6) since players brought existing knowledge about the variables in my study. This contributed to a ceiling effect in learning outcomes since the correct answers to some questions were congruent with common knowledge, for example: “Does depression increase or reduce wellbeing?”. Additionally, the level of this pre-test knowledge may have varied across the conditions. These factors reduced my ability to detect learning outcomes and attribute to game or non-game condition assignment.

These two issues could have been controlled for by adapting the methods I used to test learning outcomes. The impact of the former issue, participants bringing in pre-existing knowledge about variables could be resolved by transforming data into obscure and unfamiliar formats. For example, in my game I might have renamed the variables in my network to fictitious diseases. This would have preserved the subject I wanted to expose participants to, public health network complexity, while preventing players from applying existing knowledge in a subsequent learning assessment. The impact of the latter issue, potential differences in pre-test knowledge, could be accounted for with additional stages of testing. Conducting pre-testing would establish baseline knowledge which could be compared to a post-test to accurately gauge learning outcomes. Therefore, implementing a combination of pre- and post-testing, or obscuring the true variables in my data, would have made it easier for me to measure the learning outcomes from my data game.

7.2 Quality of evidence in games research

A general issue was the poor quality of evidence for the effects of adding game features to educational and research programs. The main problem I outlined in the previous chapter (6) was the lack of well-constructed experimental research (M. Brown et al., 2016; Johnson et al., 2016; Kara, 2021). This is evidenced by reviews of research quality since these typically report that studies satisfy half the criteria for robust research, such as using appropriate designs, statistical analysis, and correctly interpreting findings (Looyestyn et al., 2017). I also combined findings across the literature to demonstrate that poor research designs consistently bias effect estimates to exaggerate the effects of game features to be larger than they really are. The consequences of this cannot be over-stated and a high-profile example is the association between videogames and violence. An official report by the American Psychological Association taskforce on violent media stated that videogame play causes violence but a recent re-examination of this determined that only low-quality research suggested such an association (Ferguson et al., 2020). Poor research quality is therefore an important and prevalent issue to address before we can understand the effects of game features. In this section I will discuss some potential solutions to this, including adopting a gold standard for evidence, forming operational definitions of games, and focussing game design on essential features only.

7.2.1 A gold standard for games research

Researchers might consider adopting a gold standard for evidence in the search for the effects of game features. A "gold standard" describes a method which is considered the most accurate and reliable way of collecting evidence in a given field (Webb et al., 2020). For example, well-constructed randomised control trials (RCTs) are considered the strongest form of evidence in epidemiology. I will present both the case for and against implementing a gold standard.

Implementing a gold standard is likely to improve awareness and adherence to best practices, and raise overall research quality. I demonstrated that a strong experimental design allowed me to

determine the effects of game features by comparing game and non-game participants (6). Following some gold standard practices more closely would have improved the strength of evidence in my studies, such as conducting pre- and post-testing, and it is likely others would benefit too. While establishing a gold standard does not guarantee that researchers will notice, follow or properly implement recommendations, it would at minimum raise awareness of good practices (Grossman & Mackenzie, 2005). Furthermore, RCT gold standards are already used in some areas of games research including gamifying healthcare interventions to improve patient outcomes (M. Brown et al., 2016; Fleming et al., 2017; Sardi et al., 2017). Therefore, a gold standard would likely have a positive impact on the evidence quality in games research.

A gold standard is not, however, suitable for all researchers to follow. There are good reasons that some researchers might not follow a gold standard for experimental design. Some researchers use qualitative methodology to describe players' experiences in rich detail, where strong evidence is marked by researchers' abilities for reflexivity and understanding individual perspectives (Aleven et al., 2010; Sasupilli et al., 2019). The criteria some researchers use for assessing strength of evidence are different, and so the gold standard for methods like RCTs would be inappropriate for qualitative researchers since it does not meet their criteria. Similarly, data game researchers often tackle software issues with solutions and tests that are objectively demonstrable, so have less need for RCT-like causal designs (Barros, 2016; Friberger et al., 2013). Therefore, one type of gold standard, such as for RCTs, would not be appropriate for all types of games researchers.

In summary, a gold standard would likely benefit experimental games researchers. It would at minimum raise awareness of best practices and improve strength of evidence overall. However, different types of gold standards would be required to account for the different perspectives and values of other researchers, for example using qualitative or software methods.

7.2.2 Operationalising game features

Researchers should operationalise the aspects of games that they believe will produce specific effects. In previous chapters (1, 5) I introduced that while there are definitions for what games and serious games are, there is disagreement over what the constituent ingredients of games are. There is disagreement over what the important features of serious games are, be they the engaging value of gameplay education (Clapper, 2018; Michael & Chen, 2006; Zyda, 2005b), the serious content (Clapper, 2018; Stoll, 1999), or both (Marc Prensky, 2001). As a result, game terminology is used to inconsistently refer to games with different features and purposes (Figure 7.1). My studies (5, 6) demonstrate an important contribution that games researchers can integrate into their research. By theoretically isolating the "active ingredients" of a game, I was able to better attribute down-stream outcomes to individual game features. Researchers should adopt similar strategies as this helps better define and observe the aspects of games that can produce beneficial effects for education and research.

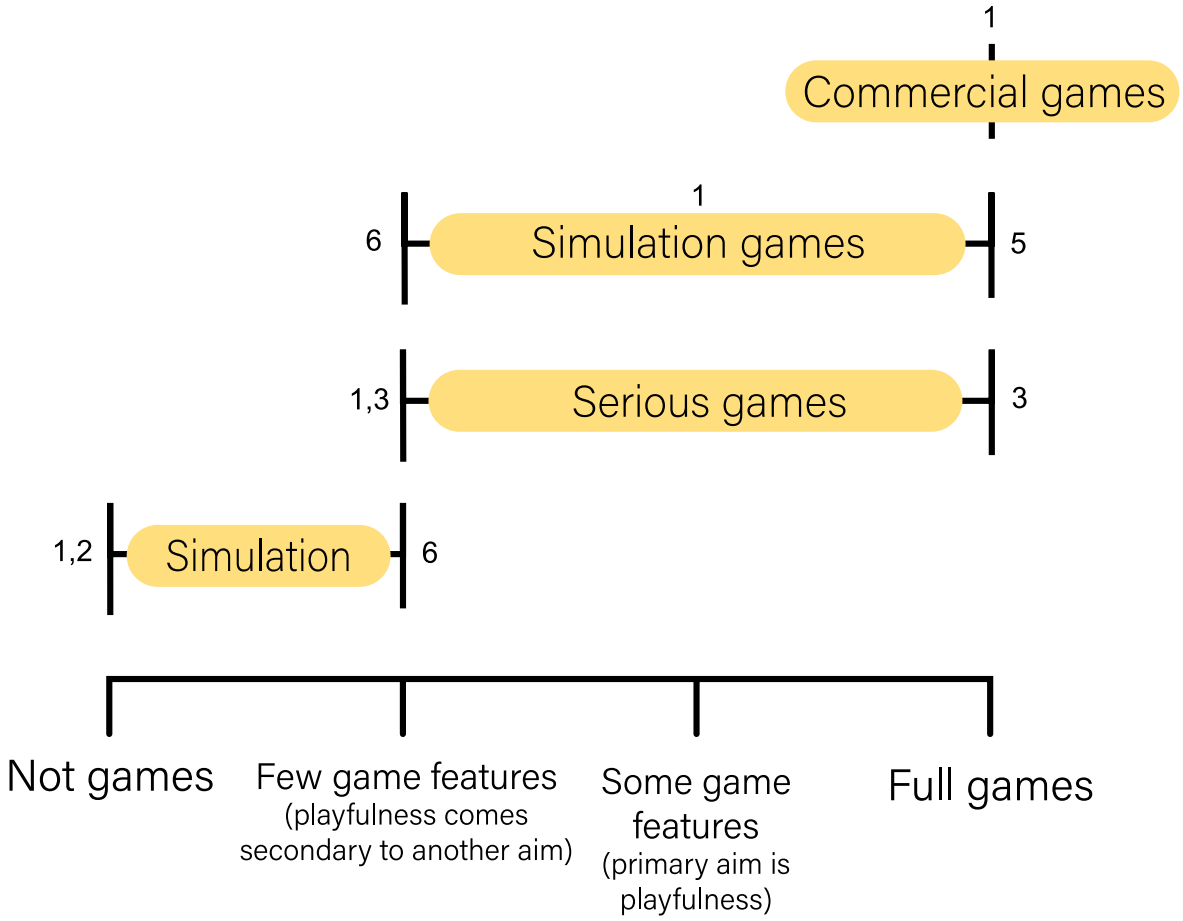


Figure 7.1 Researchers use the same terms in overlapping ways to describe different types of games that have different implementations of game. References indicate use of terms: [1](Qin et al., 2010), [2](Ricciardi et al., 2014), [3](Sailer & Homner, 2020), [4](Schrier, 2016), [5](Wardaszko, 2018), [6](Geurts et al., 2007).

7.2.3 Style over substance

One final aspect of the games research process is to consider to what degree aesthetics contribute to effective game design. "Style over substance" refers to the general notion that a game might look aesthetically pleasing but not deliver a substantial play experience. I will briefly outline this issue and suggest that while large game development companies have dedicated artistic resources, small scale serious games might be bogged down with a focus on achieving aesthetic fidelity.

The effort spent on game aesthetics might be better spent on designing more compelling gameplay loops. I conclude from my game development process (5) that I should have directed more effort to refining my gameplay loop, with more extensive playtesting, and I could have achieved this by designing fewer sound and graphic elements. A counterargument might be that players'

appreciation of the aesthetic elements was one of the most prevalent themes to emerge from feedback, so clearly was an enjoyable aspect of gameplay. It is argued the audio-visual theme of a game helps immerse players (Hunicke et al., 2004), such as in the fantasy setting of World of Warcraft where sounds and visuals are used to transport players to a new world (Blizzard Entertainment, 2005). However, there is no direct evidence that it contributes to outcomes like learning or research. This is not a well-researched area but comparing case studies of serious games would indicate that the most popular citizen science game of all time (FoldIt) does not incorporate an extensive audio-visual theme, whereas other games do (e.g., Play to Cure)(Figure 7.2). It is important to note that these citizen science games did not have large development teams because while the argument of misdirecting resources may be true for teams with shared responsibilities, this is not true for established professional teams, such as the developers of World of Warcraft, who hire dedicated art and sound specialists. Therefore, researchers should perform a cost-benefit analysis before investing limited resources into audio-visual components for data games.

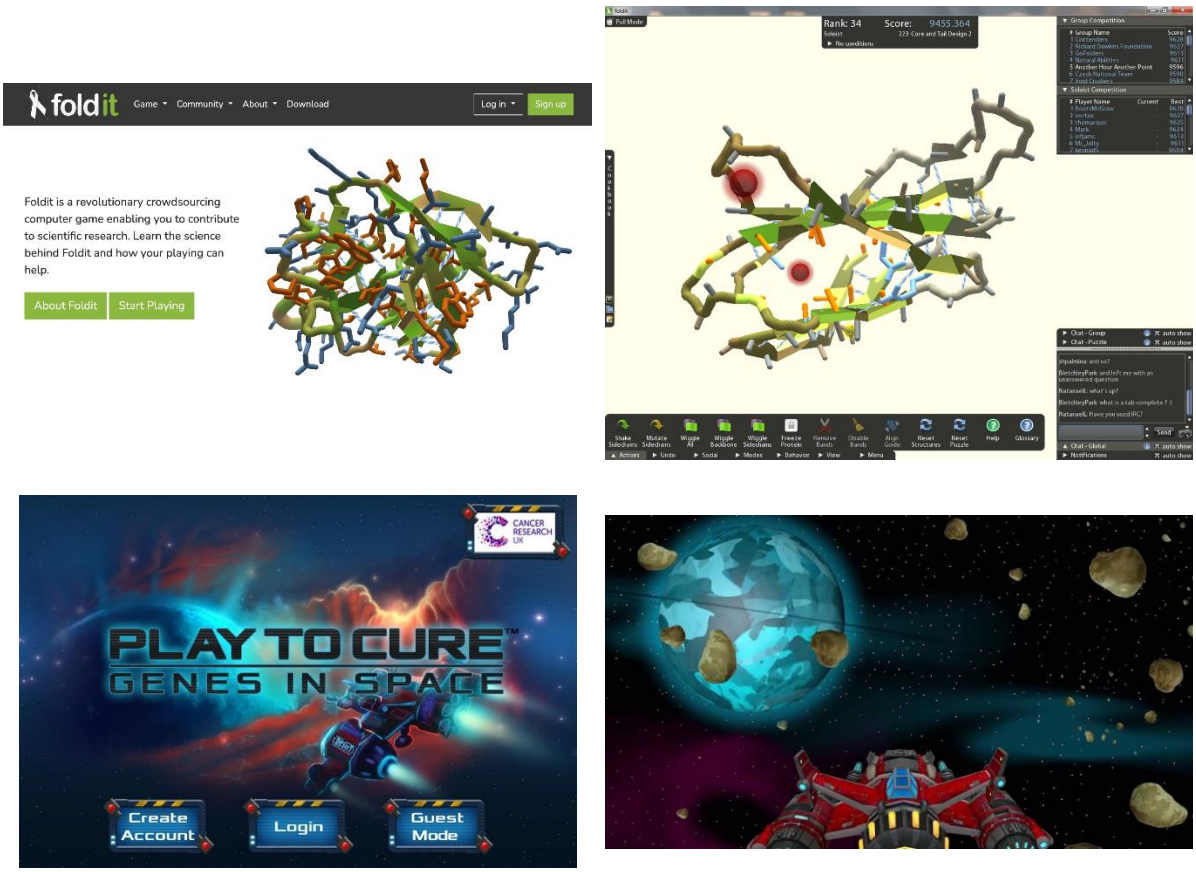


Figure 7.2 Many serious games, like FoldIt (top), have simple graphics but some, like Play to Cure (bottom), have more complex graphics. It is not clear whether differences in graphical treatments affect the outcomes, since both games were very successful at engaging players in research projects.

7.2.4 Interim summary

In this section I outlined three issues in games research that could be improved with specific actions. Games research is a broad multi-disciplinary field but researchers investigating the effects of game features should adopt gold standard experimental methodology, develop operational definitions of what game features they are manipulating, and identify the most important aspects of games to ensure research efforts do not focus on style over substance.

7.3 Reflecting on games in practice

As part of my PhD experience, I worked in public health placements. As someone with skills relevant to ongoing relief efforts, I felt I should contribute, so I volunteered with the EndCoronaVirus.org charity (2020) and completed a placement with Public Health Wales (2021). These experiences taught me important lessons about applying visualisation and gamification into real-world public heath settings. In particular, future serious games should ensure gameplay is matched to be relevant to the underlying topic or dataset, and to ensure that information is conveyed in appropriate ways to avoid information overload or over-simplification.

Between June 2020 and March 2021, I volunteered as a software developer for EndCoronaVirus.org. I helped develop two games intended to communicate new economic modelling about the effects of COVID-19 lockdowns. The first game set players as the leader of a fictional country who could use a dashboard to enact COVID-19 lockdown policies (“Pandemic Game”, Elias et al., 2020). The player would win by finding a combination of policies that would limit the spread of infection while being cost effective. This first game was delivered to policy aides in Ireland and Israel who reported enjoying the game and felt it communicated the data in an interesting and relevant way. The second game also put players in a role leading a country, but was designed for a general public audience (“OutBreak”, Moreno-Stokoe et al., 2021). Players navigated a series of events in a narrative

adventure game, they responded to events that occurred over the course of the pandemic, and tried to achieve a balance between implementing policies which were popular, economical and effective in limiting illness. Player choices would have consequences later in the story, based on survey data and modelling, and they would win if they navigated four events and achieved the best scenario outcome. However, this game was not received well, as players from the general public did not feel the political role was relevant to their everyday lives. This related to findings from my user interviews (5) which support the importance of relevant gameplay since MR researchers commented that it would be important for my data game to be relevant to a real-world problem. It is also possible that players of my game, who took a similar role, also felt similar. Taken together, these experiences taught me valuable lessons in terms of the high-level decision making process for designing a relevant game. This is relevant for future research since data games should communicate information in a context that is relevant for the player.

Between March 2020 and May 2021 I worked with Public Health Wales on data visualisation projects relating to COVID-19. This taught me the importance of selecting appropriate visualisations since some are effective for some purposes but not others. In the introduction chapter (1), I introduced the difference between infographics, which focus on communicating key messages, and visualisations, which facilitate the user to explore a dataset and arrive at their own conclusions. In an initial project, I helped develop an interactive map visualisation that assisted policy makers in identifying the areas of Wales that were most vulnerable to COVID-19 (di Cara et al., 2021). This tool did not communicate one single message but instead users analysed vulnerability according to their own criteria, selecting from overlays and risk variables they felt were relevant. In this case it was most appropriate to provide an interactive visualisation that allowed users to customise and view data in detail. In contrast with this case, in a second project with Public Health Wales during an internship, I developed an accessible infographic report that summarised key points in a simple manner. I worked as a researcher on a placement where I produced a report on the provision of healthcare for the homeless population in Wales (J. Song et al., 2022). This required a detailed and

diligent analysis of summary care records, but most readers are not expected to make it past the first page where “take-away” messages are presented. These messages were intended to be communicated to a broad audience, and to make my results immediately clear, I opted to use simple and targeted infographic visuals. These experiences emphasised the importance of developing effective visualisations to facilitate specific inferences, such as the micro, relationship and macro inferences for which I developed the network visualisation tool (4).

In summary, working in public health during the pandemic reinforced some of the lessons I learned during my thesis and helped me appreciate how to best communicate information through gameplay. Volunteering with EndCoronaVirus.org demonstrated that setting players in relevant roles can contextualise information in terms of the real-life scenario it is intended to be applied in, and working with Public Health Wales demonstrated the importance of tailoring visualisation strategies for specific purposes.

7.4 Strengths and limitations

In this section I will summarise the key strengths and limitations I outlined across my studies (2-6). I will draw out the three main contributions my thesis made, and how some limitations present opportunities for future research.

First, using MR allowed me to build a network dataset describing the relationships among a wide range of variables related to wellbeing, mental and physical health. The MR method was essential to my studies, and using it to obtain evidence for a causal network constitutes a core strength of my thesis. However, the types of psychological variables I studied tend not to have strong instrumental variables and are at higher risk of producing estimates biased by pleiotropy. These factors reduced the reliability of my estimates, since I cannot be certain that the relationships I found, and did not find, for psychological variables represent true causal pathways. Improving the availability and quality of psychological variables in MR is therefore a core priority to be addressed by future research.

Second, I conducted software design and development processes in a systematic manner (4,5). The strength of this was borrowing elements from established academic frameworks and performing user research. This allowed me to collect and analyse information about existing software, my intended users, and evaluate what decision decisions would be most effective in a more rigorous and reliable manner. However, game development is still an emerging field and there are not well established methods for performing academic game design. As a result, I focussed more on some areas, such as constructing a valid experimental manipulation, than others, such as conducting user research. One of the main limitations of my studies was in the user research processes. The COVID-19 pandemic affected the manner in which I conducted playtesting, resulting in fewer opportunities for detailed in-person playtesting sessions, but I could have also conducted more targeted user interviews at the beginning stages of the user research. Addressing these limitations would help future research design a more effective game product. There is therefore scope to improve my design process by strengthening its weaker aspects while maintaining its stronger aspects.

Finally, I demonstrated that adding game features to a public health simulation engaged players to spend more time exploring and experimenting with an underlying dataset (6). I used a robust experimental design to investigate the outcomes of game features and found strong evidence that game features encouraged players to engage for longer. This adds to accumulating experimental evidence that supports the motivational effects of games. However, I did not find strong evidence for the learning and research effects of games and this may be due to methodological issues. My evidence was limited by a few key factors including not controlling for pre-test knowledge, not ensuring all outcome measures were available for both game and non-game conditions, as well as using an MCQ assessment which showed a strong ceiling effect. These issues could be addressed by spending longer designing and pilot testing the materials prior to conducting the experimental study. Future research should continue investigating the outcomes of game features by experimentally comparing game and non-game conditions.

In summary, my thesis contributed to the study of data games and network complexity in three key ways. I developed a dataset representing the network complexity in public health, transformed this into a data game through an academic game design process, and demonstrated that game features can help engage players in this data. Future research can build on this work by further formalising a systematic process of game design and developing effective materials for testing the effects of game features on learning and research outcomes.

7.5 Future directions

In this final section, I will close my thesis by outlining the direction for future research that follows on from my studies and that I believe would be most beneficial for the field of data games. I outlined previously (6) that games have been used to collect useful data that has contributed to real research projects, including cancer and genetics research, and this could be extended to MR as well. A data game could help evaluate the plausibility of causal relationships in maps of the human phenome. By crowd-sourcing opinions from epidemiologists with different backgrounds and specialities, a data game could engage researchers to judge whether, according to their knowledge, causal relationships are plausible or not. Fully validating the putative causal estimates in maps of the human phenome would be a very large undertaking, requiring manual analysis and triangulation of evidence for thousands of estimates (Hemani, Bowden, et al., 2017), but a data game could provide some preliminary evidence to shortlist some of the most plausible relationships that have not received attention before. This would therefore be a useful output that would advance our understanding of public health. My studies indicate that a data game solution to this problem would be technologically feasible and such a game would benefit from the groundwork I have laid in this thesis in terms of visualising MR network data (4), transforming it into gameplay (5) and demonstrating the effects of gameplay on engagement (6). Furthermore, my interviews with MR researchers (5) suggested that they would be interested in playing a game designed for researchers. This is corroborated by previous research involving researchers at the University of Bristol which found that

a gamified research project, using prediction markets (Munafo et al., 2015), resulted in crowd-sourced judgements that predicted which 70% accuracy which academic papers would successfully replicate (Thompson & Munafo, 2019). Therefore, my thesis has established precedent, means and motive to develop a data game to help validate causal maps of the human phenome.

7.6 Conclusion

My thesis investigated the network complexity in causal pathways between health variables, and explored novel approaches for understanding them. My overall aim was to develop methods that help us understand the complexity of public health and I believe my findings achieve this because my initial chapters set up the empirical foundation for this problem and obtained a dataset describing network complexity which was explored using novel software approaches in subsequent chapters. The use of MR allowed me to estimate the causal effects between 16 physical and mental health variables (2,3), though the methods for investigating psychological variables are currently limited. I then developed visualisation, simulation and game methods of exploring this dataset (4, 5) using a design process that serves as an example for future research to continue building academic game design frameworks. However, my user research could have been more targeted and better structured according to clear goals, as this would have helped me design a data game that was more relevant to the real-life problems MR researchers face. Finally, I experimentally tested whether game features achieve outcomes related to motivation, education and research over traditional non-game mediums (6). I found strong evidence that game features can improve player engagement with an underlying dataset, though my design was limited in ways that prevented me from making strong inferences about the educational or research outcomes of gameplay. Future research can iterate on my methods by ensuring players' prior levels of knowledge do not interfere with learning outcomes, and by ensuring participants engage in identical activities across game and non-game conditions.

I draw three final conclusions. First, MR and visualisation methods can help us document the network complexity in public health. Second, visualisation methods help us communicate this

complexity. Third, interactive simulation and game methods can help us better understand this complexity through exploring and experimenting with data as part of engaging experiences.

My hope is that, by demonstrating the potential for engaging participants with data games and highlighting exciting avenues of exploration for combining data games with MR, this thesis inspires research into a new generation of data games in population health science.

References

Albrecht, J. S., Wickwire, E. M., Vadlamani, A., Scharf, S. M., & Tom, S. E. (2019). Trends in insomnia diagnosis and treatment among medicare beneficiaries, 2006–2013. *The American Journal of Geriatric Psychiatry*, *27*(3), 301–309. https://doi.org/10.1016/j.jagp.2018.10.017

Alessi, C., Martin, J. L., Fiorentino, L., Fung, C. H., Dzierzewski, J. M., Rodriguez Tapia, J. C., Song, Y., Josephson, K., Jouldjian, S., & Mitchell, M. N. (2016). Cognitive behavioral therapy for insomnia in older veterans using nonclinician sleep coaches: Randomized controlled trial. *Journal of the American Geriatrics Society*, *64*(9), 1830–1838. https://doi.org/10.1111/jgs.14304

Aleven, V., Myers, E., Easterday, M., & Ogan, A. (2010). Toward a framework for the analysis and design of educational games. *2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, 69–76. https://doi.org/10.1109/DIGITEL.2010.55

Alhola, P., & Polo-Kantola, P. (2007). Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric Disease and Treatment*, *3*(5), 553–567. http://www.ncbi.nlm.nih.gov/pubmed/19300585

Anacleto, R., Badoni, S., Parween, S., Butardo, V. M., Misra, G., Cuevas, R. P., Kuhlmann, M., Trinidad, T. P., Mallillin, A. C., Acuin, C., Bird, A. R., Morell, M. K., & Sreenivasulu, N. (2019). Integrating a genome-wide association study with a large-scale transcriptome analysis to predict genetic regions influencing the glycaemic index and texture in rice. *Plant Biotechnology Journal*, *17*(7), 1261–1275. https://doi.org/10.1111/pbi.13051

Anderson, L. W., Bloom, B. S., & others. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Andreatta, P. B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., Stern, S., & Frankel, J. (2010). Virtual reality triage training provides a viable solution for disaster-preparedness. *Academic Emergency Medicine*, *17*(8), 870–876. https://doi.org/10.1111/j.1553-2712.2010.00728.x

Badsha, Md. B., & Fu, A. Q. (2019). Learning causal biological networks with the principle of mendelian randomization. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00460

Barker, D., & Clegg, R. (1994). *Fast-track: A rad approach*. Addison Wesley Longman .

Barros, G. A. B. (2016). *Exploration of Open Data through Procedural Content Generation*. 1–2.

Barros, G. A. B., Liapis, A., & Togelius, J. (2015). Data adventures. *Proceedings of the FDG Workshop on Procedural Content Generation in Games*.

Barros, G. A. B., Liapis, A., & Togelius, J. (2016). Murder mystery generation from open data. *Proceedings of the 7th International Conference on Computational Creativity, ICCC 2016*, 197–204.

Barros, G. A. B., & Togelius, J. (2015). Balanced Civilization map generation based on Open Data. *2015 IEEE Congress on Evolutionary Computation, CEC 2015 - Proceedings*, 1482–1489. https://doi.org/10.1109/CEC.2015.7257063

Bartels, M. (2015). Genetics of wellbeing and its components satisfaction with life, happiness, and quality of life: a review and meta-analysis of heritability studies. *Calcified Tissue International*, *96*(3), 137–156. https://doi.org/10.1007/s10519-015-9713-y

Bartle, R. (1996). Hearts, clubs, diamonds, spades: players who suit muds. *Journal of MUD Research*, *1*(1), 19. https://www.hayseed.net/MOO/JOVE/bartle.html%0Ahttps://www.researchgate.net/profile/Richard_Bartle/publication/247190693_Hearts_clubs_diamonds_spades_Players_who_suit_MUDs/links/540058700cf2194bc29ac4f2.pdf

Baselmans, B. M. L., Jansen, R., Ip, H. F., van Dongen, J., Abdellaoui, A., van de Weijer, M. P., Bao, Y., Smart, M., Kumari, M., Willemsen, G., Hottenga, J.-J., Boomsma, D. I., de Geus, E. J. C., Nivard, M. G., & Bartels, M. (2019). Multivariate genome-wide analyses of the well-being spectrum. *Nature Genetics*, *51*(3), 445–451. https://doi.org/10.1038/s41588-018-0320-8

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. van der. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 205395172110138. https://doi.org/10.1177/20539517211013868

Beneš, H., Kurella, B., Kummer, J., Kazenwadel, J., Selzer, R., & Kohnen, R. (1999). Rapid onset of action of levodopa in restless legs syndrome: a double-blind, randomized, multicenter, crossover trial. *Sleep*, *22*(8), 1073–1081. https://doi.org/10.1093/sleep/22.8.1073

Bien, S. A., & Peters, U. (2019). Moving from one to many: insights from the growing list of pleiotropic cancer risk genes. *British Journal of Cancer*, *120*(12), 1087–1089. https://doi.org/10.1038/s41416-019-0475-9

Biggs, J. (2003). Aligning teaching for constructing learning John Biggs Keywords What is constructive alignment ? Defining the ILOs. *Education*, *94*(11), 112106. http://egusdsecondaryed.pbworks.com/f/aligning_teaching_for_constructing_learning.pdf

Bishir, M., Bhat, A., Essa, M. M., Ekpo, O., Ihunwo, A. O., Veeraraghavan, V. P., Mohan, S. K., Mahalakshmi, A. M., Ray, B., Tuladhar, S., Chang, S., Chidambaram, S. B., Sakharkar, M. K., Guillemin, G. J., Qoronfleh, M. W., & Ojcius, D. M. (2020). Sleep deprivation and neurological disorders. *BioMed Research International*, *2020*, 1–19. https://doi.org/10.1155/2020/5764017

Bisson, C., & Luckner, J. (1996). Fun in Learning: the pedagogical role of fun in adventure education. *Journal of Experiential Education*, *19*(2), 108–112. https://doi.org/10.1177/105382599601900208

Blizzard Entertainment. (2005). *World of Warcraft*. https://worldofwarcraft.com/en-gb/

Boggiss, A. L., Consedine, N. S., Brenton-Peters, J. M., Hofman, P. L., & Serlachius, A. S. (2020). A systematic review of gratitude interventions: Effects on physical health and health behaviors. *Journal of Psychosomatic Research*, *135*, 110165. https://doi.org/10.1016/j.jpsychores.2020.110165

Bou Sleiman, M., Jha, P., Houtkooper, R., Williams, R. W., Wang, X., & Auwerx, J. (2020). The gene-regulatory footprint of aging highlights conserved central regulators. *Cell Reports*, *32*(13), 108203. https://doi.org/10.1016/j.celrep.2020.108203

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. https://doi.org/10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314. https://doi.org/10.1002/gepi.21965

Bowden, J., del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N. A., & Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *International Journal of Epidemiology*, dyw220. https://doi.org/10.1093/ije/dyw220

Bowden, J., Fabiola Del Greco, M., Minelli, C., Smith, G. D., Sheehan, N. A., & Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample mendelian randomization analyses using MR-Egger regression: The role of the I 2 statistic. *International Journal of Epidemiology*, *45*(6), 1961–1974. https://doi.org/10.1093/ije/dyw220

Bowden, J., Spiller, W., del Greco M, F., Sheehan, N., Thompson, J., Minelli, C., & Davey Smith, G. (2018). Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *International Journal of Epidemiology*, *47*(4), 1264–1278. https://doi.org/10.1093/ije/dyy101

Bradford Hill, A. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Braun, V., & Clarke, V. (2012). Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Brice, J., Buck, N., & Prentice-Lane, E. (1993). *British Household Panel Survey User Manual*. University of Essex.

Brion, M.-J. A., Shakhbazov, K., & Visscher, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *International Journal of Epidemiology*, *42*(5), 1497–1501. https://doi.org/10.1093/ije/dyt179

Brown, B., & Knowles, D. (2020). *Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization*. 1–34. https://doi.org/10.1101/2020.06.18.160176

Brown, M., O'Neill, N., van Woerden, H., Eslambolchilar, P., Jones, M., & John, A. (2016). Gamification and Adherence to Web-Based Mental Health Interventions: A Systematic Review. *JMIR Mental Health*, *3*(3), e39. https://doi.org/10.2196/mental.5710

Buckingham, S. D. (2008). Scientific software: seeing the SNPs between us. *Nature Methods*, *5*(10), 903–908. https://doi.org/10.1038/nmeth1008-903

Bulpitt, C. J. (1997). Quality of life as an outcome measure. *Postgraduate Medical Journal*, *73*(864), 613–616. https://doi.org/10.1136/pgmj.73.864.613

Burgess, S., Bowden, J., Fall, T., Ingelsson, E., & Thompson, S. (2017). Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants. *Epidemiology*, *28*(1), 30–42. https://doi.org/10.1097/EDE.0000000000000559

Burgess, S., Butterworth, A., & Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, *37*(7), 658–665. https://doi.org/10.1002/gepi.21758

Burgess, S., Daniel, R. M., Butterworth, A. S., & Thompson, S. G. (2015a). Network Mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways. *International Journal of Epidemiology*, *44*(2), 484–495. https://doi.org/10.1093/ije/dyu176

Burgess, S., Daniel, R. M., Butterworth, A. S., & Thompson, S. G. (2015b). Network Mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways. *International Journal of Epidemiology*, *44*(2), 484–495. https://doi.org/10.1093/ije/dyu176

Burgess, S., & Thompson, S. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, *40*(3), 755–764. https://doi.org/10.1093/ije/dyr036

Burgess, S., & Thompson, S. G. (2015). *Mendelian Randomisation* (1st ed.). CRC Press.

Burgess, S., & Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal of Epidemiology*, *32*(5), 377–389. https://doi.org/10.1007/s10654-017-0255-x

Buttussi, F., Pellis, T., Cabas Vidani, A., Pausler, D., Carchietti, E., & Chittaro, L. (2013). Evaluation of a 3D serious game for advanced life support retraining. *International Journal of Medical Informatics*, *82*(9), 798–809. https://doi.org/10.1016/j.ijmedinf.2013.05.007

Cai, J., Wohn, D. Y., & Freeman, G. (2019). Who purchases and why? Explaining motivations for in-game purchasing in the online survival game fortnite. *CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 391–396. https://doi.org/10.1145/3311350.3347196

Cancer Research UK. (2015). *Citizen science games*. Https://Www.Cancerresearchuk.Org/Get-Involved/Citizen-Science/the-Projects#citizenscience2.

Cannon, H. M., Feinstein, A. H., & Friesen, D. P. (2010). Managing complexity: Applying the conscious-competence model to experiential learning. *Developments in Business Simulations and Experiential Learning, 37*, 172–182.

Cappuccio, F. P., Miller, M. A., & Lockley, S. W. (2010). *Sleep, health and society*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199566594.001.0001

Carr, A., Cullen, K., Keeney, C., Canning, C., Mooney, O., Chinseallaigh, E., & O'Dowd, A. (2021). Effectiveness of positive psychology interventions: a systematic review and meta-analysis. *The Journal of Positive Psychology*, *16*(6), 749–769. https://doi.org/10.1080/17439760.2020.1818807

Carter, A. R., Sanderson, E., Hammerton, G., Richmond, R. C., Davey Smith, G., Heron, J., Taylor, A. E., Davies, N. M., & Howe, L. D. (2021). Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *European Journal of Epidemiology*, *36*(5), 465–478. https://doi.org/10.1007/s10654-021-00757-1

Carter, M., Moore, K., Mavoa, J., Horst, H., & gaspard, luke. (2020). Situating the appeal of fortnite within children's changing play cultures. *Games and Culture*, *15*(4), 453–471. https://doi.org/10.1177/1555412020913771

CD Projekt. (2020). *Management board report on the activities of the cd projekt group and cd projekt s.a. in 2020*.

CD Projekt Red. (2015). *The Witcher*. https://www.thewitcher.com/en/witcher3

Cechanowicz, J., Gutwin, C., Brownell, B., & Goodfellow, L. (2013). Effects of gamification on participation and data quality in a real-world market research domain. *Proceedings of the First International*

Conference on Gameful Design, Research, and Applications, 58–65. https://doi.org/10.1145/2583008.2583016

Centers for Disease Control and Prevention. (2013). *Plague Inc.* CDC Public Health Matters Blog. https://blogs.cdc.gov/publichealthmatters/2013/04/plague-inc/

Chan, A. W., Yu, D. S., Choi, K., Lee, D. T., Sit, J. W., & Chan, H. Y. (2016). Tai chi qigong as a means to improve night-time sleep quality among older adults with cognitive impairment: a pilot randomized controlled trial. *Clinical Interventions in Aging*, *Volume 11*, 1277–1286. https://doi.org/10.2147/CIA.S111927

Cheng, S., L. Egues, A., & Cohen-Brown, G. (2018). Visualizing Medicine: Mapping Connections with Plague Inc. To Learn in the Interdisciplinary Classroom. In *Interdisciplinary Place-Based Learning in Urban Education* (pp. 111–132). Springer International Publishing. https://doi.org/10.1007/978-3-319-66014-1_6

Chignon, A., Bon-Baret, V., Boulanger, M.-C., Li, Z., Argaud, D., Bossé, Y., Thériault, S., Arsenault, B. J., & Mathieu, P. (2020). Single-cell expression and Mendelian randomization analyses identify blood genes associated with lifespan and chronic diseases. *Communications Biology*, *3*(1), 206. https://doi.org/10.1038/s42003-020-0937-x

Choi, K. W., Chen, C.-Y., Stein, M. B., Klimentidis, Y. C., Wang, M.-J., Koenen, K. C., & Smoller, J. W. (2019). Assessment of Bidirectional Relationships Between Physical Activity and Depression Among Adults. *JAMA Psychiatry*, *76*(4), 399. https://doi.org/10.1001/jamapsychiatry.2018.4175

Chou, Y.-K. (2014). *"Actionable gamification." Beyond Points, Badges, and Leaderboards*.

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, *32*(4), 556–577. https://doi.org/10.1002/sim.5408

Clapper, T. C. (2018). Serious games are not all serious. *Simulation and Gaming*, *49*(4), 375–377. https://doi.org/10.1177/1046878118789763

Coburn, C. (2014). Play to Cure: Genes in Space. *The Lancet Oncology*, *15*(7), 688. https://doi.org/10.1016/S1470-2045(14)70259-1

Cornelis, M. C., Byrne, E. M., Esko, T., Nalls, M. A., Ganna, A., Paynter, N., Monda, K. L., Amin, N., Fischer, K., Renstrom, F., Ngwa, J. S., Huikari, V., Cavadino, A., Nolte, I. M., Teumer, A., Yu, K., Marques-Vidal, P., Rawal, R., Manichaikul, A., … Chasman, D. I. (2015). Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Molecular Psychiatry*, *20*(5), 647–656. https://doi.org/10.1038/mp.2014.107

Cornelis, M., & Munafo, M. (2018). Mendelian Randomization Studies of Coffee and Caffeine Consumption. *Nutrients*, *10*(10), 1343. https://doi.org/10.3390/nu10101343

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. Harper & Row.

Daan, S., Beersma, D. G., & Borbely, A. A. (1984). Timing of human sleep: recovery process gated by a circadian pacemaker. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *246*(2), R161–R183. https://doi.org/10.1152/ajpregu.1984.246.2.R161

Das, K. v., Jones-Harrell, C., Fan, Y., Ramaswami, A., Orlove, B., & Botchwey, N. (2020). Understanding subjective well-being: perspectives from psychology and public health. *Public Health Reviews*, *41*(1), 25. https://doi.org/10.1186/s40985-020-00142-5

Dashti, H. S., Jones, S. E., Wood, A. R., Lane, J. M., van Hees, V. T., Wang, H., Rhodes, J. A., Song, Y., Patel, K., Anderson, S. G., Beaumont, R. N., Bechtold, D. A., Bowden, J., Cade, B. E., Garaulet, M., Kyle, S. D., Little, M. A., Loudon, A. S., Luik, A. I., … Saxena, R. (2019). Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nature Communications*, *10*(1), 1100. https://doi.org/10.1038/s41467-019-08917-4

Davey Smith, G., & Hemani, G. (2014a). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), 89–98. https://doi.org/10.1093/hmg/ddu328

Davies, N. M., Holmes, M. v., & Davey Smith, G. (2018). Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ (Online)*, *362*. https://doi.org/10.1136/bmj.k601

de Freitas, S., & Jarvis, S. (2006). A framework for developing serious games to meet learner needs. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*.

de Geus, E. J. C. (2021). A genetic perspective on the association between exercise and mental health in the era of genome-wide association studies. *Mental Health and Physical Activity*, *20*, 100378. https://doi.org/10.1016/j.mhpa.2020.100378

de Puy, W. H. (1883). *People's Cyclopedia of Universal Knowledge*. The Jones Bros. Publishing Co.

de Sousa Borges, S., Durelli, V. H. S., Reis, H. M., & Isotani, S. (2014). A systematic mapping on gamification applied to education. *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, 216–222. https://doi.org/10.1145/2554850.2554956

Deak, M., & Epstein, L. J. (2009). The history of polysomnography. *Sleep Medicine Clinics*, *4*(3), 313–321. https://doi.org/10.1016/j.jsmc.2009.04.001

Deci, E., & Ryan, R. (2012). *The Oxford Handbook of Human Motivation* (R. M. Ryan, Ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399820.001.0001

Deleuze, G., & Guattri, F. (1972). *Capitalism and Schizophrenia*. Les Éditions de Minuit.

Deterding, S. (2015). s. *Human–Computer Interaction*, *30*(3–4), 294–335. https://doi.org/10.1080/07370024.2014.993471

Deterding, S., Khaled, R., Nacke, L. E., & Dixon, D. (2011). Gamification: toward a definition. *CHI 2011 Gamification Workshop Proceedings*, 12–15.

di Cara, N. H., Song, J., Maggio, V., Moreno-Stokoe, C., Tanner, A. R., Woolf, B., Davis, O. S., & Davies, A. (2021). Mapping population vulnerability and community support during COVID-19: a case study from Wales. *International Journal of Population Data Science*, *5*(4), 1409. https://doi.org/10.23889/ijpds.v5i4.1409

di Tomasso, D. (2011). *Beyond gamification: Architecting engagement through game design thinking*. http://www.slideshare.net/ditommaso/beyond-gamification-architectingengagement-through-game-design-thinking

Diehl, L. A., Souza, R. M., Alves, J. B., Gordan, P. A., Esteves, R. Z., Jorge, M. L. S. G., & Coelho, I. C. M. (2013). Insuonline, a serious game to teach insulin therapy to primary care physicians: design of the game and a randomized controlled trial for educational validation. *JMIR Research Protocols*, *2*(1), e5. https://doi.org/10.2196/resprot.2431

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*(3), 542–575. https://doi.org/10.1037/0033-2909.95.3.542

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and open questions in the science of subjective well-being. *Collabra: Psychology*, *4*(1). https://doi.org/10.1525/collabra.115

Diener, E., Oishi, S., & Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour*, *2*(4), 253–260. https://doi.org/10.1038/s41562-018-0307-6

Diener, E., Pressman, S. D., Hunter, J., & Delgadillo-Chase, D. (2017). If, why, and when subjective well-being influences health, and future needed research. *Applied Psychology: Health and Well-Being*, *9*(2), 133–167. https://doi.org/10.1111/aphw.12090

Dixon, S., Morgan, K., Mathers, N., Thompson, J., & Tomeny, M. (2006). Impact of cognitive behavior therapy on health-related quality of life among adult hypnotic users with chronic insomnia. *Behavioral Sleep Medicine*, *4*(2), 71–84. https://doi.org/10.1207/s15402010bsm0402_1

Djaouti, D., Alvarez, J., Jessel, J.-P., & Rampnoux, O. (2011). Origins of serious games. In *Serious Games and Edutainment Applications* (pp. 25–43). Springer London. https://doi.org/10.1007/978-1-4471-2161-9_3

Doherty, A., Smith-Byrne, K., Ferreira, T., Holmes, M. v., Holmes, C., Pulit, S. L., & Lindgren, C. M. (2018). GWAS identifies 14 loci for device-measured physical activity and sleep duration. *Nature Communications*, *9*(1), 5257. https://doi.org/10.1038/s41467-018-07743-4

Driscoll, H. C., Serody, L., Patrick, S., Maurer, J., Bensasi, S., Houck, P. R., Mazumdar, S., Nofzinger, E. A., Bell, B., Nebes, R. D., Miller, M. D., & Reynolds, C. F. (2008). Sleeping well, aging well: a descriptive and cross-sectional study of sleep in "successful agers" 75 and older. *The American Journal of Geriatric Psychiatry*, *16*(1), 74–82. https://doi.org/10.1097/JGP.0b013e3181557b69

Driver, H. S., & Taylor, S. R. (2000). Exercise and sleep. *Sleep Medicine Reviews*, *4*(4), 387–402. https://doi.org/10.1053/smrv.2000.0110

Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Players, F., Stoddard, B. L., Popovic, Z., & Baker, D. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology*, *30*(2), 190–192. https://doi.org/10.1038/nbt.2109

Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental health and academic success in college. *The B.E. Journal of Economic Analysis & Policy*, *9*(1). https://doi.org/10.2202/1935-1682.2191

Elias, B., Felix, B., Moreno-Stokoe, C. M., Frankel, S., Martinovici, A., & Bar-Yam, Y. (2020). *"Pandemic Game": COVID-19 game for policy makers*. http://pandemic-game-prod.s3-website.us-east-2.amazonaws.com

Elliott, B. (2007). Anything is possible: Managing feature creep in an innovation rich environment. *2007 IEEE International Engineering Management Conference*, 304–307. https://doi.org/10.1109/IEMC.2007.5235049

Elsworth, B., Mitchell, R., Raistrick, C., Paternoster, L., Hemani, G., & Gaunt, T. (2019). *MRC IEU UK Biobank GWAS pipeline version 2*. https://doi.org/https://doi.org/10.5523/bris.pnoat8cxo0u52p6ynfaekeigi

Epic Games. (2017). *Fortnite*. https://www.epicgames.com/fortnite/en-US/home

Espie, C. A., Fleming, L., Cassidy, J., Samuel, L., Taylor, L. M., White, C. A., Douglas, N. J., Engleman, H. M., Kelly, H.-L., & Paul, J. (2008). Randomized controlled clinical effectiveness trial of cognitive behavior therapy compared with treatment as usual for persistent insomnia in patients with cancer. *Journal of Clinical Oncology*, *26*(28), 4651–4658. https://doi.org/10.1200/JCO.2007.13.9006

Espie, C. A., MacMahon, K. M. A., Kelly, H.-L., Broomfield, N. M., Douglas, N. J., Engleman, H. M., McKinstry, B., Morin, C. M., Walker, A., & Wilson, P. (2007). Randomized clinical effectiveness trial of nurse-administered small-group cognitive behavior therapy for persistent insomnia in general practice. *Sleep*, *30*(5), 574–584. https://doi.org/10.1093/sleep/30.5.574

Etchels, P. (2019). *Lost in a good game: Why we play video games and what they can do for us*. Icon Books.

Evans, D. M., Brion, M. J. A., Paternoster, L., Kemp, J. P., McMahon, G., Munafò, M., Whitfield, J. B., Medland, S. E., Montgomery, G. W., Timpson, N. J., st. Pourcain, B., Lawlor, D. A., Martin, N. G., Dehghan, A., Hirschhorn, J., & Davey Smith, G. (2013). Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genetics*, *9*(10), e1003919. https://doi.org/10.1371/journal.pgen.1003919

Evans, D. M., & Davey Smith, G. (2015). Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annual Review of Genomics and Human Genetics*, *16*(1), 327–350. https://doi.org/10.1146/annurev-genom-090314-050016

Even, S. (2011). *Graph Algorithms*. Cambridge University Press.

Fava, M., McCall, W. V., Krystal, A., Wessel, T., Rubens, R., Caron, J., Amato, D., & Roth, T. (2006). Eszopiclone co-administered with fluoxetine in patients with insomnia coexisting with major depressive disorder. *Biological Psychiatry*, *59*(11), 1052–1060. https://doi.org/10.1016/j.biopsych.2006.01.016

Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology*, *12*(1), 14. https://doi.org/10.1186/s12982-015-0037-4

Ferguson, C. J., Copenhaver, A., & Markey, P. (2020). Reexamining the findings of the american psychological association's 2015 task force on violent media: A Meta-Analysis. *Perspectives on Psychological Science*, *15*(6), 1423–1443. https://doi.org/10.1177/1745691620927666

Fernandez-Mendoza, J., & Vgontzas, A. N. (2013). Insomnia and its impact on physical and mental health. *Current Psychiatry Reports*, *15*(12), 418. https://doi.org/10.1007/s11920-013-0418-8

Fibiger, H. C., & Phillips, A. G. (1988). Mesocorticolimbic dopamine systems and reward. *Annals of the New York Academy of Sciences*, *537*(1 The Mesocorti), 206–215. https://doi.org/10.1111/j.1749-6632.1988.tb42107.x

Firth, J., Torous, J., Nicholas, J., Carney, R., Rosenbaum, S., & Sarris, J. (2017). Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *Journal of Affective Disorders*, *218*(January), 15–22. https://doi.org/10.1016/j.jad.2017.04.046

Fleming, T. M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S. N., Cheek, C., Lucassen, M., Lau, H. M., Pollmuller, B., & Hetrick, S. (2017). Serious games and gamification for mental health: Current

status and promising directions. *Frontiers in Psychiatry*, *7*(JAN). https://doi.org/10.3389/fpsyt.2016.00215

Foster, R., Peirson, S., Wulff, K., Winnebeck, E., Vetter, C., & Roenneberg, T. (2013). *Sleep and Circadian Rhythm Disruption in Social Jetlag and Mental Illness* (pp. 325–346). https://doi.org/10.1016/B978-0-12-396971-2.00011-7

Francisco-Aparicio, A., Gutiérrez-Vela, F. L., Isla-Montes, J. L., & Sanchez, J. L. G. (2013). *Gamification: Analysis and Application* (pp. 113–126). https://doi.org/10.1007/978-1-4471-5445-7_9

Free Ice Cream. (2017). *HiveMind2030*. https://freeicecream.co.uk/2030-hive-mind/

Free Ice Cream, & Davis, O. S. P. (2018). *Playable Data for Human Health*. http://staging.freeicecream.co.uk/?p=111

Free Ice Cream, & Overseas Development Institute (2017). *2030 Hive Mind by Free Ice Cream*. NESTA. https://www.nesta.org.uk/feature/smarter-policy-through-simulation/2030-hive-mind-by-free-ice-cream/

Freimer, N., & Sabatti, C. (2003). The human phenome project. *Nature Genetics*, *34*(1), 15–21. https://doi.org/10.1038/ng0503-15

Friberger, M. G., & Togelius, J. (2012). Generating interesting Monopoly boards from open data. *2012 IEEE Conference on Computational Intelligence and Games, CIG 2012*, 288–295. https://doi.org/10.1109/CIG.2012.6374168

Friberger, M. G., & Togelius, J. (2013). Bar chart ball, a data game. *Fdg2013.Org*. https://dspace.mah.se/handle/2043/16942%5Cnhttp://www.fdg2013.org/program/posters/poster18_togelius_friberger.pdf

Friberger, M. G., Togelius, J., Borg Cardona, A., & Ermacora, M. (2013). Data games. *Proceedings of the The Fourth Workshop on Procedural Content Generation in Games, ACM Digital Library*, 1–8.

Fullerton, T. (2018). *Game Design Workshop*. A K Peters/CRC Press. https://doi.org/10.1201/b22309

Ganesh, A., Massoulie, L., & Towsley, D. (2005). The effect of network topology on the spread of epidemics. *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, *2*, 1455–1466. https://doi.org/10.1109/INFCOM.2005.1498374

Gao, X., Meng, L.-X., Ma, K.-L., Liang, J., Wang, H., Gao, Q., & Wang, T. (2019). The bidirectional causal relationships of insomnia with five major psychiatric disorders: A Mendelian randomization study. *European Psychiatry*, *60*, 79–85. https://doi.org/10.1016/j.eurpsy.2019.05.004

Garbarino, S., Lanteri, P., Durando, P., Magnavita, N., & Sannita, W. (2016a). Co-morbidity, mortality, quality of life and the healthcare/welfare/social costs of disordered sleep: a rapid review. *International Journal of Environmental Research and Public Health*, *13*(8), 831. https://doi.org/10.3390/ijerph13080831

Garber, L. L., Hyatt, E. M., & Boya, Ü. Ö. (2017). Gender differences in learning preferences among participants of serious business games. *The International Journal of Management Education*, *15*(2), 11–29. https://doi.org/10.1016/j.ijme.2017.02.001

Genetics of Personality Consortium, de Moor, M. H. M., van den Berg, S. M., Verweij, K. J. H., Krueger, R. F., Luciano, M., Arias Vasquez, A., Matteson, L. K., Derringer, J., Esko, T., Amin, N., Gordon, S. D.,

Hansell, N. K., Hart, A. B., Seppälä, I., Huffman, J. E., Konte, B., Lahti, J., Lee, M., … Boomsma, D. I. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry*, *72*(7), 642–650. https://doi.org/10.1001/jamapsychiatry.2015.0554

Gentry, S. V., Gauthier, A., L'Estrade Ehrstrom, B., Wortley, D., Lilienthal, A., Tudor Car, L., Dauwels-Okutsu, S., Nikolaou, C. K., Zary, N., Campbell, J., & Car, J. (2019). Serious gaming and gamification education in health professions: systematic review. *Journal of Medical Internet Research*, *21*(3), e12994. https://doi.org/10.2196/12994

Geurts, J. L. A., Duke, R. D., & Vermeulen, P. A. M. (2007). Policy gaming for strategy and change. *Long Range Planning*, *40*(6), 535–558. https://doi.org/10.1016/j.lrp.2007.07.004

Gogus, A. (2012). Bloom's Taxonomy of Learning Objectives. In *Encyclopedia of the Sciences of Learning* (pp. 469–473). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_141

Goldberg, D. (2015). *The state of play: Creators and critics on video game culture*. Seven Stories Press.

Goldenberg, F., Hindmarch, I., Joyce, C. R. B., le Gal, M., Partinen, M., & Pilate, C. (1994). Zopiclone, sleep and health-related quality of life. *Human Psychopharmacology: Clinical and Experimental*, *9*(4), 245–251. https://doi.org/10.1002/hup.470090403

Gottesmann, C. (2002). GABA mechanisms and sleep. *Neuroscience*, *111*(2), 231–239. https://doi.org/10.1016/S0306-4522(02)00034-9

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass.)*, *10*(1), 37–48. http://www.ncbi.nlm.nih.gov/pubmed/9888278

Greenspoon, P.J., Saklofske, D. H. (2001). Toward an integration of subjective well-being and psychopathology. *Social Indicators Research*, *54*, 81–108. https://doi.org/https://doi.org/10.1023/A:1007219227883

Groh, F. (2012). Gamification: state of the art definition and utilization. *Proceedings of the 4th Seminar on Research Trends in Media Informatics (RTMI'12)*, 39–46. http://vts.uni-ulm.de/docs/2012/7866/vts_7866_11380.pdf

Grossman, J., & Mackenzie, F. J. (2005). The randomized controlled trial: gold standard, or merely standard? *Perspectives in Biology and Medicine*, *48*(4), 516–534. https://doi.org/10.1353/pbm.2005.0092

Guin, T. D.-L., Baker, R., Mechling, J., & Ruyle, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research*, *54*(5), 613–633. https://doi.org/10.2501/IJMR-54-5-613-633

Haack, M., & Mullington, J. M. (2005). Sustained sleep restriction reduces emotional and physical well-being. *Pain*, *119*(1–3), 56–64. https://doi.org/10.1016/j.pain.2005.09.011

Hajak, G., Clarenbach, P., Fischer, W., Haase, W., Bandelow, B., Adler, L., & Rüther, E. (1995). Effects of hypnotics on sleep quality and daytime well-being. Data from a comparative multicentre study in outpatients with insomnia. *European Psychiatry*, *10*(S3), 173s–179s. https://doi.org/10.1016/0924-9338(96)80100-3

Hajak, G., Clarenbach, P., Fischer, W., Haase, W., & Rüther, E. (1994). Zopiclone improves sleep quality and daytime well-being in insomniac patients. *International Clinical Psychopharmacology*, *9*(4), 251–262. https://doi.org/10.1097/00004850-199400940-00004

Hajak, G., Cluydts, R., Declerck, A., Estivill, S. E., Middleton, A., Sonka, K., & Unden, M. (2002). Continuous versus non-nightly use of zolpidem in chronic insomnia: results of a large-scale, double-blind, randomized, outpatient study. *International Clinical Psychopharmacology*, *17*(1), 9–17. https://doi.org/10.1097/00004850-200201000-00002

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - A literature review of empirical studies on gamification. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Harms, J., Seitz, D., Wimmer, C., Kappel, K., & Grechenig, T. (2015). Low-cost gamification of online surveys. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 109–113. https://doi.org/10.1145/2793107.2793146

Harrison, L., Reinecke, K., & Chang, R. (2015). Infographic Aesthetics. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1187–1190. https://doi.org/10.1145/2702123.2702545

Hartwig, F. P., Davey Smith, G. & Bowden, J. (2017). Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption. *Int. J. Epidemiol*, *46*, 1985–1998.

Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C., & Smith, G. D. (2016). Best (but oft-forgotten) practices: The design, analysis, and interpretation of Mendelian randomization studies. *American Journal of Clinical Nutrition*, *103*(4), 965–978. https://doi.org/10.3945/ajcn.115.118216

Hemani, G. (2022). Constructing deconvolved causal graphs from GWAS summary data. *Working Paper*.

Hemani, G., Bowden, J., & Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, *27*(R2), R195–R208. https://doi.org/10.1093/hmg/ddy163

Hemani, G., Bowden, J., Haycock, P., Zheng, J., Davis, O., Flach, P., Gaunt, T., & Smith, G. D. (2017). Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *BioRxiv*, *10*, 173682. https://doi.org/10.1101/173682

Hemani, G., Tilling, K., & Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genetics*, *13*(11), e1007081. https://doi.org/10.1371/journal.pgen.1007081

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., & Haycock, P. C. (2018a). The MR-base platform supports systematic causal inference across the human phenome. *ELife*, *7*, 1–29. https://doi.org/10.7554/eLife.34408

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Hindmarch, I. (1995). Effects of zopiclone on quality of life in insomnia. *European Psychiatry*, *10*(S3), 91s–94s. https://doi.org/10.1016/0924-9338(96)80087-3

Holland, J. H. (1992). Complex adaptive systems. *Daedalus*, *121*(1), 17–30.

Hom, M. A., Chu, C., Rogers, M. L., & Joiner, T. E. (2020). A meta-analysis of the relationship between sleep problems and loneliness. *Clinical Psychological Science*, *8*(5), 799–824. https://doi.org/10.1177/2167702620922969

Howell, A. E., Robinson, J. W., Wootton, R. E., McAleenan, A., Tsavachidis, S., Ostrom, Q. T., Bondy, M., Armstrong, G., Relton, C., Haycock, P., Martin, R. M., Zheng, J., & Kurian, K. M. (2020). Testing for causality between systematically identified risk factors and glioma: a Mendelian randomization study. *BMC Cancer*, *20*(1), 508. https://doi.org/10.1186/s12885-020-06967-2

Howell, R. T., Kern, M. L., & Lyubomirsky, S. (2007). Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review*, *1*(1), 83–136. https://doi.org/10.1080/17437190701492486

Hunicke, R., Leblanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *AAAI Workshop - Technical Report*, *WS-04-04*, 1–5.

Iasiello, M., & Joep van, A. (2020). Mental health and/or mental illness: A scoping review of the evidence and implications of the dual-continua model of mental health. *Evidence Base: A Journal of Evidence Reviews in Key Policy Areas*, *1*.

Ioannidis, J. P. A. (2016). Exposure-wide epidemiology: revisiting Bradford Hill. *Statistics in Medicine*, *35*(11), 1749–1762. https://doi.org/10.1002/sim.6825

ISFE. (2021). Europe Video Game Industry Key Facts 2021. *Interactive Software Federation of Europe Reports*, 1–20. https://www.isfe.eu/data-key-facts/key-facts-about-europe-s-video-games-sector/

Jackowska, M., Brown, J., Ronaldson, A., & Steptoe, A. (2016). The impact of a brief gratitude intervention on subjective well-being, biology and sleep. *Journal of Health Psychology*, *21*(10), 2207–2217. https://doi.org/10.1177/1359105315572455

Jaehne, A., Unbehaun, T., Feige, B., Lutz, U. C., Batra, A., & Riemann, D. (2012). How smoking affects sleep: A polysomnographical analysis. *Sleep Medicine*, *13*(10), 1286–1292. https://doi.org/10.1016/j.sleep.2012.06.026

Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., de Leeuw, C. A., Benjamins, J. S., Muñoz-Manchado, A. B., Nagel, M., Savage, J. E., Tiemeier, H., White, T., Tung, J. Y., Hinds, D. A., Vacic, V., Wang, X., Sullivan, P. F., van der Sluis, S., … Posthuma, D. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature Genetics*, *51*(3), 394–403. https://doi.org/10.1038/s41588-018-0333-3

Jaspers, M., Steen, T., Bos, C., & Geenen, M. (2004). The think aloud method: a guide to user interface design. *International Journal of Medical Informatics*, *73*(11–12), 781–795. https://doi.org/10.1016/j.ijmedinf.2004.08.003

Jean-Louis, G., Kripke, D. F., & Ancoli-Israel, S. (2000). Sleep and quality of well-being. *Sleep*, *23*(8), 1115–1121. http://www.ncbi.nlm.nih.gov/pubmed/11145326

Jermann, F., Perroud, N., Favre, S., Aubry, J.-M., & Richard-Lepouriel, H. (2022). Quality of life and subjective sleep-related measures in bipolar disorder and major depressive disorder. *Quality of Life Research*, *31*(1), 117–124. https://doi.org/10.1007/s11136-021-02929-8

Jiang, J. (2020). “I Never Know What to Expect”: Aleatory Identity Play in Fortnite and Its Implications for Multimodal Composition. *Computers and Composition*, *55*, 102550. https://doi.org/10.1016/j.compcom.2020.102550

Johnson, D., Deterding, S., Kuhn, K. A., Staneva, A., Stoyanov, S., & Hides, L. (2016). Gamification for health and wellbeing: A systematic review of the literature. *Internet Interventions*, *6*, 89–106. https://doi.org/10.1016/j.invent.2016.10.002

Johnson, E. O., Roehrs, T., Roth, T., & Breslau, N. (1998). Epidemiology of alcohol and medication as aids to sleep in early adulthood. *Sleep*, *21*(2), 178–186. https://doi.org/10.1093/sleep/21.2.178

Joldersma, C., & Geurts, Jac. L. A. (1998). Simulation/gaming for policy development and organizational change. *Simulation & Gaming*, *29*(4), 391–399. https://doi.org/10.1177/104687819802900402

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*(4), 63–85. https://doi.org/10.1007/BF02300500

Kalueff, A. v., & Nutt, D. J. (2007). Role of GABA in anxiety and depression. *Depression and Anxiety*, *24*(7), 495–517. https://doi.org/10.1002/da.20262

Kamat, M. A., Blackshaw, J. A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A. S., & Staley, J. R. (2019). PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics*, *35*(22), 4851–4853. https://doi.org/10.1093/bioinformatics/btz469

Kanode, C. M., & Haddad, H. M. (2009). Software engineering challenges in game development. *2009 Sixth International Conference on Information Technology: New Generations*, 260–265. https://doi.org/10.1109/ITNG.2009.74

Kapp, K. (2012). *The Gamification of Learning and Instruction. Game-Based Methods and Strategies for Training and Education*. Pfeiffer.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial. *Journal of Clinical Epidemiology*, *54*(6), 541–549. https://doi.org/10.1016/S0895-4356(00)00347-4

Kara, N. (2021). A systematic review of the use of serious games in science education. *Contemporary Educational Technology*, *13*(2), ep295. https://doi.org/10.30935/cedtech/9608

Katikireddi, S. V., Green, M. J., Taylor, A. E., Davey Smith, G., & Munafò, M. R. (2018). Assessing causal relationships using genetic proxies for exposures: an introduction to Mendelian randomization. *Addiction*, *113*(4), 764–774. https://doi.org/10.1111/add.14038

Kaviya, K. (2014). The Blue Brain. *IJAICT*, *1*(3). https://doi.org/01.0401/ijaict.2014.03.06

Kawrykow, A., Roumanis, G., Kam, A., Kwak, D., Leung, C., Wu, C., Zarour, E., Sarmenta, L., Blanchette, M., & Waldispühl, J. (2012). Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS ONE*, *7*(3), e31362. https://doi.org/10.1371/journal.pone.0031362

Keeling, M. J., & Eames, K. T. D. (2005). Networks and epidemic models. *Journal of The Royal Society Interface*, *2*(4), 295–307. https://doi.org/10.1098/rsif.2005.0051

Kinmonth, A. L., Woodcock, A., Griffin, S., Spiegal, N., & Campbell, M. J. (1998). Randomised controlled trial of patient centred care of diabetes in general practice: impact on current wellbeing and future disease risk. *BMJ*, *317*(7167), 1202–1208. https://doi.org/10.1136/bmj.317.7167.1202

Klabbers, J. H. G. (2003). Gaming and simulation: principles of a science of design. *Simulation and Gaming*, *34*(4), 569–591. https://doi.org/10.1177/1046878103258205

Knutson, K. L., Spiegel, K., Penev, P., & van Cauter, E. (2007). The metabolic consequences of sleep deprivation. *Sleep Medicine Reviews*, *11*(3), 163–178. https://doi.org/10.1016/j.smrv.2007.01.002

Krieger, N., & Smith, G. D. (2016). The tale wagged by the DAG: Broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, *45*(6), 1787–1808. https://doi.org/10.1093/ije/dyw114

Kristensson, P., Matthing, J., & Johansson, N. (2008). Key strategies for the successful involvement of customers in the co-creation of new technology-based services. *International Journal of Service Industry Management*, *19*(4), 474–491. https://doi.org/10.1108/09564230810891914

Krystal, A. D. (2007). Treating the health, quality of life, and functional impairments in insomnia. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *3*(1), 63–72. http://www.ncbi.nlm.nih.gov/pubmed/17557457

Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plots--rational approach to visualizing networks. *Briefings in Bioinformatics*, *13*(5), 627–644. https://doi.org/10.1093/bib/bbr069

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Kunz, W., & Rittel, H. W. J. (1972). Information science: On the structure of its problems. *Information Storage and Retrieval*, *8*(2), 95–98. https://doi.org/10.1016/0020-0271(72)90011-3

Kwak, D., Kam, A., Becerra, D., Zhou, Q., Hops, A., Zarour, E., Kam, A., Sarmenta, L., Blanchette, M., & Waldispühl, J. (2013). Open-Phylo: a customizable crowd-computing platform for multiple sequence alignment. *Genome Biology*, *14*(10), R116. https://doi.org/10.1186/gb-2013-14-10-r116

Kyle, S. D., Morgan, K., & Espie, C. A. (2010). Insomnia and health-related quality of life. *Sleep Medicine Reviews*, *14*(1), 69–82. https://doi.org/10.1016/j.smrv.2009.07.004

Lai, C. C. W. (2018). The mediating role of sleep quality in the relationship between personality and subjective well-being. *SAGE Open*, *8*(2), 215824401877313. https://doi.org/10.1177/2158244018773139

Lancel, M., Stroebe, M., & Eisma, M. C. (2020). Sleep disturbances in bereavement: A systematic review. *Sleep Medicine Reviews*, *53*, 101331. https://doi.org/10.1016/j.smrv.2020.101331

Lane, N. D., Lin, M., Mohammod, M., Yang, X., Lu, H., Cardone, G., Ali, S., Doryab, A., Berke, E., Campbell, A. T., & Choudhury, T. (2014). Bewell: sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Networks and Applications*, *19*(3), 345–359. https://doi.org/10.1007/s11036-013-0484-5

Larsson, D., & Goldberg, L. (2015). *The state of play : Sixteen Voices on Video Games*. Seven Stories Press.

Lau, E. Y. Y., Hui, C. H., Lam, J., & Cheung, S.-F. (2017). Sleep and optimism: A longitudinal study of bidirectional causal relationship and its mediating and moderating variables in a Chinese student sample. *Chronobiology International*, *34*(3), 360–372. https://doi.org/10.1080/07420528.2016.1276071

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., & Rathouz, P. J. (2008). Self-reported and measured sleep duration. *Epidemiology*, *19*(6), 838–845. https://doi.org/10.1097/EDE.0b013e318187a7b0

Lawlor, D. A. (2016). Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology*, *45*(3), 908–915. https://doi.org/10.1093/ije/dyw127

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*(8), 1133–1163. https://doi.org/10.1002/sim.3034

LeBlanc, M., Zuber, V., Thompson, W. K., Andreassen, O. A., Frigessi, A., & Andreassen, B. K. (2018). A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC Genomics*, *19*(1), 494. https://doi.org/10.1186/s12864-018-4859-7

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Leppek, K., Byeon, G. W., Kladwang, W., Wayment-Steele, H. K., Kerr, C. H., Xu, A. F., Kim, D. S., Topkar, V. v., Choe, C., Rothschild, D., Tiu, G. C., Wellington-Oguri, R., Fujii, K., Sharma, E., Watkins, A. M., Nicol, J. J., Romano, J., Tunguz, B., Diaz, F., … Das, R. (2022). Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nature Communications*, *13*(1), 1536. https://doi.org/10.1038/s41467-022-28776-w

Lewinsohn, P. M., Redner, J., & John R., S. (1991). The relationship between life satisfaction and psychosocial variables: New perspectives. *Subjective Well-Being: An Interdisciplinary Perspective*, *21*, 141–169.

Lewsey, F. (2020). *Cambridge game 'pre-bunks' coronavirus conspiracies*. https://www.cam.ac.uk/stories/goviral

Li, F., Fisher, K. J., Harmer, P., Irbe, D., Tearse, R. G., & Weimer, C. (2004). Tai chi and self-rated quality of sleep and daytime sleepiness in older adults: a randomized controlled trial. *Journal of the American Geriatrics Society*, *52*(6), 892–900. https://doi.org/10.1111/j.1532-5415.2004.52255.x

Liapis, A., Yannakakis, G. N., Nelson, M. J., Preuss, M., & Bidarra, R. (2019). Orchestrating game generation. *IEEE Transactions on Games*, *11*(1), 48–68. https://doi.org/10.1109/TG.2018.2870876

Lima, M. (2013). *Visual Complexity: Mapping Patterns of Information*.

Lins, L., & Carvalho, F. M. (2016). SF-36 total score as a single measure of health-related quality of life: Scoping review. *SAGE Open Medicine*, *4*, 205031211667172. https://doi.org/10.1177/2050312116671725

Liu, X., Li, R., Lanza, S. T., Vasilenko, S. A., & Piper, M. (2013). Understanding the role of cessation fatigue in the smoking cessation process. *Drug and Alcohol Dependence*, *133*(2), 548–555. https://doi.org/10.1016/j.drugalcdep.2013.07.025

Locke, E. A., & Latham, G. P. (1994). Goal setting theory. *Motivation: Theory and Research*, *13*(29).

Lofgren, E. T., & Feff, N. H. (2007). Personal View The untapped potential of virtual game worlds to shed light on real world epidemics. *Infection.Thelancet.Com*, *7*(September).

Looyestyn, J., Kernot, J., Boshoff, K., Ryan, J., Edney, S., & Maher, C. (2017). Does gamification increase engagement with online programs? A systematic review. *PLOS ONE*, *12*(3), e0173403. https://doi.org/10.1371/journal.pone.0173403

Lucero, A., Holopainen, J., Ollila, E., Suomela, R., & Karapanos, E. (2013). The Playful Experiences (PLEX) framework as a guide for expert evaluation. *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces, DPPI 2013*, 221–230. https://doi.org/10.1145/2513506.2513530

Lumsden, J., Skinner, A., Coyle, D., Lawrence, N., & Munafo, M. (2017). Attrition from web-based cognitive testing: a repeated measures comparison of gamification techniques. *Journal of Medical Internet Research*, *19*(11), e395. https://doi.org/10.2196/jmir.8473

Lumsden, J., Skinner, A., Woods, A. T., Lawrence, N. S., & Munafò, M. (2016a). The effects of gamelike features and test location on cognitive test performance and participant enjoyment. *PeerJ*, *4*, e2184. https://doi.org/10.7717/peerj.2184

Luzzatto, L. (2012). SICKLE CELL ANAEMIA AND MALARIA. *Mediterranean Journal of Hematology and Infectious Diseases*, *4*(1), e2012065. https://doi.org/10.4084/mjhid.2012.065

Lyubomirsky, S., & Lepper, H. S. (1995). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*(2), 137–155. https://doi.org/https://doi.org/10.1023/A:1006824100041

MacKlin, C., Edwards, M., Wargaski, J., & Yang Li, K. (2009). Dataplay: Mapping game mechanics to traditional data visualization. *Breaking New Ground: Innovation in Games, Play, Practice and Theory - Proceedings of DiGRA 2009*.

Magnello, M. E. (2012). Victorian statistical graphics and the iconography of Florence Nightingale's polar area graph. *BSHM Bulletin: Journal of the British Society for the History of Mathematics*, *27*(1), 13–37. https://doi.org/10.1080/17498430.2012.618102

Magnusson, A., & Boivin, D. (2003). Seasonal affective disorder: an overview. *Chronobiology International*, *20*(2), 189–207. https://doi.org/10.1081/CBI-120019310

Malliarakis, C., Maya, S., & Xinogalos, S. (2014). Designing educational games for computer programming: A holistic framework. *The Electronic Journal of E-Learning*, *12*(3), 281–181. https://files.eric.ed.gov/fulltext/EJ1035677.pdf

Mancini, M. E., Soar, J., Bhanji, F., Billi, J. E., Dennett, J., Finn, J., Ma, M. H.-M., Perkins, G. D., Rodgers, D. L., Hazinski, M. F., Jacobs, I., Morley, P. T., Aufderheide, T. P., Atkins, D. L., Barelli, A., Baubin, M., Bernhard, M., Botha, M., Brennan, N., … Yuen, T. (2010). Part 12: Education, Implementation, and Teams. *Circulation*, *122*(16_suppl_2). https://doi.org/10.1161/CIRCULATIONAHA.110.971143

Marc Prensky. (2001). Fun, play and games: what makes games engaging. In *Digital Game-Based Learning* . McGraw-Hill.

Marczewski, A. (2018). Player and user types hexad. In *Even Ninja Monkeys Like to Play: Gamification, Game Thinking and Motivational Design* (1st ed., Issue October, pp. 1–25). CreateSpace Independent Publishing Platform. https://doi.org/1514745666

Marks, N., & Shah, H. (2004). A well-being manifesto for a flourishing society. *Journal of Public Mental Health*, *3*(4), 9–15. https://doi.org/10.1108/17465729200400023

Marlatt, R. (2020). Capitalizing on the craze of fortnite: toward a conceptual framework for understanding how gamers construct communities of practice. *Journal of Education*, *200*(1), 3–11. https://doi.org/10.1177/0022057419864531

Martin, J. L., Song, Y., Hughes, J., Jouldjian, S., Dzierzewski, J. M., Fung, C. H., Rodriguez Tapia, J. C., Mitchell, M. N., & Alessi, C. A. (2017). A four-session sleep intervention program improves sleep for older adult day health care participants: results of a randomized controlled trial. *Sleep*, *40*(8). https://doi.org/10.1093/sleep/zsx079

Mathur, M. B., & VanderWeele, T. J. (2019). Finding common ground in meta-analysis "wars" on violent video games. *Perspectives on Psychological Science*, *14*(4), 705–708. https://doi.org/10.1177/1745691619850104

Mavandadi et al. (2012a). Crowd-sourced BioGames: managing the big data problem for next-generation lab-on-a-chip platforms. *Lab on a Chip*, *12*(20), 4102. https://doi.org/10.1039/c2lc40614d

Mavandadi et al. (2012b). BioGames: a platform for crowd-sourced biomedical image analysis and telediagnosis. *Games for Health Journal*, *1*(5), 373–376. https://doi.org/10.1089/g4h.2012.0054

Mavandadi et al. (2012c). A mathematical framework for combining decisions of multiple experts toward accurate and remote diagnosis of malaria using tele-microscopy. *PLoS ONE*, *7*(10), e46192. https://doi.org/10.1371/journal.pone.0046192

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., Padmanabhan, S., Nielsen, K., & Ozcan, A. (2012). Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. *PLoS ONE*, *7*(5), e37245. https://doi.org/10.1371/journal.pone.0037245

McKee, A. (2016). What Is Fun? In *FUN!* (pp. 29–40). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-49179-4_3

McLeod, K. S. (2000). Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine*, *50*(7–8), 923–935. https://doi.org/10.1016/S0277-9536(99)00345-7

Meadows, D. H. (2008). *Systems thinking: A primer*. Chelsea green publishing.

Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. N. (2013). Do points, levels and leaderboards harm intrinsic motivation? *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, 66–73. https://doi.org/10.1145/2583008.2583017

Meng, X.-H., Chen, X.-D., Greenbaum, J., Zeng, Q., You, S.-L., Xiao, H.-M., Tan, L.-J., & Deng, H.-W. (2018). Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone*, *113*, 41–48. https://doi.org/10.1016/j.bone.2018.05.012

Michael, D., & Chen, S. (2006). *Serious games: Games that educate, train, and inform*. Thomson Course Technology.

Milstein, B., Homer, J., Briss, P., Burton, D., & Pechacek, T. (2011). Why behavioral and environmental interventions are needed to improve health at lower cost. *Health Affairs*, *30*(5), 823–832. https://doi.org/10.1377/hlthaff.2010.1116

Milstein, B., Homer, J., & Hirsch, G. (2009). The "healthbound" policy simulation game: an adventure in us health reform. *International System Dynamics Conference*.

Milstein, B., Homer, J., & Hirsch, G. (2010). Analyzing national health reform strategies with a dynamic simulation model. *American Journal of Public Health*, *100*(5), 811–819. https://doi.org/10.2105/AJPH.2009.174490

Minelli, C., del Greco M., F., van der Plaat, D. A., Bowden, J., Sheehan, N. A., & Thompson, J. (2021). The use of two-sample methods for Mendelian randomization analyses on single large datasets. *International Journal of Epidemiology*, *50*(5), 1651–1659. https://doi.org/10.1093/ije/dyab084

Mitchell, S. J. (2010). *Positive psychology and sleep: the influence of an internet-based exercise*.

Mitgutsch, K., & Alvarado, N. (2012). Purposeful by design?: a serious game design assessment framework. *Proceedings of the International Conference on the Foundations of Digital Games - FDG '12*, 121. https://doi.org/10.1145/2282338.2282364

Mojang Studios. (2011). *Minecraft*. https://www.minecraft.net/en-us

Mora, A., Riera, D., Gonzalez, C., & Arnedo-Moreno, J. (2015). A Literature Review of Gamification Design Frameworks. *VS-Games 2015 - 7th International Conference on Games and Virtual Worlds for Serious Applications*, *September*. https://doi.org/10.1109/VS-GAMES.2015.7295760

Moreno, J. L. (1933). Psychological and social organization of groups in the community. *Proceedings & Addresses. American Association on Mental Deficiency*, *38*, 224–242.

Moreno-Stokoe, C. M., Elias, B., Felix, B., Harmaala, L., Rudel, S., Frankel, S., Martinovici, A., & Bar-Yam, Y. (2021). *OutBreak: A five minute game to return to normality*. https://outbreak.endcoronavirus.org

Morin, C. M., Koetter, U., Bastien, C., Ware, J. C., & Wooten, V. (2005). Valerian-hops combination and diphenhydramine for treating insomnia: a randomized placebo-controlled clinical trial. *Sleep*, *28*(11), 1465–1471. https://doi.org/10.1093/sleep/28.11.1465

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. v, Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., … DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990. https://doi.org/10.1038/ng.2383

Munafò, M. R., Higgins, J. P. T., & Davey Smith, G. (2021). Triangulating Evidence through the Inclusion of Genetically Informed Designs. *Cold Spring Harbor Perspectives in Medicine*, *11*(8), a040659. https://doi.org/10.1101/cshperspect.a040659

Munafo, M. R., Pfeiffer, T., Altmejd, A., Heikensten, E., Almenberg, J., Bird, A., Chen, Y., Wilson, B., Johannesson, M., & Dreber, A. (2015). Using prediction markets to forecast research evaluations. *Royal Society Open Science*, *2*(10), 150287. https://doi.org/10.1098/rsos.150287

Ndemic Creations. (2012). *Plague Inc*. https://www.ndemiccreations.com/en/22-plague-inc

Ndemic Creations, & World Health Organisation. (2021). *Plague Inc: The Cure*.

Neurath, O., & Kleinschmidt, H. E. (1939). Health education by isotype. *Am J Public Health Nations Health.*, *29*(5), 548–549.

Neurath, O., & Neurath, M. (1955). *Leprosy*. Isotype collection, University of Reading.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., Armasu, S. M., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., … Farrall, M. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*(10), 1121–1130. https://doi.org/10.1038/ng.3396

Nowowiejska, J., Baran, A., & Flisiak, I. (2021). Mutual relationship between sleep disorders, quality of life and psychosocial aspects in patients with psoriasis. *Frontiers in Psychiatry*, *12*. https://doi.org/10.3389/fpsyt.2021.674460

Nutt, D., Wilson, S., & Paterson, L. (2008). Sleep disorders as core symptoms of depression. *Dialogues in Clinical Neuroscience*, *10*(3), 329–336. https://doi.org/10.31887/DCNS.2008.10.3/dnutt

Ohayon, M. M. (2011). Epidemiological overview of sleep disorders in the general population. *Sleep Medicine Research*, *2*(1), 1–9. https://doi.org/10.17241/smr.2011.2.1.1

Okbay, A., Baselmans, B. M. L., de Neve, J.-E., Turley, P., Nivard, M. G., Fontana, M. A., Meddens, S. F. W., Linnér, R. K., Rietveld, C. A., Derringer, J., Gratten, J., Lee, J. J., Liu, J. Z., de Vlaming, R., Ahluwalia, T. S., Buchwald, J., Cavadino, A., Frazier-Wood, A. C., Furlotte, N. A., … Cesarini, D. (2016a). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, *48*(6), 624–633. https://doi.org/10.1038/ng.3552

O'Kearney, R., & Pech, M. (2014). General and sleep-specific worry in insomnia. *Sleep and Biological Rhythms*, *12*(3), 212–215. https://doi.org/10.1111/sbr.12054

O'Loughlin, J., Casanova, F., Jones, S. E., Hagenaars, S. P., Beaumont, R. N., Freathy, R. M., Watkins, E. R., Vetter, C., Rutter, M. K., Cain, S. W., Phillips, A. J. K., Windred, D. P., Wood, A. R., Weedon, M. N., & Tyrrell, J. (2021). Using Mendelian Randomisation methods to understand whether diurnal preference is causally related to mental health. *Molecular Psychiatry*, *26*(11), 6305–6316. https://doi.org/10.1038/s41380-021-01157-3

Omvik, S., Sivertsen, B., Pallesen, S., Bjorvatn, B., Havik, O. E., & Nordhus, I. H. (2008). Daytime functioning in older patients suffering from chronic insomnia: Treatment outcome in a randomized controlled trial comparing CBT with Zopiclone. *Behaviour Research and Therapy*, *46*(5), 623–641. https://doi.org/10.1016/j.brat.2008.02.013

Ong, A. D. (2010). Pathways linking positive emotion and health in later life. *Current Directions in Psychological Science*, *19*(6), 358–362. https://doi.org/10.1177/0963721410388805

Ozcan, A. (2014). Educational Games for Malaria Diagnosis. *Science Translational Medicine*, *6*(233). https://doi.org/10.1126/scitranslmed.3009172

Pahlavan, A., & Ghasem, A. (2022). Effectiveness of positive psychology interventions on death anxiety and sleep quality of female patients with multiple sclerosis (MS). *Research in Clinical Psychology and Counselling*, *9*(2), 28–42.

Pajitnov, A. (1984). *Tetris*. https://tetris.com/play-tetris

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Computers & Education*, *52*(1), 1–12. https://doi.org/10.1016/j.compedu.2008.06.004

Parackal, M., & Parackal, S. (2017). Implication of alcohol consumption on aggregate wellbeing. *Perspectives in Public Health*, *137*(4), 220–226. https://doi.org/10.1177/1757913916669538

Paradox Interactive. (2013). *Europa Universalis IV*. https://eu4.paradoxwikis.com/Europa_Universalis_4_Wiki

Paradox Interactive. (2016). *Hearts of Iron IV*. https://www.paradoxinteractive.com/games/hearts-of-iron-iv/about

Park, S., & Kim, S. (2018). Patterns among 754 gamification cases: Content Analysis for Gamification Development. *JMIR Serious Games*, *6*(4), e11336. https://doi.org/10.2196/11336

Pasman, J. A., Smit, D. J. A., Kingma, L., Vink, J. M., Treur, J. L., & Verweij, K. J. H. (2020). Causal relationships between substance use and insomnia. *Drug and Alcohol Dependence*, *214*, 108151. https://doi.org/10.1016/j.drugalcdep.2020.108151

Pasquier, P., Mérat, S., Malgras, B., Petit, L., Queran, X., Bay, C., Boutonnet, M., Jault, P., Ausset, S., Auroy, Y., Perez, J. P., Tesnière, A., Pons, F., & Mignon, A. (2016). A serious game for massive training and assessment of french soldiers involved in forward combat casualty care (3d-sc1): Development and deployment. *JMIR Serious Games*, *4*(1), 1–10. https://doi.org/10.2196/games.5340

Payne, J. D., & Nadel, L. (2004). Sleep, dreams, and memory consolidation: The role of the stress hormone cortisol. *Learning & Memory*, *11*(6), 671–678. https://doi.org/10.1101/lm.77104

Pearl, J. (1982). Reverend bayes on inference engines: a distributed hierarchical approach. *Proceedings of the Second AAAI Conference on Artificial Intelligence*, 133–136.

Peplow, M. (2016). Citizen science lures gamers into Sweden's Human Protein Atlas. *Nature Biotechnology*, *34*(5), 452–453. https://doi.org/10.1038/nbt0516-452c

Perach, R., Allen, C. K., Kapantai, I., Madrid-Valero, J. J., Miles, E., Charlton, R. A., & Gregory, A. M. (2019). The psychological wellbeing outcomes of nonpharmacological interventions for older persons with insomnia symptoms: A systematic review and meta-analysis. *Sleep Medicine Reviews*, *43*, 1–13. https://doi.org/10.1016/j.smrv.2018.09.003

Pierce, B. L., Ahsan, H., & VanderWeele, T. J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology*, *40*(3), 740–752. https://doi.org/10.1093/ije/dyq151

Positech. (2013). *Democracy 3*. https://www.positech.co.uk/democracy3/

Preist, C., Massung, E., & Coyle, D. (2014). Competing or aiming to be average? Normification as a means of engaging digital volunteers. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 1222–1233. https://doi.org/10.1145/2531602.2531615

Public Health England. (2020). *Wellbeing and mental health: Applying All Our Health*. https://www.gov.uk/government/publications/wellbeing-in-mental-health-applying-all-our-health/wellbeing-in-mental-health-applying-all-our-health

Qin, J., Chui, Y. P., Pang, W. M., Choi, K. S., & Heng, P. A. (2010). Learning blood management in orthopedic surgery through gameplay. *IEEE Computer Graphics and Applications*, *30*(2), 45–57. https://doi.org/10.1109/MCG.2009.83

Rangan, R., Watkins, A. M., Chacon, J., Kretsch, R., Kladwang, W., Zheludev, I. N., Townley, J., Rynge, M., Thain, G., & Das, R. (2021). *De novo* 3D models of SARS-CoV-2 RNA elements from consensus experimental secondary structures. *Nucleic Acids Research*, *49*(6), 3092–3108. https://doi.org/10.1093/nar/gkab119

Redwine, S., & Riddle, W. (1985). Software technology maturation. *Proceedings of the 8th International Conference on Software Engineering*, *ICSE '85*.

Reimer, M. A., & Flemons, W. W. (2003). Quality of life in sleep disorders. *Sleep Medicine Reviews*, *7*(4), 335–349. https://doi.org/10.1053/smrv.2001.0220

Relton, C. L., & Davey Smith, G. (2012). Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology*, *41*(1), 161–176. https://doi.org/10.1093/ije/dyr233

Ricciardi, F., de Paolis, L. T., Huai, Y., Zhang, W., Chen, Z., Zhao, F., Wang, W., Dang, K., Xue, K., Gao, Y., Jiang, S., Miao, Z., Li, M., Hao, Q., Chen, C., & Qian, A. (2014). A comprehensive review of serious games in health professions. *International Journal of Computer Games Technology*, *2014*. https://doi.org/10.1155/2014/787968

Richmond, R. C., & Davey Smith, G. (2019). Commentary: Orienting causal relationships between two phenotypes using bidirectional Mendelian randomization. *International Journal of Epidemiology*, *48*(3), 907–911. https://doi.org/10.1093/ije/dyz149

Rittel, H., & M. Weber. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*.

Robson, D. (2010). Mental health and smoking: effects on wellbeing. *British Journal of Wellbeing*, *1*(9), 22–26. https://doi.org/10.12968/bjow.2010.1.9.22

Rogers, P. J. (2007). Caffeine, mood and mental performance in everyday life. *Nutrition Bulletin*, *32*(s1), 84–89. https://doi.org/10.1111/j.1467-3010.2007.00607.x

Roungas, B., & Dalpiaz, F. (2016). *A Model-Driven Framework for Educational Game Design* (pp. 1–11). https://doi.org/10.1007/978-3-319-40216-1_1

Røysamb, E., & Nes, R. B. (2018). *The genetics of wellbeing*. UT: DEF Publishers. https://nobascholar.com/chapters/72/download.pdf

Rundo, J. V., & Downey, R. (2019). *Polysomnography* (pp. 381–392). https://doi.org/10.1016/B978-0-444-64032-1.00025-4

Saddichha, S. (2010). Diagnosis and treatment of chronic insomnia. *Annals of Indian Academy of Neurology*, *13*(2), 94. https://doi.org/10.4103/0972-2327.64628

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

Sailer, M., & Homner, L. (2020). The gamification of learning: a meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Salen, K., & Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals*. The MIT Press.

Salzmann, A., Chaturvedi, N., & Garfield, V. (2021). *The relationship between cognitive function and sleep duration: a Mendelian randomisation study*. https://doi.org/https://doi.org/10.1101/2020.09.08.20190611

Sanderson, E., Davey Smith, G., Windmeijer, F., & Bowden, J. (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, *48*(3), 713–727. https://doi.org/10.1093/ije/dyy262

Sardi, L., Idri, A., & Fernández-Alemán, J. L. (2017). A systematic review of gamification in e-Health. *Journal of Biomedical Informatics*, *71*, 31–48. https://doi.org/10.1016/j.jbi.2017.05.011

Sasupilli, M., Bokil, P., & Punekar, R. M. (2019). *Game Design Frameworks and Evaluating Techniques for Educational Games: A Review* (pp. 277–286). https://doi.org/10.1007/978-981-13-5974-3_24

Sateia, M. J. (2014). International classification of sleep disorders-third edition. *Chest*, *146*(5), 1387–1394. https://doi.org/10.1378/chest.14-0970

Scharf, M., Erman, M., Rosenberg, R., Seiden, D., McCall, W. V., Amato, D., & Wessel, T. C. (2005). A 2-week efficacy and safety study of eszopiclone in elderly patients with primary insomnia. *Sleep*, *28*(6), 720–727. https://doi.org/10.1093/sleep/28.6.720

Schreier, J. (2017). *Blood, Sweat, and Pixels: The Triumphant, Turbulent Stories Behind How Video Games Are Made*. HarperPB.

Schreier, J. (2019). How bioware's anthem went wrong. *Kotaku*.

Schrier, K. (2016). *Knowledge games: How playing games can solve problems, create insight, and make change*. JHU Press.

Schunk, D. H., Meece, J., & Pintrick, P. (2012). *Motivation in Education: Theory, Research, and Applications*. Pearson.

Sedgwick, P. (2012). Multiple significance tests: the Bonferroni correction. *BMJ*, *344*(jan25 4), e509–e509. https://doi.org/10.1136/bmj.e509

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115.2*(163).

Sella, E., Miola, L., Toffalini, E., & Borella, E. (2021). The relationship between sleep quality and quality of life in aging: a systematic review and meta-analysis. *Health Psychology Review*, 1–23. https://doi.org/10.1080/17437199.2021.1974309

Shaw, M. (2001). The coming-of-age of software architecture research. *ICSE '01: Proceedings of the 23rd International Conference on Software Engineering*. https://doi.org/doi/10.5555/381473.381549

Shneiderman, B. (2003). The eyes have it: a task by data type taxonomy for information visualizations. In *The Craft of Information Visualization* (pp. 364–371). Elsevier. https://doi.org/10.1016/B978-155860915-0/50046-9

Sicart, M. (2008). Defining game mechanics. *The International Journal of Computer Game Research*, *8*(2).

Simonofski, A., Zuiderwijk, A., Clarinval, A., & Hammedi, W. (2022). Tailoring open government data portals for lay citizens: A gamification theory approach. *International Journal of Information Management*, *65*(July 2021), 102511. https://doi.org/10.1016/j.ijinfomgt.2022.102511

Singh, F. A. M. B. J. W. (2017). Open-Phylo: a customizable crowd-computing platform for multiple sequence alignment. *In Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2017)*, 177–186.

Skiena, S. S. (2012). Sorting and searching. In *The Algorithm Design Manual* (pp. 103–144). Springer London. https://doi.org/10.1007/978-1-84800-070-4_4

Smith, G. D., & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), 89–98. https://doi.org/10.1093/hmg/ddu328

Snel, J., & Lorist, M. M. (2011). *Effects of caffeine on sleep and cognition* (pp. 105–117). https://doi.org/10.1016/B978-0-444-53817-8.00006-2

Soeffing, J. P., Lichstein, K. L., Nau, S. D., McCrae, C. S., Wilson, N. M., Aguillard, R. N., Lester, K. W., & Bush, A. J. (2008). Psychological treatment of insomnia in hypnotic-dependant older adults. *Sleep Medicine*, *9*(2), 165–171. https://doi.org/10.1016/j.sleep.2007.02.009

Song, J., Moreno-Stokoe, C., Grey, C., & Alisha, D. (2022). Health of individuals with lived experience of homelessness in Wales, during the COVID-19 pandemic. *Public Health Wales Publications*.

Song, L., Li, H., Wang, J., Xie, J., Chen, G., Liang, T., Wang, Y., Ye, L., Wang, X., Kuang, X., Ren, M., Ye, J., Tang, Y., Ji, K., Liao, W., & Zhang, X. (2022). Educational attainment could be a protective factor against obstructive sleep apnea: a study based on Mendelian randomization. *Journal of Thoracic Disease*, *14*(1), 210–215. https://doi.org/10.21037/jtd-21-945

Stallman, H. M., & Kohler, M. (2016). Prevalence of sleepwalking: a systematic review and meta-analysis. *PLOS ONE*, *11*(11), e0164769. https://doi.org/10.1371/journal.pone.0164769

Sterne, J. A. C. (2001). Sifting the evidence---what's wrong with significance tests? Another comment on the role of statistical methods. *BMJ*, *322*(7280), 226–231. https://doi.org/10.1136/bmj.322.7280.226

Stiasny, K., Röbbecke, J., Schüler, P., & Oertel, W. H. (2000). Treatment of idiopathic restless legs syndrome (RLS) with the D2-agonist cabergoline--an open clinical trial. *Sleep*, *23*(3), 349–354. http://www.ncbi.nlm.nih.gov/pubmed/10811379

Stoll, C. (1999). *High-tech heretic: Reflections of a computer contrarian.* First Anchor Books.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., Smith, K., Revaz, B., Finnbogason, B., Szantner, A., & Lundberg, E. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature Biotechnology*, *36*(9), 820–832. https://doi.org/10.1038/nbt.4225

Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious Games: An Overview*. http://his.diva-portal.org/smash/record.jsf?pid=diva2%3A2416&dswid=4653

Suttorp, M. M., Siegerink, B., Jager, K. J., Zoccali, C., & Dekker, F. W. (2015). Graphical presentation of confounding in directed acyclic graphs. *Nephrology Dialysis Transplantation*, *30*(9), 1418–1423. https://doi.org/10.1093/ndt/gfu325

Sward, K. A., Richardson, S., Kendrick, J., & Maloney, C. (2008). Use of a web-based game to teach pediatric content to medical students. *Ambulatory Pediatrics*, *8*(6), 354–359. https://doi.org/10.1016/j.ambp.2008.07.007

System Dynamics. (2010). *HealthBound*. https://systemdynamics.org/healthbound/

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, A. E., Fluharty, M. E., Bjørngaard, J. H., Gabrielsen, M. E., Skorpen, F., Marioni, R. E., Campbell, A., Engmann, J., Mirza, S. S., Loukola, A., Laatikainen, T., Partonen, T., Kaakinen, M., Ducci, F., Cavadino,

A., Husemoen, L. L. N., Ahluwalia, T. S., Jacobsen, R. K., Skaaby, T., … Munafò, M. R. (2014). Investigating the possible causal association of smoking with depression and anxiety using Mendelian randomisation meta-analysis: the CARTA consortium. *BMJ Open*, *4*(10), e006141. https://doi.org/10.1136/bmjopen-2014-006141

Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty. *Epidemiology*, *22*(5), 745. https://doi.org/10.1097/EDE.0b013e318225c2be

Thompson, J., & Munafo, M. (2019). Using prediction markets to estimate ratings of academic research quality in a mock Research Excellence Framework exercise. *MetaArxiv Pre-Print*.

Thulasiraman, K. ; S. M. N. S. (2011). *Graphs: Theory and Algorithms*. John Wiley & Sons.

Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, *42*(5), 441–447. https://doi.org/10.1038/ng.571

Tondello, G., Premsukh, H., & Nacke, L. (2018). A theory of gamification principles through goal-setting theory. *Proceedings of the 51st Hawaii International Conference on System Sciences*, *October 2017*. https://doi.org/10.24251/hicss.2018.140

Tozatto-Maio, K., Girot, R., Ly, I. D., Silva Pinto, A. C., Rocha, V., Fernandes, F., Diagne, I., Benzerara, Y., Dinardo, C. L., Soler, J. P., Kashima, S., Araujo, I. L., Kenzey, C., Fonseca, G. H. H., Rodrigues, E. S., Volt, F., Jarduli, L., Ruggeri, A., Mariaselvam, C., … Tamouza, R. (2020). Polymorphisms in inflammatory genes modulate clinical complications in patients with sickle cell disease. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.02041

Tullis, T., & Albert, B. (2008). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Elsevier.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., Magnusson, P., Oskarsson, S., Johannesson, M., Visscher, P. M., Laibson, D., Cesarini, D., Neale, B. M., & Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*(2), 229–237. https://doi.org/10.1038/s41588-017-0009-4

Valve. (2013). *Dota 2*. https://www.dota2.com/home

van Bilsen, A., Bekebrede, G., & Mayer, I. (2010). Understanding complex adaptive systems by playing games. *Informatics in Education*, *9*(1), 1–18. https://doi.org/10.15388/infedu.2010.01

van de water, A., Holmes, A., & Hurley, D. (2011). Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review. *Journal of Sleep Research*, *20*(1pt2), 183–200. https://doi.org/10.1111/j.1365-2869.2009.00814.x

van den Berg, S. M., de Moor, M. H. M., McGue, M., Pettersson, E., Terracciano, A., Verweij, K. J. H., Amin, N., Derringer, J., Esko, T., van Grootheest, G., Hansell, N. K., Huffman, J., Konte, B., Lahti, J., Luciano, M., Matteson, L. K., Viktorin, A., Wouda, J., Agrawal, A., … Boomsma, D. I. (2014). Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory. *Behavior Genetics*, *44*(4), 295–313. https://doi.org/10.1007/s10519-014-9654-x

Verbeek, I. H., Konings, G. M., Aldenkamp, A. P., Declerck, A. C., & Klip, E. C. (2006). Cognitive behavioral treatment in clinically referred chronic insomniacs: group versus individual treatment. *Behavioral Sleep Medicine*, *4*(3), 135–151. https://doi.org/10.1207/s15402010bsm0403_1

Vhaduri, S., & Poellabauer, C. (2018). Impact of different pre-sleep phone use patterns on sleep quality. *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 94–97. https://doi.org/10.1109/BSN.2018.8329667

Wade, A. (2010). The societal costs of insomnia. *Neuropsychiatric Disease and Treatment*, 1. https://doi.org/10.2147/NDT.S15123

Walker, M. P. (2009). The role of sleep in cognition and emotion. *Annals of the New York Academy of Sciences*, *1156*(1), 168–197. https://doi.org/10.1111/j.1749-6632.2009.04416.x

Walsh, J. K., Krystal, A. D., Amato, D. A., Rubens, R., Caron, J., Wessel, T. C., Schaefer, K., Roach, J., Wallenstein, G., & Roth, T. (2007). Nightly treatment of primary insomnia with eszopiclone for six months: effect on sleep, quality of life, and work limitations. *Sleep*, *30*(8), 959–968. https://doi.org/10.1093/sleep/30.8.959

Walsh, J. K., Roth, T., Randazzo, A., Erman, M., Jamieson, A., Scharf, M., Schweitzer, P. K., & Ware, J. C. (2000). Eight weeks of non-nightly use of zolpidem for primary insomnia. *Sleep*, *23*(8), 1–10. https://doi.org/10.1093/sleep/23.8.1h

Wang, H.-Y., & Wang, Y.-S. (2008). Gender differences in the perception and acceptance of online games. *British Journal of Educational Technology*, ???-??? https://doi.org/10.1111/j.1467-8535.2007.00773.x

Wardaszko, M. (2018). Interdisciplinary approach to complexity in simulation game design and implementation. *Simulation and Gaming*, *49*(3), 263–278. https://doi.org/10.1177/1046878118777809

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483. http://www.ncbi.nlm.nih.gov/pubmed/1593914

Wayment-Steele, H. K., Kim, D. S., Choe, C. A., Nicol, J. J., Wellington-Oguri, R., Watkins, A. M., Parra Sperberg, R. A., Huang, P.-S., Participants, E., & Das, R. (2021). Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Research*, *49*(18), 10604–10617. https://doi.org/10.1093/nar/gkab764

Weaver, W. (1948). Science and complexity. *American Scientist*.

Webb, P., Bain, C., & Page, A. (2020). *Essential epidemiology* (4th ed.). Cambridge University Press.

Werbach, K., & Hunter, D. (2012). *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press.

White, C. A., Uttl, B., & Holder, M. D. (2019). Meta-analyses of positive psychology interventions: The effects are much smaller than previously reported. *PLOS ONE*, *14*(5), e0216588. https://doi.org/10.1371/journal.pone.0216588

Winn, B. M. (2009). The design, play, and experience framework. In *Handbook of Research on Effective Electronic Gaming in Education* (pp. 1010–1024). IGI Global. https://doi.org/10.4018/978-1-59904-808-6.ch058

Winskell, K., Sabben, G., & Obong'o, C. (2019). Interactive narrative in a mobile health behavioral intervention (tumaini): Theoretical grounding and structure of a smartphone game to prevent hiv

among young africans. In *JMIR Serious Games* (Vol. 7, Issue 2). JMIR Publications Inc. https://doi.org/10.2196/13037

Wood, D. F. (2003). ABC of learning and teaching in medicine: Problem based learning. *BMJ*, *326*(7384), 328–330. https://doi.org/10.1136/bmj.326.7384.328

Wootton, R. E., Greenstone, H. S. R., Abdellaoui, A., Denys, D., Verweij, K. J. H., Munafò, M. R., & Treur, J. L. (2021). Bidirectional effects between loneliness, smoking and alcohol use: evidence from a Mendelian randomization study. *Addiction*, *116*(2), 400–406. https://doi.org/10.1111/add.15142

Wootton, R. E., Jones, H. J., & Sallis, H. M. (2022). Mendelian randomisation for psychiatry: how does it work, and what can it tell us? *Molecular Psychiatry*, *27*(1), 53–57. https://doi.org/10.1038/s41380-021-01173-3

Wootton, R. E., Lawn, R. B., Millard, L. A. C., Davies, N. M., Taylor, A. E., Munafò, M. R., Timpson, N. J., Davis, O. S. P., Davey Smith, G., & Haworth, C. M. A. (2018). Evaluation of the causal effects between subjective wellbeing and cardiometabolic health: mendelian randomisation study. *BMJ*, k3788. https://doi.org/10.1136/bmj.k3788

Wootton, R. E., Richmond, R. C., Stuijfzand, B. G., Lawn, R. B., Sallis, H. M., Taylor, G. M. J., Hemani, G., Jones, H. J., Zammit, S., Davey Smith, G., & Munafò, M. R. (2020). Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychological Medicine*, *50*(14), 2435–2443. https://doi.org/10.1017/S0033291719002678

World Health Organisation. (2012). The World Health Organization Quality of Life (WHOQOL). *The World Health Organisation Publications*. https://www.who.int/tools/whoqol

World Health Organisation. (2018). *International classification of diseases for mortality and morbidity statistics (11th Revision)*.

World Health Organisation. (2021). *Health topics: Public health services*. https://www.euro.who.int/en/health-topics/Health-systems/public-health-services#:~:text=Public Health is defined as,Acheson%2C 1988%3B WHO).&text=Public health focuses on the,the eradication of particular diseases.

World Health Organization. (1993). *The ICD-10 Classification of Mental and Behavioural Disorders*. World Health Organization.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., … Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*(5), 668–681. https://doi.org/10.1038/s41588-018-0090-3

Wright, S. M., & Aronne, L. J. (2012). Causes of obesity. *Abdominal Radiology*, *37*(5), 730–732. https://doi.org/10.1007/s00261-012-9862-x

Wulff, K., Gatti, S., Wettstein, J. G., & Foster, R. G. (2010). Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, *11*(8), 589–599.

Yang, J., Yan, B., Zhao, B., Fan, Y., He, X., Yang, L., Ma, Q., Zheng, J., Wang, W., Bai, L., Zhu, F., & Ma, X. (2020). Assessing the Causal Effects of Human Serum Metabolites on 5 Major Psychiatric Disorders. *Schizophrenia Bulletin*, *46*(4), 804–813. https://doi.org/10.1093/schbul/sbz138

Yardley, L., Morrison, L., Bradbury, K., & Muller, I. (2015). The person-based approach to intervention development: Application to digital health-related behavior change interventions. *Journal of Medical Internet Research*, *17*(1), e30. https://doi.org/10.2196/jmir.4055

Yee, N. (2006). Motivations for play in online games. *Cyberpsychology and Behavior*, *9*(6), 772–775. https://doi.org/10.1089/cpb.2006.9.772

Yusoff, A., Crowder, R., Gilbert, L., & Wills, G. (2009). A conceptual framework for serious games. *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, 21–23. https://doi.org/10.1109/ICALT.2009.19

Zhang, F., Cao, H., Zhang, X., Xu, Y., Tian, L., Yuan, J., Wang, G., & Baranova, A. (2019). Investigating the associations of major depressive disorder with various health outcomes in the context of common genetic variants: a mendelian randomisation study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3494404

Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M., & Smith, G. D. (2017). Recent Developments in Mendelian Randomization Studies. *Current Epidemiology Reports*, *4*(4), 330–345. https://doi.org/10.1007/s40471-017-0128-6

Zhou, F., Guo, Y., Wang, Z., Liu, S., & Xu, H. (2021). Assessing the causal associations of insomnia with depressive symptoms and subjective well-being: a bidirectional Mendelian randomization study. *Sleep Medicine*, *87*, 85–91. https://doi.org/10.1016/j.sleep.2021.08.025

Zichermann, G., & Linder, J. (2010). *Game-Based Marketing: Inspire Customer Loyalty Through Rewards, Challenges, and Contests*. John Wiley & Sons.

Zielinski, M., McKenna, J., & McCarley, R. (2016). Functions and mechanisms of sleep. *AIMS Neuroscience*, *3*(1), 67–104. https://doi.org/10.3934/Neuroscience.2016.1.67

Ziporyn, T. D., Malow, B. A., Oakes, K., & Wahlstrom, K. L. (2017). Self-report surveys of student sleep and well-being: a review of use in the context of school start times. *Sleep Health*, *3*(6), 498–507. https://doi.org/10.1016/j.sleh.2017.09.002

Zwass, V. (2010). Co-creation: toward a taxonomy and an integrated research perspective. *International Journal of Electronic Commerce*, *15*(1), 11–48. https://doi.org/10.2753/JEC1086-4415150101

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer*, *38*(9), 25–32.

Appendix

Items in this appendix are listed by the empirical chapters they are referenced in (chapters 2-6).

2

2.1 Search terms for literature review

I searched PubMed and Scopus for articles published any time in any language which had titles matching the search terms: (insomnia, sleep, sleepless*) AND (wellbeing, well-being, well being, happiness, satisfaction, swb, or quality of life).

2.2 Instruments for wellbeing and insomnia

Table 2.1 Genetic variants used as instruments in my MR analysis

Wellbeing

SNP RSID	Effect allele	Non-effect allele	b	R^2	Proxy SNP
rs3756290	G	A	0.01772	0.000115	
rs2075677	G	A	-0.02201	0.000177	
rs4958581	C	T	-0.0134	8.09E-05	

Insomnia

SNP RSID	Effect allele	Non-effect allele	b	R^2	Proxy SNP
rs12310246	A	G	0.010302	7.80E-05	
rs728017	G	A	0.005409	2.73E-05	
rs11090039	A	G	0.007447	4.44E-05	
rs10865954	C	T	-0.01035	9.36E-05	
rs4643373	C	T	-0.00671	3.70E-05	
rs9527083	A	G	-0.00976	8.14E-05	
rs1519102	C	G	-0.00718	4.32E-05	
rs3131638	G	A	0.006921	3.29E-05	
rs7566062	T	C	0.009066	5.62E-05	
rs72820274	A	G	0.005392	2.76E-05	rs6707445
rs34036083	C	T	0.006767	4.03E-05	rs10865356
rs7040224	G	A	-0.00603	3.08E-05	
rs2491124	C	T	-0.00927	8.23E-05	
rs17005118	A	G	0.008087	4.98E-05	
rs224029	C	T	0.007136	4.81E-05	
rs4588900	A	G	0.006433	4.05E-05	
rs66674044	T	A	0.009659	4.47E-05	rs12444979
rs176644	T	G	0.008052	6.10E-05	
rs2398144	A	C	0.005496	2.79E-05	

rs2815757	T	C	0.008833	4.75E-05	rs1460940
rs3184470	A	G	-0.00589	3.11E-05	rs3743909
rs6465151	T	C	0.012546	6.24E-05	
rs35322724	A	C	0.00636	3.84E-05	rs8056764
rs16903122	T	C	0.007353	3.98E-05	
rs2867690	C	T	-0.00764	3.40E-05	
rs10800992	T	C	0.006038	3.54E-05	rs12402747
rs28611339	T	G	0.012743	7.13E-05	rs7014570
rs56133505	A	G	0.006336	3.89E-05	rs12146545
rs2903385	A	G	0.007059	4.87E-05	rs10010325
rs10947690	G	A	0.009191	6.37E-05	
rs6967168	G	T	0.008071	4.76E-05	rs6956407
rs67501351	G	C	-0.00773	4.46E-05	rs2521480
rs6119267	G	C	0.008461	6.00E-05	
rs671985	A	G	-0.00528	2.70E-05	rs1214478
rs1031654	A	C	-0.01042	6.83E-05	
rs984306	T	C	-0.00842	5.18E-05	
rs7571486	A	G	-0.0065	3.11E-05	rs6729029
rs10947987	T	C	-0.00668	4.33E-05	rs6902650
rs11588755	A	G	-0.00874	7.48E-05	rs11590708
rs12924275	T	C	0.006229	2.97E-05	rs4985101
rs17025198	A	G	0.009004	5.18E-05	
rs13138995	G	A	-0.00592	3.23E-05	
rs324017	C	A	-0.00988	7.93E-05	rs324015
rs10761240	A	G	-0.00902	7.61E-05	
rs314281	C	T	0.009659	9.08E-05	rs314280
rs11605348	A	G	-0.00838	6.25E-05	rs12419692
rs2792990	C	G	0.008509	3.53E-05	rs733329
rs6606731	A	T	0.009525	5.53E-05	
rs1861412	A	G	0.010935	0.000115	
rs62213452	T	G	0.005908	2.75E-05	rs1396777
rs72773790	C	T	-0.00604	3.14E-05	rs12684650
rs116466468	C	T	-0.00855	5.24E-05	rs11686762
rs12917449	C	A	0.009935	6.03E-05	
rs34214423	C	A	-0.00716	3.11E-05	rs12325489
rs11650304	G	C	-0.01097	3.03E-05	rs17617360
rs6888135	A	C	0.007787	5.94E-05	
rs7475916	G	C	0.006056	3.27E-05	
rs10947428	C	T	0.00865	4.95E-05	
rs524859	A	G	-0.00882	7.03E-05	
rs9889282	C	A	0.005565	2.87E-05	
rs12614369	G	A	-0.00826	4.03E-05	
rs60565673	G	T	0.01016	9.49E-05	rs1261073
rs8181889	A	G	-0.00621	3.63E-05	
rs9316619	C	T	-0.00901	4.60E-05	
rs2737240	G	A	-0.00635	3.27E-05	rs2737245

rs11756035	C	G	0.010132	4.50E-05	
rs12251016	T	A	0.006007	3.20E-05	rs10828247
rs7486418	T	G	0.006624	3.88E-05	
rs7214267	A	G	-0.01048	0.000105	rs6503422
rs12790660	C	T	0.006962	4.11E-05	
rs62264767	C	A	-0.01106	6.03E-05	rs6795060
rs3902952	T	C	0.006863	2.83E-05	rs12595958
rs12912299	T	C	-0.00588	3.34E-05	
rs6702604	G	A	0.009966	9.49E-05	rs10494048
rs9563886	C	T	0.005733	3.08E-05	
rs12187443	C	T	-0.00683	4.05E-05	
rs7168238	G	C	-0.01231	4.11E-05	
rs2838787	A	G	-0.00737	5.07E-05	
rs55972276	A	C	0.010277	4.89E-05	rs13174833
rs908668	T	C	0.007585	3.73E-05	

2.3 Additional MR plots

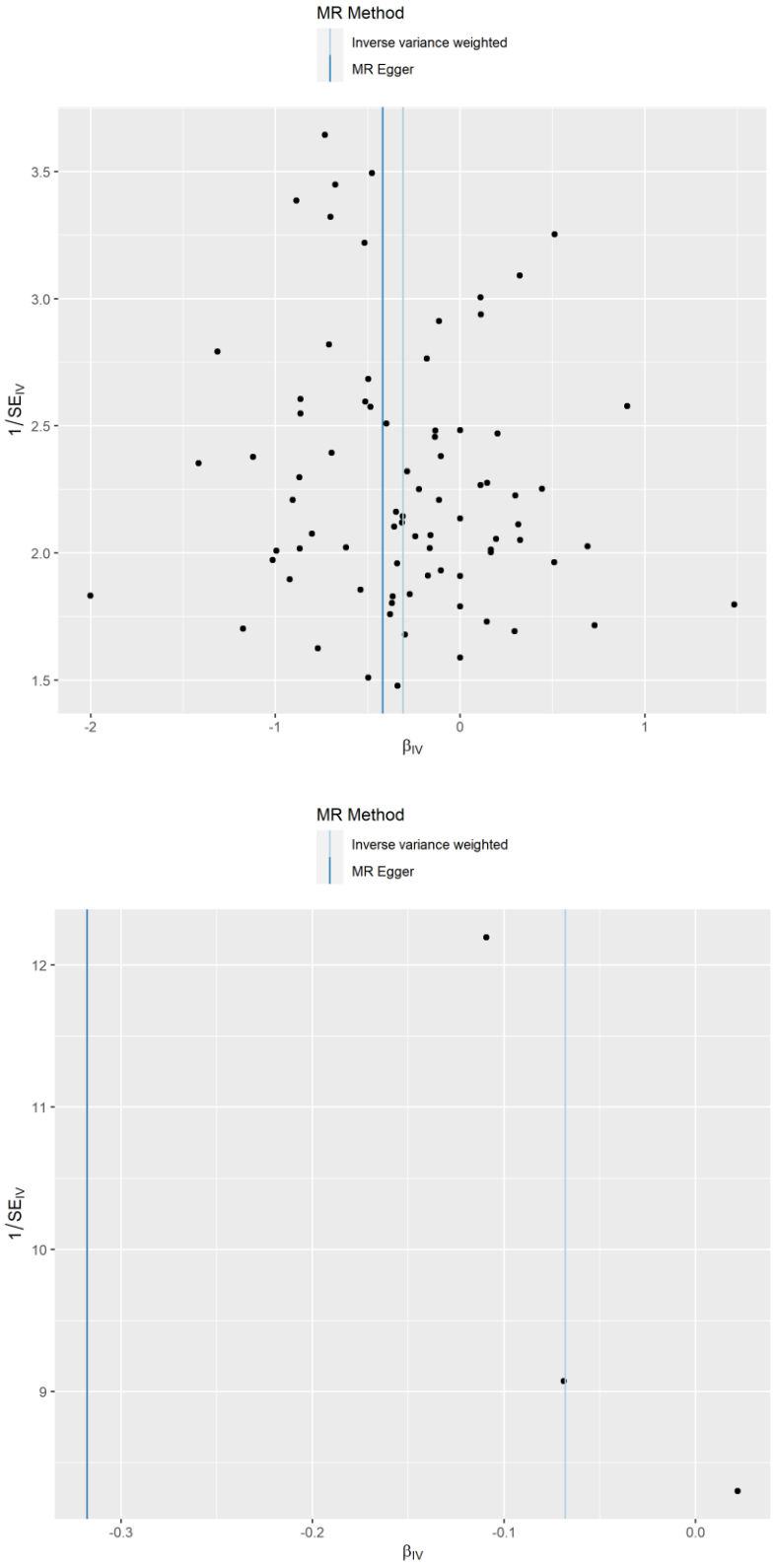
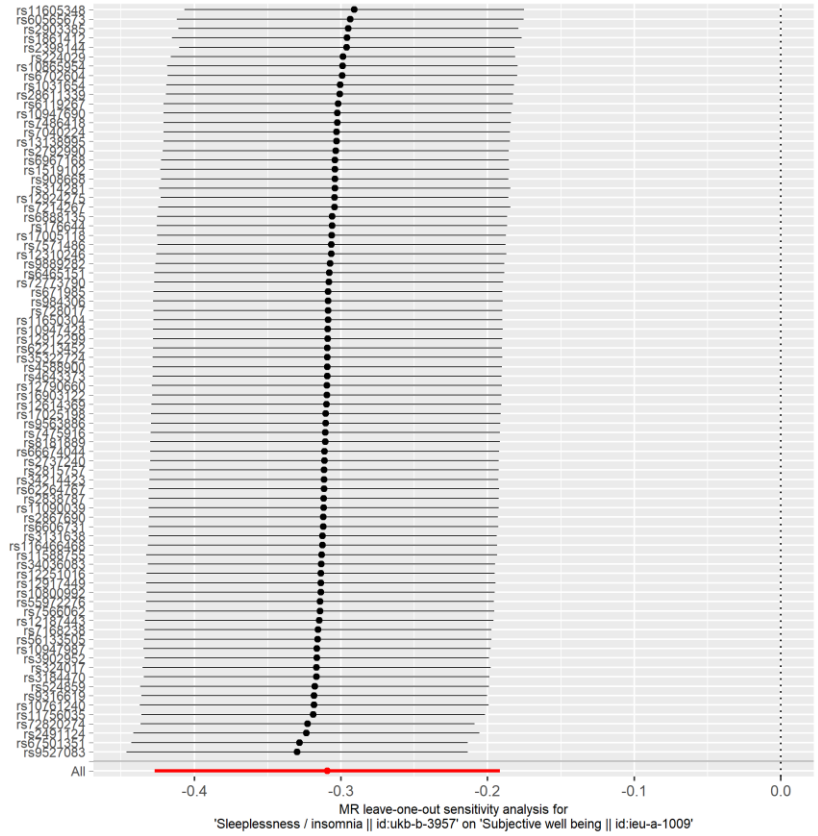


Figure 2.1 Funnel plots are used to identify if an analysis is biased by directional pleiotropy. In this cause the plots for insomnia on wellbeing (top) and wellbeing on insomnia (bottom) indicate that no substantial number

of variants bias estimation towards a negative effect, to the left of the graph, or towards a positive effect, to the right of the graph.



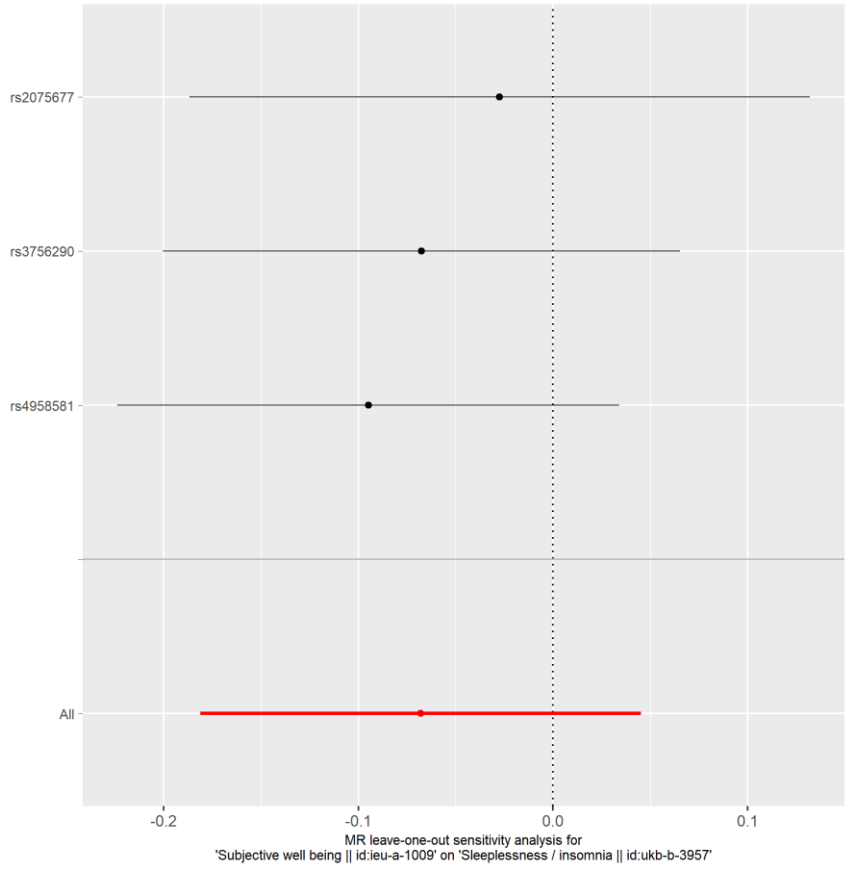


Figure 2.2 Leave-one-out analysis is used to identify if any individual variants drive effect estimation. Plots for insomnia on wellbeing (top) and wellbeing on insomnia (bottom) indicate that no single variant has an effect much larger than the others in either analysis.

2.4 Comparing instrument strengths

Instrument strengths were compared to previous studies (Jansen et al., 2019; Zhou et al., 2021) to understand the relative strength of my instruments for insomnia and wellbeing. Both studies used instruments for insomnia and wellbeing, but insomnia is only measured in compatible units for comparison in one study (Jansen et al., 2019) and wellbeing instruments are only reported in the other (Zhou et al., 2021), so these are analysed separately.

Instrument strength for insomnia was calculated by obtaining instrument-exposure associations for 116 SNPs associated with insomnia ($P<5x10^{-05}$)(Supplementary table 6)(Jansen et al., 2019). First, alleles were aligned to predict an increase in insomnia, and the mean instrument-exposure odds ratio was taken (mean=1.05, sd=0.021, min=1.03, max=1.23) as an indication of mean instrument strength. In order to compare with the R^2 values in the present study odds ratios were converted to R^2 values (using ESCAL effect converter: https://www.escal.site/). The result was a mean instrument-exposure association of R^2 = 0.017%. Since the authors do not present which of these variants were used as instruments (n=88) for insomnia in MR, it is assumed that obtained instruments had a mean instrument strength similar to this estimate (95% confidence interval of instrument-exposure associations: 0.014% - 0.020%).

Instrument strength for wellbeing was calculated by obtaining instrument-exposure associations (Supplementary table 5)(Zhou et al., 2021). As for insomnia, alleles were aligned to predict wellbeing, and the mean instrument-exposure association was taken (mean=0.013, SD=0.004, min=0.01, max=0.03). This resulted in a mean instrument-exposure b=0.013 which represents a 0.013 SD increase in wellbeing for effect alleles. To compare compatible units, the association statistic presented in this study, betas, were compared to comparable instrument-exposure betas in the present study.

Values in the present study for comparison are $R^2_{insomnia}$ = 0.005% (mean=4.98x10^{-05}, SD=2.04x10^{-05}, min=2.70x10^{-05}, max=1.15x10^{-04}) and $b_{wellbeing}$ = 0.018 (mean=0.018, sd=0.001, min=0.01, max=0.022). Comparing R^2 values for insomnia indicate that my mean instruments explain less variance in insomnia (0.005% vs 0.017%) but stronger association with wellbeing (b=0.018 vs b=0.013).

3

3.1 Effects in network analysis

Table 3.1 All effect estimates which reached network-wide significance (*P*<2.15x10^{-04})

Exposure	Outcome	nSNP	b	se	*P*
Alcohol	BMI	100	0.22	0.05	3.66x10^{-05}
Alcohol	Education	99	-0.19	0.03	8.18x10^{-12}
Alcohol	Intelligence	100	-0.49	0.08	4.70x10^{-10}
Alcohol	Loneliness	100	0.02	0.01	1.76x10^{-05}
Alcohol	Not socialising	100	0.08	0.01	6.98x10^{-30}
BMI	Alcohol	460	0.28	0.02	4.51x10^{-36}
BMI	CHD	451	0.41	0.04	1.14x10^{-29}
BMI	Coffee intake	460	0.09	0.01	7.51x10^{-29}
BMI	Diabetes	126	0.97	0.10	1.86x10^{-23}
BMI	Education	452	-0.18	0.01	1.04x10^{-38}
BMI	Exercise	460	-0.09	0.02	2.26x10^{-05}
BMI	Intelligence	458	-0.29	0.04	6.04x10^{-11}
BMI	Loneliness	458	0.03	0.00	1.89x10^{-21}
BMI	Not socialising	458	0.03	0.00	4.27x10^{-15}
BMI	Insomnia	460	0.05	0.01	1.01x10^{-08}
BMI	Smoking	343	1.88	0.31	8.38x10^{-10}
BMI	Worry	458	-0.04	0.01	4.17x10^{-14}
Coffee intake	BMI	40	0.72	0.18	4.44x10^{-05}
Coffee intake	Exercise	40	-0.37	0.09	2.79x10^{-05}
Depression	Loneliness	36	0.05	0.01	4.13x10^{-12}
Depression	Neuroticism	36	0.16	0.04	1.22x10^{-05}
Depression	Insomnia	36	0.09	0.02	8.05x10^{-06}
Depression	Worry	36	0.09	0.02	2.77x10^{-09}
Diabetes	CHD	39	0.11	0.03	9.26x10^{-05}
Education	Alcohol	318	-0.49	0.03	8.26x10^{-69}
Education	BMI	318	-0.35	0.03	3.22x10^{-40}
Education	CHD	316	-0.47	0.05	3.50x10^{-19}
Education	Coffee intake	318	0.10	0.01	7.21x10^{-19}
Education	Exercise	318	-0.24	0.04	5.32x10^{-10}
Education	Intelligence	318	1.69	0.06	2.31x10^{-182}
Education	Loneliness	318	-0.06	0.01	4.93x10^{-23}
Education	Neuroticism	313	-0.17	0.03	3.35x10^{-09}
Education	Not socialising	318	-0.12	0.01	6.00x10^{-70}

Education	Insomnia	318	-0.11	0.01	$4.93x10^{-20}$
Education	Smoking	250	-2.18	0.46	$2.25x10^{-06}$
Education	Wellbeing	250	0.08	0.02	$1.7x10^{-04}$
Exercise	Intelligence	20	-0.75	0.14	$1.05x10^{-07}$
Intelligence	Alcohol	79	-0.10	0.02	$8.00x10^{-08}$
Intelligence	CHD	78	-0.12	0.03	$6.80x10^{-05}$
Intelligence	Coffee intake	79	0.03	0.01	$8.07x10^{-08}$
Intelligence	Education	77	0.19	0.01	$6.38x10^{-53}$
Intelligence	Exercise	79	-0.12	0.02	$1.81x10^{-08}$
Intelligence	Loneliness	79	-0.01	0.00	$2.1x10^{-04}$
Intelligence	Not socialising	79	-0.02	0.00	$8.28x10^{-08}$
Loneliness	Depression	2	5.84	0.78	$5.80x10^{-14}$
Loneliness	Neuroticism	16	1.36	0.32	$2.67x10^{-05}$
Loneliness	Insomnia	16	0.66	0.17	$7.11x10^{-05}$
Loneliness	Worry	16	0.47	0.07	$1.31x10^{-11}$
Not socialising	Alcohol	10	2.00	0.37	$6.66x10^{-08}$
Not socialising	Depression	1	-5.61	0.90	$4.26x10^{-10}$
Not socialising	Education	10	-2.07	0.34	$1.18x10^{-09}$
Not socialising	Intelligence	10	-4.37	1.06	$4.12x10^{-05}$
Not socialising	Loneliness	10	0.25	0.05	$3.00x10^{-06}$
Insomnia	CHD	80	0.88	0.19	$2.29x10^{-06}$
Insomnia	Depression	6	2.96	0.31	$1.43x10^{-21}$
Insomnia	Education	80	-0.27	0.07	$6.12x10^{-05}$
Insomnia	Loneliness	80	0.13	0.02	$3.58x10^{-10}$
Insomnia	Not socialising	80	0.13	0.02	$4.02x10^{-07}$
Insomnia	Wellbeing	80	-0.31	0.06	$2.66x10^{-07}$
Insomnia	Worry	80	0.19	0.04	$1.44x10^{-07}$
Worry	Loneliness	67	0.19	0.02	$1.74x10^{-14}$
Worry	Neuroticism	63	0.77	0.13	$4.17x10^{-09}$
Worry	Insomnia	67	0.36	0.05	$2.43x10^{-13}$
Worry	Wellbeing	50	-0.41	0.08	$8.47x10^{-08}$

3.2 Steiger testing

Table 3.2 Results from Steiger testing in network MR analysis

Exposure	Outcome	Mean R2 (exposure)	Mean R2 (outcome)	Steiger pval	Steiger fail
Alcohol	BMI	0.011	0.008	1.22E-14	
Alcohol	Education	0.011	0.002	8.61x10-209	
Alcohol	Intelligence	0.011	0.005	9.74x10-36	
Alcohol	Loneliness	0.011	0.001	0	
Alcohol	Not socialising	0.011	0.001	1.08x10-265	
BMI	Alcohol	0.062	0.008	0	
BMI	CHD	0.061	0.005	0	
BMI	Coffee intake	0.062	0.004	0	
BMI	Diabetes	0.024	0.008	3.74x10-65	
BMI	Education	0.061	0.006	0	
BMI	Exercise	0.062	0.002	0	

BMI	Intelligence	0.062	0.012	0
BMI	Loneliness	0.062	0.003	0
BMI	Not socialising	0.062	0.002	0
BMI	Sleeplessness	0.062	0.004	0
BMI	Smoking	0.048	0.006	1.81x10-270
BMI	Worry	0.062	0.004	0
Coffee intake	BMI	0.007	0.007	0.287537 *
Coffee intake	Exercise	0.007	0.000	3.92x10-203
Depression	Loneliness	0.003	0.000	2.91x10-50
Depression	Neuroticism	0.003	0.000	1.43x10-30
Depression	Sleeplessness	0.003	0.001	3.41x10-40
Depression	Worry	0.003	0.001	7.64x10-23
Diabetes	CHD	0.031	0.001	4.16x10-283
Education	Alcohol	0.020	0.005	0
Education	BMI	0.020	0.008	7.45x10-178
Education	CHD	0.020	0.003	8.48x10-262
Education	Coffee intake	0.020	0.002	0
Education	Exercise	0.020	0.002	0
Education	Intelligence	0.020	0.021	0.49823 *
Education	Loneliness	0.020	0.002	0
Education	Neuroticism	0.020	0.002	5.86x10-264
Education	Not socialising	0.020	0.003	0
Education	Sleeplessness	0.020	0.002	0
Education	Smoking	0.016	0.004	6.28x10-57
Education	Wellbeing	0.016	0.001	0
Exercise	Intelligence	0.002	0.002	0.792904 *
Intelligence	Alcohol	0.021	0.002	9.57x10-282
Intelligence	CHD	0.021	0.001	7.82x10-248
Intelligence	Coffee intake	0.021	0.001	0
Intelligence	Education	0.021	0.004	3.53x10-196
Intelligence	Exercise	0.021	0.001	0
Intelligence	Loneliness	0.021	0.001	0
Intelligence	Not socialising	0.021	0.001	0
Loneliness	Depression	0.000	0.000	0.737645
Loneliness	Neuroticism	0.001	0.000	2.82E-14
Loneliness	Sleeplessness	0.001	0.000	1.37E-17
Loneliness	Worry	0.001	0.000	3.12x10-22
Not socialising	Alcohol	0.001	0.000	0.00023
Not socialising	Depression	0.000	0.000	0.681406 *
Not socialising	Education	0.001	0.001	0.897781 *
Not socialising	Intelligence	0.001	0.001	0.061122 *
Not socialising	Loneliness	0.001	0.000	1.4E-17
Sleeplessness	CHD	0.004	0.001	3.6x10-35
Sleeplessness	Depression	0.000	0.000	0.508669
Sleeplessness	Education	0.004	0.001	1.46x10-79
Sleeplessness	Loneliness	0.004	0.001	8.81x10-70
Sleeplessness	Not socialising	0.004	0.001	2.17x10-75
Sleeplessness	Wellbeing	0.004	0.000	2.65x10-68

Sleeplessness	Worry	0.004	0.001	2.34x10-44
Worry	Loneliness	0.006	0.001	1.28x10-136
Worry	Neuroticism	0.006	0.001	9.92x10-60
Worry	Sleeplessness	0.006	0.001	3.22x10-121
Worry	Wellbeing	0.005	0.000	6.6x10-102

Note: Comparisons which failed the Steiger test are marked with an asterisk (*) and further research is required to obtain better evidence that the causal effect exists in the direction reported in this study.

3.3 Follow-up analysis of genetic associations

In order to further investigate any sources of confounding, I tested the overlap in instruments used for each of the 16 exposures. The aim of this follow-up analysis was to identify if any of the factors had instruments which were associated with the other factors in analysis. Initially, I searched for overlap in the RSIDs of the 1305 single nucleotide variants used as instruments in the network analysis. This did not reveal any identical instruments which were used for multiple exposures. This approach had limited power to detect overlap in instruments because it is common to "clump" variants very often occur together into a single lead variant, and this clumping could hide an underlying overlap in variants. I expanded my search by obtaining variants associated with each of the 16 factors without clumping.

Searching GWAS summary datasets again with a liberal threshold for association ($P<5x10^{-06}$) and no clumping returned a much larger number of variants associated with the 16 factors in analysis (n=301,036). Comparing RSID overlap in this sample revealed that a third (32%) of the variants associated with one exposure were also associated with at least one other factor in the network as well. Breaking these results down per exposure (**Table 3.2**) indicates that some factors, such as exercise (74% overlap), showed more genetic overlap than others, such as BMI (18% overlap) so may have been differently at risk of pleiotropy. Furthermore, instruments for insomnia (4%) were in high linkage disequilibrium with the instruments used for other factors in analysis ($R^2>.8$). These findings are important because they support the sensitivity testing and indicate that a minority of effect estimates may have been biased by pleiotropy.

Table 3.2 Number of variants associated with one or more exposures in analysis

	Variants (SNPs)	Associated with at least one other exposure (P<5x10^{-6})	%
Total	301036	95522	32%
Wellbeing	163	78	48%
Insomnia	9953	4158	42%
Depression	267	99	37%
Worry	21726	13166	61%
Alcohol	19835	11345	57%
Smoking	174	43	25%
Education	72359	19846	27%
BMI	141187	26093	18%
Intelligence	18671	11649	62%
Loneliness	5675	4178	74%
Exercise	4959	3351	68%
Not socialising	2751	1585	58%
Neuroticism	36	1	3%
Coffee intake	7589	2991	39%
Diabetes	1220	341	28%
CHD	4424	756	17%

3.4 Mediation analysis calculations

Steps taken to perform mediation analysis in network MR:

1. Identify mediating pathways

$$x_{sleeplessness} \rightarrow z_{\text{depression}} \rightarrow y_{wellbeing}$$
$$x_{sleeplessness} \rightarrow z_{\text{worry}} \rightarrow y_{wellbeing}$$
$$x_{sleeplessness} \rightarrow z_{\text{education}} \rightarrow y_{wellbeing}$$

2. Calculate the indirect effects

$$\hat{\beta}x \Longrightarrow \mathrm{y}_{\text{depression}} = 2.9 * -0.11 = -0.32$$

$$\hat{\beta}x \Longrightarrow \mathrm{y}_{\text{worry}} = 0.17 * -0.42 = -0.07$$

$$\hat{\beta}x \Longrightarrow \mathrm{y}_{\text{education}} = -0.33 * 0.07 = -0.023$$

3. Subtract indirect effect from total effect estimate to give direct estimate

$$\hat{\beta}x^1 \rightarrow y^1 = -0.31 - \sum(-0.32, -0.07, -0.02)$$

$$\textbf{Direct effect} = \mathbf{0.10}$$

4

4.1 Literature review

4.1.1 Methods

Defining terms

I began the review by defining what was meant by the "current practice" of researchers who "visualise" results containing "network relationships":

- I defined "Current practice" as the content included in academic papers.
- I defined "Visualisation" as any content presented in figures as part of the main paper. Conversely "non-visualisation" content was defined as content presented in tables
- I defined "Network relationships" with the intent to capture a large range of content where researchers described associated relationships: A network relationship constituted the effect of a single factor on multiple others, the effect of multiple factors on a single factor, or the effects of multiple factors on multiple other factors. This definition was not restricted to causal analyses, and included associations, but examples of this definition are displayed in Figure 4.1.

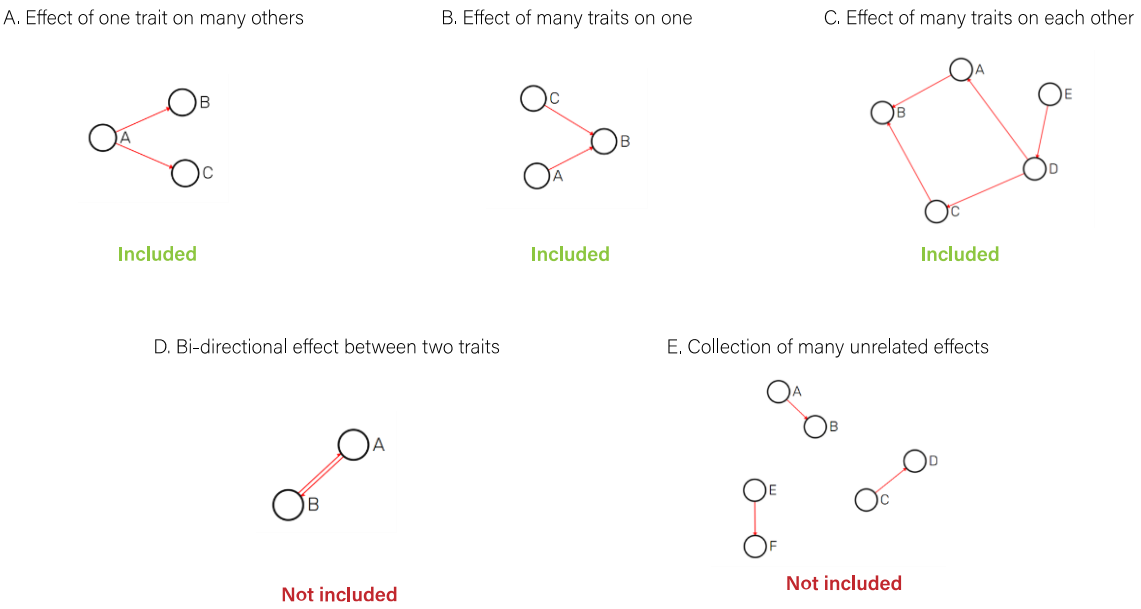


Figure 4.1. Examples of causal effect data which would and would not meet the criteria for "network relationships"

Search strategy

I searched Scopus to identify academic papers for analysis. I searched titles, abstracts and keywords for the terms "network" AND ("MR" OR "mendelian randomisation" OR "mendelian randomization" OR "mendelianrandomisation" OR "mendelianrandomization"). Results included academic papers published in English between 2008 and October 2021.

Screening criteria

I screened results for relevance for the following criteria:

> **Network studies**: Results were included in analysis if the article considered or analysed network relationships consistent with the definition above.
>
> **Academic works**: The full range of academic modalities were included: Empirical studies, software notes, review papers, books and chapters, conference papers and posters, opinion pieces and editorials.

I screened all articles identified by the search strategy to determine whether they were relevant based on the criteria above. In most cases it was clear from the title and abstract whether the term "network" was used in a relevant manner, to refer to network relationships, or in an irrelevant manner, for example, referencing networks of contributing researchers or databases. In the cases where there was not enough information to make this judgement from the abstract alone the full text was read. Generally, I erred on the side of caution and included papers for analysis to ensure relevant papers were not excluded.

Data extraction

Following screening, I extracted data to understand how authors used tables and figures to convey network relationships. For each table and figure data was extracted using a standardised form. Surrounding content was used for context. For each article, the standardised form recorded the frequency with which tables and graphs were used to present network relationships in the main body of the text (i.e., not supplementary materials or appendices since not all papers had one). For this purpose, figures were categorised into sixteen sub-types. Generally, different categories were visually distinct so no formal criteria for categorisation was used. The exception to this was Forest and dot plots which are conceptually distinct but were implemented in identical manners so were grouped together. In cases of ambiguity I referred to the authors' categorisation of graph type. Lastly the section of the paper which a table or figure sat was used to determine whether the author used it to present background and theory (in the introduction) or results (in the results). Examples of tables and figures were also saved for demonstrative purposes.

Included works

My search produced 121 articles matching search terms. After screening twenty-nine remained for analysis (see Table 4.1 for included articles). Of these papers twenty-seven were articles, one was a review (Richmond, 2016) and one was an editorial (Morgan, 2017). Results were sourced from a range of publications including Frontiers (n=3), BMC Biomedicine Central (n=3), and Nature (n=2). Authors applied mendelian randomisation to a range of topics: Genetics, Epigenetics and Genomics (n=10), Epidemiology and Public Health (n=10), cellular and systemic Biology (n=8), as well as Neuroscience (n=1). This search was originally conducted in April 2019 and was expanded in October 2021 which resulted in an 88% increase in included articles (from 18 to 34).

See Figure 4.2 for search strategy and Figure 4.3 for data extraction process.

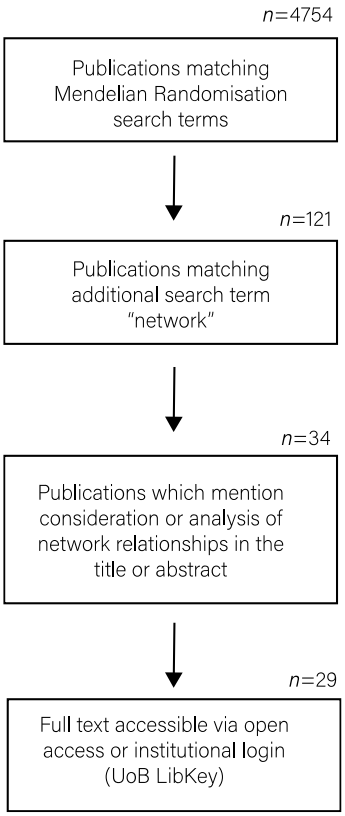


Figure 4.2 Process for article discovery and screening process

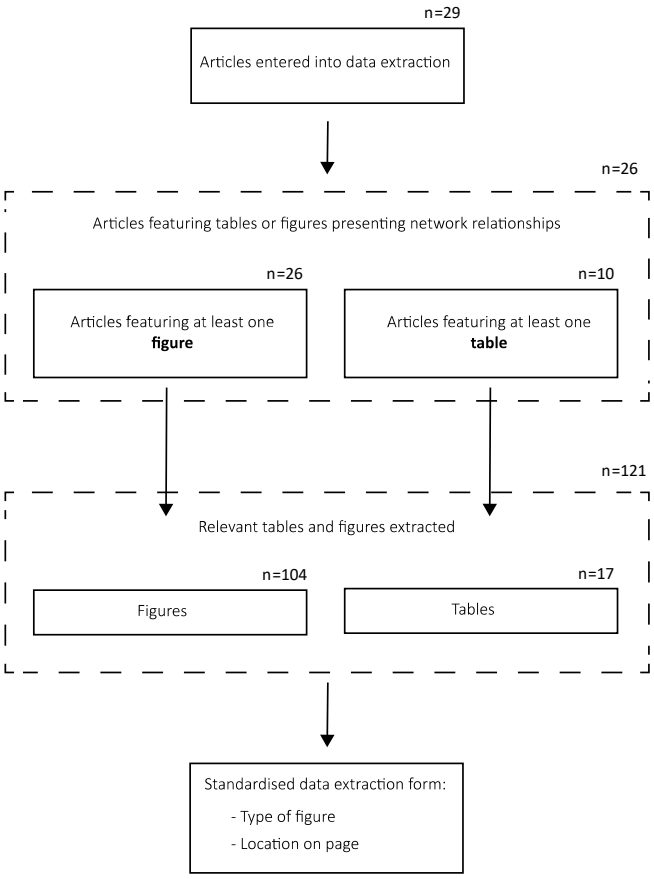


Figure 4.3 Data extraction process

Table 4.1 Included articles in review

First author (year)	Title	DOI
Anacleto (2019)	Integrating a genome-wide association study with a large-scale transcriptome analysis to predict genetic regions influencing the glycaemic index and texture in rice	10.1111/pbi.13051
Badsha (2019)	Learning causal biological networks with the principle of Mendelian randomization	10.3389/fgene.2019.00460

Badsha (2021)	MRPC: An R Package for Inference of Causal Graphs	10.3389/fgene.2021.651812
Bandres-Ciga (2020)	Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease	10.1007/s00401-020-02181-3
Burgess (2015)	Network Mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways	10.1093/ije/dyu176
Evans (2015)	Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality	10.1146/annurev-genom-090314-050016
Hou (2020)	Exploring the causal pathway from ischemic stroke to atrial fibrillation: A network Mendelian randomization study	10.1186/s10020-019-0133-y
Howey (2020)	Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data	10.1371/journal.pgen.1008198
Hu (2020)	Exploring the causal pathway from body mass index to coronary heart disease: a network Mendelian randomization study	10.1177/2040622320909040
Huai (2020)	A Comprehensive Analysis of MicroRNAs in Human Osteoporosis	10.3389/fendo.2020.516213

Li (2021)	Causal effect of sex hormone-binding globulin and testosterone on coronary heart disease: A multivariable and network Mendelian randomization analysis	10.1016/j.ijcard.2021.06.037
Liao (2020)	Exploring the causal pathway from omega-6 levels to coronary heart disease: A network Mendelian randomization study	10.1016/j.numecd.2019.09.013
Liu (2021a)	Genetically predicted insomnia in relation to 14 cardiovascular conditions and 17 cardiometabolic risk factors: A mendelian randomization study	10.1161/JAHA.120.020187
Liu (2021b)	Novel DNA methylation loci and genes showing pleiotropic association with Alzheimer's dementia: a network Mendelian randomization analysis	10.1080/15592294.2021.1959735
Luijk (2018)	Genome-wide identification of directed gene networks using large-scale population genomics data	10.1038/s41467-018-05452-6
Meng (2018)	Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions	10.1016/j.bone.2018.05.012
Morgan (2017)	Network Mendelian Randomization Study Design to Assess Factors Mediating the Causal Link between Telomere Length and Heart Disease	10.1161/CIRCRESAHA.117.311387
Noordam (2020)	Proteome-wide assessment of diabetes mellitus in Qatari identifies IGFBP-2 as a risk factor already with early glycaemic disturbances	10.1016/j.abb.2020.108476
Richmond (2016)	Challenges and novel approaches for investigating molecular mediation	10.1093/hmg/ddw197

Si (2020)	Causal Pathways from Body Components and Regional Fat to Extensive Metabolic Phenotypes: A Mendelian Randomization Study	10.1002/oby.22857
Sieurin (2021)	Neuroticism, Smoking, and the Risk of Parkinson's Disease	10.3233/JPD-202522
Davey Smith (2014)	Mendelian randomization: Genetic anchors for causal inference in epidemiological studies	10.1093/hmg/ddu328
Thom (2020)	Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes	10.1186/s12920-020-00742-9
Timpson (2011)	C-reactive protein levels and body mass index: Elucidating direction of causation through reciprocal Mendelian randomization	10.1038/ijo.2010.137
Yang (2020)	Assessing the causal effects of human serum metabolites on 5 major psychiatric disorders	10.1093/schbul/sbz138
Yazdani (2019)	Genome analysis and pleiotropy assessment using causal networks with loss of function mutation and metabolomics	10.1186/s12864-019-5772-4
Yazdani (2020)	Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network	10.1186/s12859-020-03753-6
Yuan (2020)	Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies	10.1038/s41467-020-17668-6

Zhan (2017)	Exploring the Causal Pathway from Telomere Length to Coronary Heart Disease: A Network Mendelian Randomization Study	10.1161/CIRCRESAHA.116.310517

4.1.2 Results

Table 4.2. Results of data extraction. Count of figures and tables used to present network relationships in MR publications

		Figures													
		All	Network graphs			Diagrams					Plots			Charts	
First author (year)	Tables		Un-directed	Directed	DAG	Dendrogram	Venn	Cord	Kite	Heatmap	Forest	Scatter	Box	Pie	Bar
Anacleto (2019)		5	1			1			1	1			1		
Badsha (2019)		9	1	3	3					1		1			
Badsha (2021)		6		3	2	1									
Bandres-Ciga (2020)		0													
Burgess (2015)	2	5		2	3										
Davey Smith (2014)	1	4			4										
Evans (2015)		1			1										

Hou (2020)	2	1		1								
Howey (2020)	1	5		3	2							
Hu (2020)	4	2		1	1							
Huai (2020)		11	5				1		1	2		2
Li (2021)		2		1	1							
Liao (2020)	1	1			1							
Liu (2021a)	1	2								2		
Liu (2021b)	2	3		1	1		1					
Luijk (2018)		3		2				1				
Meng (2018)		1	1									
Morgan (2017)		0										
Noordam (2020)		2								2		
Richmond (2016)		5		1	4							
Si (2020)		7			1			1		5		
Sieurin (2021)	1	2			1						1	
Thom (2020)		8				2			2		2	2
Timpson (2011)		1		1								

Yang (2020)	2	2						1		1				
Yazdani (2019)		5			5									
Yazdani (2020)		10		3	3				2		1	1	1	
Yuan (2020)		0												
Zhan (2017)	1	1			1									

Content analysis

The features of network graphs used to present network relationships in MR papers were summarised to understand current practice in MR visualisation (Table 4.3). Ten different styles of network graph were used for this analysis:

- "DAG 1" - Yazdani (2019), Figure 7
- "DAG 2" - Yazdani (2020), Figures 3 & 4
- "DAG 3" - Howey (2020), Figures 1, 8 & 9
- "Cyclic graph 1" - Badsha: Figures 5 & 7 (2019), 4 & 5 (2021)
- "Cyclic graph 2" - Yazdani (2020), Figure 7
- "Cyclic graph 3" - Hu (2020), Figure 2
- "Undirected graph 1" - Anacleto (2019), Figure 5
- "Undirected graph 2" - Meng 2018, Figure 4
- "Undirected graph 3 - Luijk (2018), Figure 2
- "Undirected graph 4" - Huai (2020), Figure 4

Table 4.3 Summary of evidence for features enabling inference

Feature	Supporting figures	Count
Design		
Coloured nodes	DAGs 1 and 2, Cyclic graph 2, and Undirected graphs 1-4	7
Node sizing	Undirected graph 4	1
Node shape	Undirected graph 4	1
Results formats		
Directional		6
- Cyclic	DAGs 1-3	3
- Acyclic	Cyclic graphs 1-3	3
Non-directional	Undirected graphs 1-4	4
Statistical parameters		
None	DAG 1, Cyclic graphs 2 and 3, Undirected graphs 1-4	7
Annotated edges	DAGs 2 and 3, Cyclic graph 1	3
Variable edge thickness	DAG 3	1
Layout		
Layout		10
- Force	DAG 1, Cyclic graph 2, and Undirected graphs 1, 3 and 4	5
- Rigid	DAGs 2 and 3, Cyclic graphs 1 and 3	4
- Circular	Undirected graph 2	1

4.2 Estimate of programming languages used for performing MR

4.2.1 Methods

I made a search of publicly available repositories on GitHub (www.github.com). I used the terms "mendelian randomisation" OR "mendelian randomization" OR "mendelianrandomisation" OR

"mendelianrandomization". Results included all software published up until October 2021. Languages were identified by the GitHub algorithmic determination of the most dominant programming language for files in each repository.

4.2.2 Results

Table 4.4 Languages of public repositories related to MR on GitHub

Language	Packages	Percentage
R	101	53%
Python	33	17%
JavaScript	10	5%
Java	8	4%
C++	14	7%
Stata	7	4%
HTML	6	3%
Perl	5	3%
Shell	5	3%

4.3 Software review

4.3.1 Methods

Search strategy

Software archives and repositories were searched to identify software used by MR researchers to visualise results. I searched archives for the statistical packages Stata, R, and Python as well as a broad search of publicly available repositories and software. I used the terms "mendelian randomisation" OR "mendelian randomization" OR "mendelianrandomisation" OR

"mendelianrandomization" although the implementation of search functions varied across software archives and repositories:

R packages were searched on the Comprehensive R Archive Network (https://cran.r-project.org/) using the "CRAN packages - general info" tag.

Stata packages were searched on the Boston College Statistical Software Components archive (SSC: https://ideas.repec.org/s/boc/bocode.html).

Python packages were searched on the Python Package Index (https://pypi.org/). Search terms were entered separately one-by-one and hits were aggregated.

Software in any programming language was searched GitHub (https://github.com/search) using the "repositories" tag.

R packages were additionally searched on Bioconductor (https://bioconductor.org/packages/devel/bioc/). The search term "mendelian" was used to capture as many results as possible due to a low number of results.

Google was searched for available software (www.google.com). The search terms were: ("mendelian randomisation" OR "mendelian randomization" OR "mendelianrandomisation" OR "mendelianrandomization") AND "visualisation" AND ("tool" OR "software"). However, it should be noted that Google will not necessarily return search results with a consistent ranking over time and for different users as search results are tailored. Only the first 80 results were examined before Google omits "results very similar to those already found".

Screening criteria

Results were screened for relevance according to the following criteria:

Software packages: Results were included in analysis if they were released as packages intended for others to use. For example, open source code for publications were excluded.

Unavailability: Results were excluded from analysis if either the software or its documentation were not accessible online

I screened results and where software was included rather than excluded when in doubt. In the cases where software had documentation, but it did not detail the capabilities of the software, it was excluded based on unavailability.

Data extraction

Following screening, data were extracted to understand the capabilities of software. A standardised data extraction form was used to record visualisation capabilities for each piece of software. Capabilities were inferred from all available documentation including package description, wiki and help pages, vignettes and publications. The standardised form was used to record whether each software was capable of producing visualisations in the sixteen categories used in the literature review.

Included works

GitHub, CRAN and Google software were pre-screened to remove duplicates and standardise search results. 152 GitHub repositories were subject to pre-screening to exclude software packages which were not intended for other researchers to use to conduct MR research. 61 were excluded since they contained code, data and written materials for published studies (n=52) or a poster (n=1), learning and teaching materials (n=5), or source code for a website (n=3). 11 GitHub repositories were excluded as duplicates since they contained code for the ten software already discovered for R, Stata and Python. 18 software for R were pre-screened to exclude general software utilities used as dependencies (n=11) which appeared in the search along with software that met search terms for Mendelian randomisation. Google results were excluded if they were academic papers (n=51) or books (n=2) which did not describe software, databases and search sites for papers (n=4) or software (n=2), talks, posters and teaching materials (n=6), laboratory or researcher personal profiles (n=7).

My search produced 93 software matching search terms. After screening 60 software remained which were available and had sufficient documentation. One piece of software discovered during the literature search (Cytoscape: https://cytoscape.org/), and one previously known software (MR Visualisation Tool: http://bristol-medical-stat.bristol.ac.uk:3838/MR-Vis/) were manually added to this search to bring the total to 62 software included in analysis. See *Figure 4.4* for the search strategy and screening processes (pre-screening activities shown in rounded rectangles).

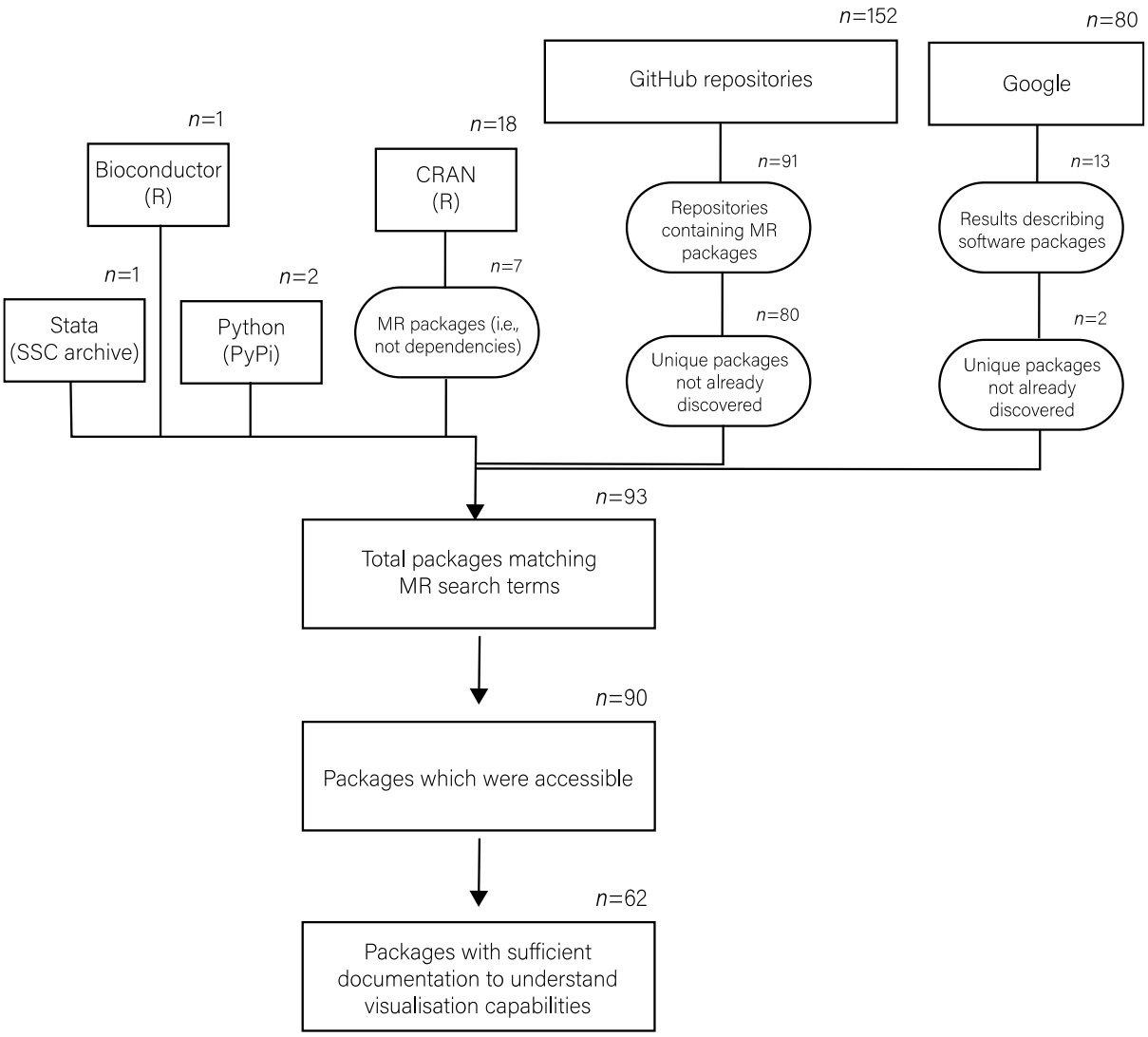


Figure 4.4 Systematic search process for software. 'Unique packages' refers to packages remaining after pre-screening where duplicates were removed.

4.3.2 Results

Table 4.5 Software included in data extraction

(Author/) Software title	Link	Database	Language
MendelianRandomization	https://cran.r-project.org/web/packages/MendelianRandomization/index.html	CRAN	r
mrbayes	https://cran.r-project.org/web/packages/mrbayes/index.html	CRAN	r
mr.raps	https://cran.r-project.org/web/packages/mr.raps/index.html	CRAN	r
MRPC	https://cran.r-project.org/web/packages/MRPC/index.html	CRAN	r
hJAM	https://cran.r-project.org/web/packages/hJAM/index.html	CRAN	r
GLIDE	https://cran.r-project.org/web/packages/GLIDE/index.html	CRAN	r
iva	https://cran.r-project.org/web/packages/iva/index.html	CRAN	r
hdose	https://pypi.org/project/hdose/	PyPi	python
pysumstats	https://pypi.org/project/pysumstats/	PyPi	python
mrrobust	https://ideas.repec.org/p/boc/usug17/14.html	Stata SSC	stata

MRCIEU/PHESANT	https://github.com/MRCIEU/PHESANT	GitHub	r
rondolab/MR-PRESSO	https://github.com/rondolab/MR-PRESSO	GitHub	r
peteryin21/py-merp	https://github.com/peteryin21/py-merp	GitHub	py
jrs95/nlmr	https://github.com/jrs95/nlmr	GitHub	
gtx	https://rdrr.io/cran/gtx/man/gtx-package.html	GitHub	r
jrs95/nlmr	https://github.com/jrs95/nlmr	GitHub	r
jingshuw/GRAPPLE	https://github.com/jingshuw/GRAPPLE	GitHub	r
gqi/MRMix	https://github.com/gqi/MRMix	GitHub	r
mikelove/mrlocus	https://github.com/mikelove/mrlocus	GitHub	r
jiazhao97/BWMR	https://github.com/jiazhao97/BWMR	GitHub	r
cnfoley/mrclust	https://github.com/cnfoley/mrclust	GitHub	r
carloscinelli/mrsensemakr	https://github.com/carloscinelli/mrsensemakr	GitHub	r
n-mounier/MRlap	https://github.com/n-mounier/MRlap	GitHub	r
liusy-jz/MRBIGR	https://github.com/liusy-jz/MRBIGR	GitHub	perl
remlapmot/ivonesample mr	https://github.com/remlapmot/ivonesamplemr	GitHub	stata
DAGitty (web)	http://www.dagitty.net/	GitHub	website
DAGitty (R)	http://www.dagitty.net/	GitHub	r / rstudio
DiagrammeR	http://rich-iannone.github.io/DiagrammeR/	GitHub	r
askieslinger/MRTool	https://github.com/askieslinger/MRTool	GitHub	
CYShapland/BESIDEMR	https://github.com/CYShapland/BESIDEMR	GitHub	r
LaiJiang/CIVMR	https://github.com/LaiJiang/CIVMR	GitHub	r
daskrohn/RBD_SMR	https://github.com/daskrohn/RBD_SMR	GitHub	r
william-denault/CFMR	https://github.com/william-denault/CFMR	GitHub	r
vaskarageorg/SCA_MR	https://github.com/vaskarageorg/SCA_MR	GitHub	r

	https://github.com/zhonghualiu/MRMiSTE		
zhonghualiu/MRMiSTERI	RI	GitHub	r
hunt-	https://github.com/hunt-		
genes/harmonize_dosage	genes/harmonize_dosage	GitHub	py
SharonLutz/reverseC	https://github.com/SharonLutz/reverseC	GitHub	r
m1sorenson/mendelian_r	https://github.com/m1sorenson/mendelia		
andomization	n_randomization	GitHub	
junghyunJJ/ggmend	https://github.com/junghyunJJ/ggmend	GitHub	r
	https://github.com/datastorm-		
visNetwork	open/visNetwork	GitHub	r
Sung-Bong-	https://github.com/Sung-Bong-		
Kang/Simple_MR	Kang/Simple_MR	GitHub	r
peteryin21/MeRP	https://github.com/peteryin21/MeRP	GitHub	py
danieliong/MRPATH	https://github.com/danieliong/MRPATH	GitHub	r
kehongjie/ImagingMR	https://github.com/kehongjie/ImagingMR	GitHub	r
WSpiller/RMVMR	https://github.com/WSpiller/RMVMR	GitHub	r
	https://github.com/XiaofengZhuCase/IMR		
XiaofengZhuCase/IMRP	P	GitHub	r
Gizmodiat/AMANDE	https://github.com/Gizmodiat/AMANDE	GitHub	shell
jalabrecque/MRchecks	https://github.com/jalabrecque/MRchecks	GitHub	
remlapmot/bpbounds	https://github.com/remlapmot/bpbounds	GitHub	r
	http://bristol-medical-		
MR Visualisation Tool	stat.bristol.ac.uk:3838/MR-Vis/	Manual	website
Cytoscape	https://cytoscape.org/	Manual	website
Epigraph DB	https://epigraphdb.org/	GitHub	website
D3	https://d3js.org/	GitHub	JavaScript
			r, python,
Tetrad	https://www.ccd.pitt.edu/	GitHub	application, website

	https://cran.r-project.org/web/packages/ggdag/vignettes/intro-to-ggdag.html	GitHub	r
ggdag			
dagR	https://cran.r-project.org/package=dagR	GitHub	r
shinyDAG	https://apps.gerkelab.com/shinyDAG/	GitHub	website
DAG program	https://hsz.dife.de/dag/	GitHub	application

Content analysis

I recorded the visualisation features present in each software. The results of this data extraction, for 33 software with visualisation features, are presented in Table 4.6.

Table 4.6 Results of data extraction. A one (1) indicates a capability was present. * = These software were general visualisation utilities which could be programmed to produce any type of visualisation.

Software (link)	**Scatter plot**	**Line chart**	**Forest plot**	**Manhattan plot**	**Bar chart**	**Dendrogram**	**Undirected graph**	**Directed (cyclic) graph**	**DAG**	**Heat map**	**Radial graph**	**Any ***
MendelianRandomization	1	1	1	1								
MR.RAPS	1											
MRPC						1		1	1			
hJAM	1									1		
GLIDE	1											
mrrobust	1	1	1	1								
MRC IEU PHESANT	1											
gtx-package		1										
MRMIX	1											

mrlocus	1										
BMWR	1	1			1						
mrclust	1										
mrsensemakr		1									
MRBIGR	1		1	1	1						
DAGitty (web)									1		
DAGitty (R)									1		
DiagrammeR							1	1	1		
CIVMR	1										
reverseC	1										
visNetwork						1	1	1	1		
MRPATH	1				1						
RMVMR	1	1									
Mrchecks		1									
MR Network Visualisation tool										1	
Cytoscape							1	1	1		1
Epigraph DB								1			
D3							1	1	1	1	1
Tetrad									1		
ggdag									1		
dagR									1		
shinyDAG									1		

5

5.1 Transcripts of interviews and discussions

5.1.1 Interviews with researchers

I conducted semi-structured interviews with three researchers at the MRC Integrative Epidemiology Unit who use MR. All researchers responded voluntarily to an advert put out asking for discussions around making an MR game. Following the interview I invited to review my notes from their meeting and clarify my understanding of the points they made. I asked each researchers some prompts including:

- "what's your research area? … aims? … motivations?"
- "do you experience any barriers to your work?"
- "what would you think of an MR game?"
- "what data from a MR game could be useful?"

Ppt 1

Research area:

MR methodology / statistics

Research aims / objectives:

Finding pleiotropy robust methods & Finding principled way of identifying & selecting valid IVs (SIVIVE framework)

Motivations for work

It's about the challenge; the puzzle of it and the mathematical problem solving

Barriers to work

Challenges: What is the best way of doing an analysis? There's so many people saying different methods are right so which should they use? Everyone's [MR researchers] looking to methodologists to tell them which to use

Finding the time is difficult; as is recruiting people with the right expertise

Also, trying to get people to agree on which methodology is right

Potential applications of games

Games can help people understand how methods work. Junior MR researchers will want to apply different methods and they're more likely to do this if they understand them.

For every method, it would be nice to make an easy to understand illustration and for people to use it in practice

[An example of a game where a place on a leader board is used to motivate engagement with educational tasks]

Knowledge from games

Can use a game to explain complex methods to non-experts

If a game is used as an analysis capture tool, then I can present the results and prove I couldn't have got these results without [it]. [But the game wouldn't be the focus]

It's like Facebook; they make it fun for you to tell them all about yourself. No one would have filled out a massive questionnaire.

Other topics:

Games as analogies (points illustrated)

Mastermind game as analogy for statistical process / challenge of selecting valid IVs; asking questions about the validity of SNPs and using that information to narrow down correct possibilities; process of testing individual points, in a sequential process of verification / challenging validity

However, in real life it wouldn't work because we don't know whether IVs are valid

The diagrams [drawn; mastermind and graph] are equivalent under the bonnet but I would prefer the graph. A journal wouldn't want it to be presented as a game because journals want it [the underlying methods] to be transparent. [Without a game] it's clearer to show what you've done.

I don't think that ultimately [in my area] if people are going to develop a new methodology, there is no shortcut to getting a PhD in maths etc. Making a game won't solve this. [Games are a] way to engage

Post-meeting further thoughts / clarification: Of course, if the main barrier to you finding an answer is computational (e.g. you have to try out 1 Billion Billion special cases to find the best one) and a game encourages thousands of people to explore some of those cases on their own computer, then the `crowd' may indeed help you to uncover a solution you couldn't easily get on your own. This principle is used in physics (https://www.skyandtelescope.com/online-resources/list-citizen-science-projects/). My point was that `citizens' (and perhaps this is a better word for lay people you are looking for) aren't going to come up with new methodology using a game

Illustrations

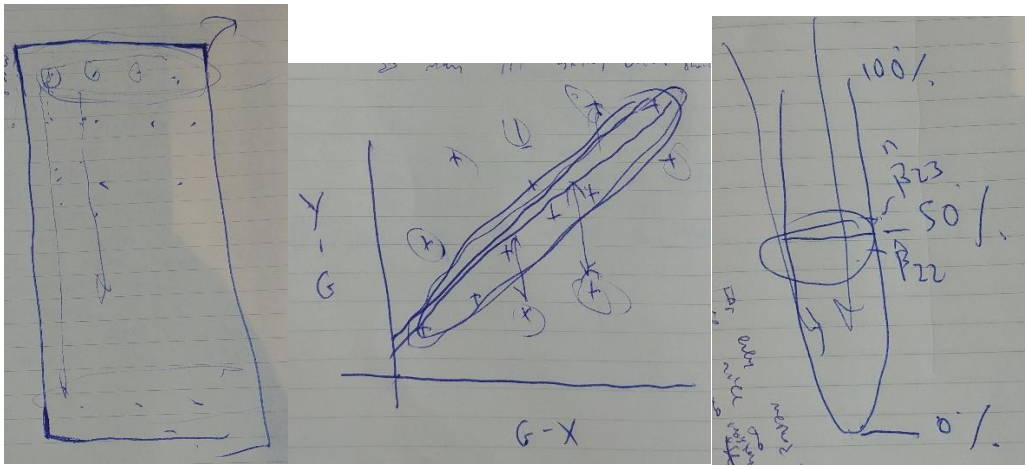


Figure 1. From left to right: A) Illustration of the game mastermind, as applied as an analogy to the scientific process of IV validation B) A graphic representation of the same context, but where in mastermind you would ask questions about the IVs to determine their validity, the mathematical solution in the line graph is working out their distance from the line 9residuals) or identifying outliers to individually explore research into C) graphic representation of a different concept, a median based method of estimating the average strength of multiple IV's effects in order to estimate their collective validity

Ppt2

Research area: Applied MR & Causes of cancer using MR

Research aims: To identify whether fatty acids increase chance of cancer

Research motivations: Different motivations on different layers. Main motivation is want to prevent cancer but also enjoy collaborating with people to answer questions. For me, MR is fun!; Enjoy the nitty gritty (e.g., getting data & performing analysis); my favourite part is writing the paper

Topics of discussion:

Barriers and difficulties in research

Although we're in a post-data world; [the availability of] data is the main limitation for my tertiary prevention MR research

Interpreting results is a challenge (multiple testing, genetic confounds, horizontal pleiotropy); challenges are increasingly methodological

Applying a game to MR

[After laying out an idea for an MR game involving crowdsourcing to verify links in the MR EvE data model] … [the concept of an MR game] Feels abstract

In a hypothesis free way you can test relationships and find ***an*** answer very quickly [implying it isn't necessarily the correct answer]

Exploration is fun – scientists like exploring. You have a question and suddenly want to go to your computer and explore it. Danger is that it doesn't necessarily give you the right answer; susceptible to data mining

[Scientists] work through hypothesis-driven methods [and this game would be more suitable for hypothesis-free / data-driven research]

Similar: IARC reviews; IARC ask a question every year and scientists sit down in a room & try to figure it out& give a score at the end for how cancerous [they've concluded] it is

[Games could be useful if] a question is too complex to address, or you want to bring outside knowledge

Could be useful if, out of many possible pathways [a risk factor could act], you want to find out which one is causal

Using additional information to find links. If I was asked to [validate MR EvE links] for my job I would conduct a literature review and run statistical tests. But as a game you're not doing all of that, it's not the whole pipeline just a specific task.

Would be interesting if you had a disease but you don't know what the cause is. Would be exciting because it's unknown; it's exploring. Or flip it around and have risk factors which you have to find out what they cause.

Could be interesting to look at examples where the risk factor might cause one thing but reduce risk for another e.g., Telomere length on CHD and cancer

Useful to separate out MR and public health impact as two separate games

Data / knowledge from games

[Reliability of the data output from a game] depends on the reliability of the crowd source

A game could help to prioritise research questions

Has to be an interesting question; gives motivation

Being grounded in a real problem makes it more interesting. If you give it context; the way you frame the question affects motivation to answer the question

Other points:

Post-meeting further thoughts / clarification:

Some starting points for the game could include

start from the disease perspective. You select one disease outcome and wish to identify all possible causes of the disease.

start from the exposure perspective. We have this exposure and want to know everything it causes.

what are the mechanisms or mediators of a known exposure-outcome association?

Framing/context determines motivation

e.g. for (1) this is a terrible disease and we don't know any causes or ways of preventing it. Or we know some causes of disease (e.g. smoking) but wish to identify additional ways of preventing that disease (because people who give up smoking still at high risk of lung cancer).

For (2), various organisations are advocating that we intervene on this exposure. What would be the consequences of intervening on this exposure?

For (3), we know X causes Disease but we don't have a practical way for intervening on X. If we could identify mediators of X on disease perhaps we could intervene on those.

Outputs of 1,2 or 3 could feed into a second game based on public health modelling of any identified interventions.

E.g. how many cases of CHD would we prevent and how many cases of cancer would we cause?

What are acceptable tradeoffs? Game could involve varying the parameters to see what happens.

Ppt3

Research area: Investigating the effects and causes of gut microbiomes (microbial cultures in the body e.g., intestine). Genetic dietics; MR

Research aims: Can MR be used to investigate the effects and causes of gut microbiomes? Add to the microbiome literature and question the nature of causality

Research motivations: Driven by interests; definitely my main motivation but I am also in the best place to ask these questions; [the IEU] is really good at evaluating methods

Barriers and difficulties in research:

In cohort studies, they give need to assess lots of people cheaply, so they give questionnaires but there are problems with this because people lie, people are asked what they are 3 months ago, and people will eat something super healthy [on the days they are asked about; they'll change behaviours]

Applying a game to MR

[Games are] more interactive; my least favourite thing to do for public engagement is giving a talk, it's not interactive and you can't ask the audience if they're following / they can't ask you questions.

Similar: Art-Scientists collaboration art installation at Hamilton House. [Games are similar to that], they're a platform for discussion *Post-meeting thought / clarification:* (this was called Creative Reactions)

The game has to be advertised as fun and helpful [i.e., scientifically valuable]. The game needs to be fun & I don't think you'll get a high rate of response if it's just admin; [this is] not interesting to researchers.

Challenging people would do it [engage them]

Public engagement is different, [this knowledge game] has to contribute to something which will help you with your research.

Similar: Hackathons get people together & playing a game & helping something

Similar: [Example of using leader boards to motivate schoolchildren to engage with an ALSPAC hand grip strength public engagement event]. [With a leader board] people were more engaged & will do it [will engage more]. It can generate discussion; is an immediate enticing point for engaging people

Data / knowledge from games

Can envision a game of education [such as the boardgame(s) another IEU researcher had made and contributed to, which educate members of the public about what the IEU does] but don't know how a game would produce an output

I think I've not seen a game used for generating output

But I think that'll be useful for people to see; people here [at the IEU] are happy to try out new things, if people see that it works they'll be happy to try it out

Not concerned with crowdsourcing as a concept

My main concern is the validity of the game; I would want to see some kind of validation or testing of its ability to produce data without bias

5.1.2 Discussions with attendees at the MR conference

I presented a poster at the MR conference (2018) and had discussions with student and researcher attendees about their broad thoughts on an MR game. I followed this up with a short discussion of my conclusions and interpretations with members of my lab group at the University of Bristol, Dynamic Genetics (marked with participant IDs not starting with C). I present below important quotes arranged into thematic categories (raw transcripts available online: https://osf.io/5c9kz/).

Scepticism about a conference game about MR for networking and learning

C2: Interesting; initially confused how you would incorporate MR EvE data into a game

C5: Has to be an efficient way of learning (for academics to use it for learning)

C5: For networking, would a game be over and above useful over key words searching conferences to find relevant people (implication: it would need to be for people to use it)

C6: Content of the game has to be intellectually simulating

C8: Would need some incentive / rewards / facilitator to get people involved though e.g., the most social person gets a reward; "Use immediate rewards (i.e., as well as an end of game prize) to get people playing, that's how all this gamification works isn't it? [I] don't know what that reward would be though…

C9: What does this do for the MR conference? (i.e., they doubted success / value)

C10: Could be good at the drinks reception; ply people with alcohol and they may be more receptive

C12: Understanding MR is different (it's a harder goal to achieve)

S1: Also how will this work for people at different levels. I think everyone is interested in networking but this is different for people at various career stages e.g. a more junior role e.g. PhD might be more interested in just attending to tick the present at a conference box than networking whereas someone more senior might be more interested in networking and then very senior might be more interested in talking to people about current projects they already have as opposed to new

collaborations. I'm massively genralising there, but I think the point was how would you make this appealing to everyone?

M2: Networking would be super cool how do you add data to that though

Some optimism about learning goal

C4: A game would make learning MR practical; practical experience helps learning

Optimism about networking goal

S1: I can definitely relate on finding networking awkard and I'm sure there are many people who just do not do this and would find a game a much better way of doing this.

C5: Networking is a more likely goal because it is a fun way to do it, and everyone hates it; this is more workable

C7: Would really appreciate it if I could play a game with my team; "Would help to bond people together, playing together

C7: Networking: Would definitely bring people in across backgrounds; "If you see something interesting (e.g., interesting interaction) you can discuss that with people; "C7: They may be more motivated to network (outside of my academic group), I normally stay with people in my area (i.e., don't leave my social / academic group)

C8: Would help facilitate networking

C12: Networking is an easy goal because it's the purpose of conferences

C13: Physical challenges with games: wifi, would need it to play the game but there isn't always wifi

M2: Could you have an app that you download which you give it your details on who you are and who you're looking to meet? Give your area and match people; "Would be nice to have the excuse to go up to people"

M1: [a game which matches you to people you want to network with] That's a good idea

O1: It's tinder for academic purposes (sentiment echoes a similar remark made by us in a project meeting)

Game could be relaxing

C7: Would say ideas you wouldn't otherwise say, in a game environment because it's less serious; might help to relax people

C11: Can have a fun game about MR mis-practice; "here are the assumptions ive violated"; "*People would be more willing to talk about these things in an informal setting like in a game*

Underlying game data needs to be relevant to players

C7: Would also like to be able to select a specific outcome (i.e., choosing cancer instead of exploring lots of irrelevant ones)

C11: Understanding MR data may only work with early career audience members, not experts

Game timing has to be appropriate to conference

C9: Do you have enough time to play this at a conference?

Conferences should take advantage of technology

M3: Can use location and camera access app like pokemon go; can put a pikachu in an unpopular talk to people go to it

M4: Can use RFID tags to track people's interactions

C3: Can integrate the app into the wider conference ecosystem i.e., by requiring people to register

5.2 Paper prototypes

In prototyping the early three ideas for the game, I constructed paper prototypes to model game mechanics and concepts in a manner which was easy to iterate on and make changes between playtesting sessions.

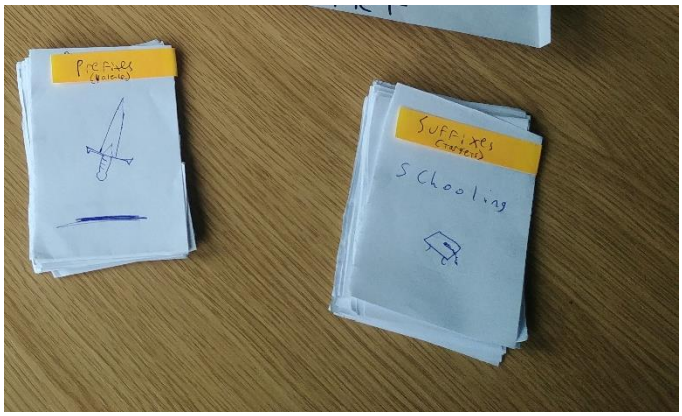


Figure 5.1 Paper prototypes represented in-game concepts, such as the factors in the network, using pieces made from paper, such as a deck of cards representing each factor. In this photo, the dungeon crawler game idea used two sets of cards to generate items in the game world with prefixes and suffixes. Prefixes, such as "a dagger of…" served to add a playful flavour to suffixes and effects such as "… increasing schooling". Different

encountered items were differently valuable in solving encounters where the player might be asked to achieve a goal such as increasing the education of the population.

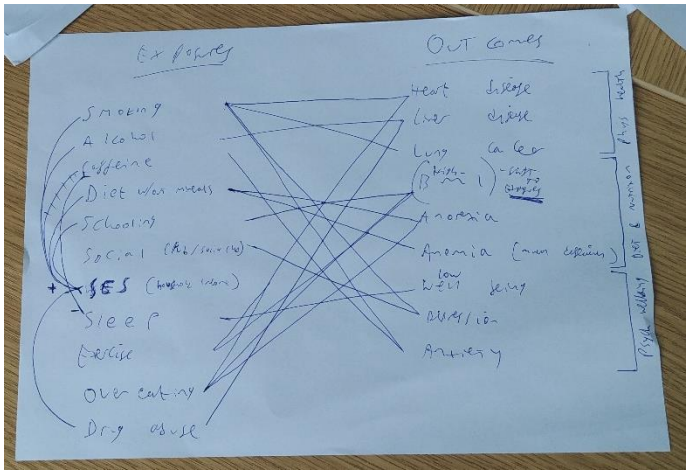


Figure 5.2 Representing network data in a paper game was difficult so one game idea, a simulation game, was not play tested as much as other ideas because it required complex calculations which took a long time to calculate by hand. This photo shows a manually drawn map recording relationships between factors in the network so that this could be used as a reference for faster determining the effects that players had when intervening on a given factor in the network.

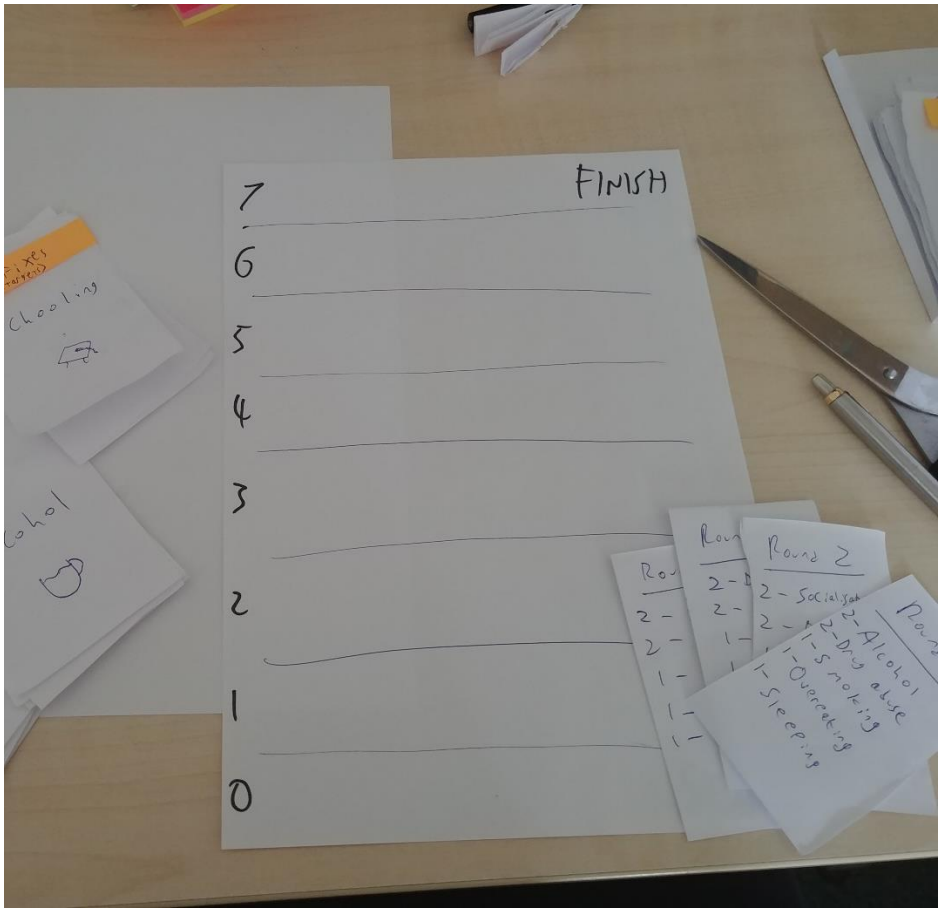


Figure 5.3 A game board made from paper allowed me to represent players on a board along with indications of their progress and the goals they had to achieve. In this photo, players of the fantasy MR idea progressed

upwards in the score board by correctly selecting the factors most likely to achieve a series of outcomes over four rounds.

5.3 Table of changes

This section documents the various changes scheduled during the development of my data game. Each of the three game prototypes underwent playtesting as well as the final data game prototype. These changes are documented in tables of changes which are presented for each of these games below. A coding scheme was used to represent the type of evidence I used to make various changes to the games in response to feedback:

- Exp - reason for change based on personal experience or knowledge of the domain
- Rep - reason for change is that this issue was often reported
- Eas - reason for change is that it is trivial to implement
- Ncr - no change required
- Nc - no immediate change required but monitor if issue re-appears and re-assess at that point
- Imp - reason for change is that this is important to achieving the overall goal of the game
- Ncon - reason for change is that it does not contradict existing evidence and theory

Additionally, changes were prioritised on a scale from 1 to 5 where higher numbers indicated lower priorities. In the final playtesting session for the final game, I checked with playtesters that my changes solved their issues and this is represented in a different format table.

Table 5.1 Table of changes for playtesting the narrative adventure prototype game

Feature	Feedback and changes	Reason for change	Priority
Playtesting session 1			
Setting	Changes: Post-apocalyptic setting where humans have forgotten medicine	EXP (expert)	3
Interventions	Feedback: Pro: Fantasy settings disarm researchers from thinking 'is this practical in real life'? (e.g., a magic potion enables you to reduce smoking rates) – magic items so you don't worry about how it caused a change, just that it did	NCR	1
Setting	Changes: Interventions change genetic presdispositions to exposures rather than affecting the exposures themselves	EXP (expert)	3
Interventions	Feedback:Need a way to force people to manipulate exposures other than the obvious and known	EXP (pers.)	1

	relationships (i.e., reducing smoking every turn to increase health) Changes: Procedurally generate interventions and randomly present them to players. Use a prefix and suffix system, where prefixes designate the diection of the effect (e.g., increasing), and the suffix designates the target of the effect (e.g., smoking), which then affects camp health as the outcome depending on the prefix and suffix combination		
Gameplay	Feedback: Pro: Functions as a real game	NCR	
Interventions	Feedback: Pro: Can scale complexity easily by adding/removing rules (e.g., 'items can now interact with each other and affect future/past interventions')	NCR	
Gameplay	Feedback: Pro: Naturally affords the use if many well-established game mechanics, such as random chance outcomes and deck building	NCR	
Gameplay	Feedback: Pro: Naturally affords progression systems (unlocking interventions / camp improvements)	NCR	
Setting	Feedback:Con: Fantasy settings look less scientific or useful	NC - Monitor and see if becomes a REP / IMP concern	
Data model	Feedback:Con: These games normally have a low-moderate level of simulation complexity, we would require greater	NC - Monitor and explore ways to make more complex	
Playtesting session 2			
Winning	Feedback:No winstate Changes: End goal is to reach the end of the board	EAS	1
Setting	Feedback:Winning has no meaning Changes: End story is that you are on a journey to migrate your camp to a safehaven for other survivors, to a large city	EAS	2
Difficulty	Feedback:No lose state or stakes Changes: Lose state when the health of your camp is reduced to zero	EAS	1
Camp health	Feedback:Onedimensional (only health stat) was not interesting, every intervention was either good or bad with no inbetween or nuance; or strategy. Also did not allow for the teaching of a successful intervention having undesired consequences on another (e.g., reducing drinking increases displacement drug taking behaviours or overeating). Changes: Have multiple different health score	EAS	1

	types which are differentially affected by interventions: physical, nutritional and psychological health (e.g., overeating increases bmi but reduces anorexia, thus reduces physical health but increases psychological health)		
Data	Feedback:Data is not based on MR eve links Changes: Use real MR EvE data	NC - No data integration available until MR EvE is back online	
Playtesting session 3			
Difficulty	Feedback:Too easy Changes: Reduce the starting camp health stats from 4 to 3 out of 7	EAS, NCON	1
Interventions	Feedback:Had large effects and combined with chance, produced a system where big changes could happen and you could win/lose in an instant Changes: Reduce the element of chance by introducing a new challenge which gives an additional intervention to chose from if passed (and one less if failed), but retain large effects of interventions to keep game tension that every decision matters	EAS, NCON	1
Camp health	Feedback:Values exceeded maximum (i.e., 6 out of 5) Changes: Make physical health a finite resource which can be reduced but never regained, but start with one more, 4 out of 7	EAS, NCON	1
Winning	Feedback:Not satisfying, no sense of achievement/victory Changes: Make game harder to win so getting to the end is a guarantee	NC - already adressed by difficulty and camp health changes	
Interventions	Feedback:Predictable effects Changes: Mechanic where camp followers do not obey intervention, producing random effects (player sugested)	NC, EXP (pers.) - Introduces too much complexity at this point but should continue to be monitored.	
Challenge tiles	Feedback:Success depends on chance (i.e., intervention items found) Changes: Give more choice of challenges so if you miss one, you are less affected by chance	NC, EXP (pers.) - challenges give an element of randomness but that would take away from feelings of strategy and planning. Continue to monitor	

Movement	Feedback:Not interesting Changes: Could restrict movement using challenges you cant cross unless you pass them	EAS, NCON (challenges)	1
	Feedback:Not interesting Changes: Could use dice (player suggested)	NC, EXP (pers.) - randomness can be disruptive to gameplay unless meaningfully implemented	
Interventions	Feedback:Surprising effects of interventions was confusing	NC, EXP (pers.) - surprising effects were part of the dataset but effects should be clear so continue to monitor.	
Playtesting session 4			
Challenge tiles	Feedback:Not very interesting Changes: Could hide whether you can pass the challenges, or the effects. Could also have a penalty for failing (i.e., being able to chose an intevention from a choise of one fewer cards increasing the chance of an unfavourable outcome)	EXP (pers.), EAS, NCON	2
Interventions	Feedback:You would rarely pick a bad intervention so sometimes you are just stuck picking the only good card, not very interactive Changes: Could separate the positive and negative items and encounter them seperately rather than chosing from them; introduce good and bad item tiles / encounters	EXP (pers.), EAS, NCON	2
Interventions	Feedback: Funny combinations of names Double and triple negative names were confusing (e.g., the dagger of poor sleep which reduces the reduction in sleep quality and therefore reduces the increase in high anxiety...) Changes: Revise naming to avoid using negatives (e.g., insomnia instead of poor sleep)	EAS, NCON, REP	1
Difficulty	Feedback:Still too easy Changes: Reduce starting health from 4 to 3 out of 7	EAS, NCON	1
Interventions	Feedback:Experts may know the relationship between outcomes and so the mechanic of hidden intervention effects wouldn't work	NC - will be revealed by later testing with expert audience	
Difficulty	Feedback:It is predictable to see if you've won if you're doing well because something would have to go very wrong for you to lose near the end with lots of health. Especially as you can chose to skip risky	EAS, NCON	2

	encounters Changes: Add random events to challenges which can make large changes and must be met before the end of the game		
Coop	Feedback:Was a signle-player experience, would be fun to play with others Changes: Add a second player character onto the board	NCON	1
Interventions	Changes: Could have morality dilemmas where you have to sacrifice one part of the health of your camp to improve it in another (player suggested)	NC - adressed by having to chose negative interventions now	
Player character	Feedback:Would be more motivating to have characters Changes: Ability to chose from different characters to play as (player suggested), perhaps with different strengths / weaknesses, or starting stat points	EXP (pers.), NCON	3
Playtesting session 5			
interventions	Feedback: confusing double negatives; confused by combinations, working out what the effects are Changes: could have directionality on the back of the effect, rather than on fromt of item	IMP	1
idea	Changes: different QR codes in different rooms, conference could be dungeons; need to scan qr code on cards; (player suggested)	NCR	
interventions	Feedback:whats the reason for items, people might chose their assortment of item to target; minus items Changes: clearly separate items from their effect on the exposure; could just have minus/plus effects;	IMP	1
interventions	Feedback:booklet of non-nutirtional diet makes sense but beads of alcohol don't. you look for meanign in the items. So "leaflet would be helpful but I don't know what beads of alcohol would do". item descriptions were distracting Changes: Adressed above	NCR	
idea	Changes: could have task of making supersoldiers by using MR and progressively improving your troops	NCR	

Table 5.2 Table of changes for the epidemic game prototype

Feature	Feedback and changes	Reason for change	Priority
Playtesting session 1			
Data model	Feedback: NCR	NCR	
Setting	Feedback: NCR	NCR	
Gameplay	Feedback:Con: Requires extra layers to turn a simulation into a full game	NCR	
Playtesting session 2			
Logic	Less clear associations / relationships / data so can't quickly guess (player suggested)IMPShould	IMP	Should
Interaction	Digitalise it? Doing things digitally may be more fun, as it would have a modern touch and it may be easier to engage (player suggested)IMPMust	IMP	Must
Interaction	(will come from further gamification)IMPMust	IMP	Must
Rules	NCR	NCR	
	Introduce some abiguity; make the choices a bit tougher to make, using some randomness or less clearly linked (player suggested)EASShould	EAS	Should
Interaction	(will come from further gamification)NCR	NCR	
Randomisation	Add randomisationEASShould	EAS	Should
Feedback	(will come from further gamification)NCR	NCR	
Interaction	(will come from further gamification)NCR	NCR	
Playtesting session 3			
Symptom upgrades	Simplify to 1 branch categoryEAS1	EAS	1
Modifier upgrades	Remove & replace with reworked modifiersNCR1	NCR	1
Stats	Simplify and conform to 1-10 scalesEAS1	EAS	1
Playtesting session 4			
Infectivity counter	store in memory and use immediately to infectEAS1	EAS	1
Upgrade points	Remove upgrade couters from game for simplicity; store in memory and use immediately to upgradeEAS1	EAS	1
Lethality counter	store in memory and use immediately to infectEAS1	EAS	1
Symptoms	Introduce more infectivity symptoms earlier, offset lethality til later. Upon review, not necessary changeNCR1	NCR	1
Cure progress	Could… Remove rule: remove but compensate for cure progress loss. BUT… Cure progression should give pressure independent of players' movesNCR	NCR	
Infecting	Introduce rule: can only infect from adjacent infected country but not razed oneEAS1	EAS	1

Difficulty	Introduce more cure progress points. Previous change, adding 1 cure progress after first death makes game harderNCR1	NCR	1
Modifier upgrades	Could require an initial investment (i.e., +1 cure progress) before you can start lowering progress, means players have to be thinking a turn aheadEAS2	EAS	2
Playtesting session 5			
Symptoms	Could produce a random sample of symptomsNCR	NCR	
Setting	Can swap to healingIMP1	IMP	1
Win condition	NCR	NCR	
	Could be cool to have a bot to generate symptoms automatically in the network, could say I want a game on smoking and it auto-populates itNCR	NCR	
Map	Another space to map non-communical disease (game is a network, may be way to map in non-comunical disease). However, at this point we can consider this for the future since it is v low priorityEXP (OD), NCOR4	EXP (OD), NCOR	4
Gameplay	NCR	NCR	
Playtesting session 6			
	Can swap to healingIMP1	IMP	1
Infecting countries	infectivitity, deaths, and in-game rewards should be proportional to populationEXP, NCON2	EXP, NCON	2
Fun	NCR	NCR	
Difficulty	needs more levels; or more symptoms (more complex game elements - addressed with change below)NCR	NCR	
Symptoms	Add more symptoms, more complex interactiosn betweens symptoms and infections to make it harder to predict game courseEXP, NCON2	EXP, NCON	2
Cure progress upgrades	and more risky upgrades like the +1/-4 cure progressEXP, NCON2	EXP, NCON	2
Co-op	Idea - Combine powers together to achieve the same goal, having half of the puzzleEXP, NCON1	EXP, NCON	1
Co-op (competitive)	Idea - for comeptitive, can play off of some countries having more population than others, giving compeittive edge; through certain copmbinations / routes you might get a ncie coutnry to infect tahts very large. i.e., starting at other ends of the map. Could kill off others coutnries, creating barriers for expansion - but inccurrs a cost such as increasing noticabilityEXP, NCON1	EXP, NCON	1

Table 5.3 Table of changes for the fantasy MR prototype game

Feature	Feedback and changes	Reason for change	Priority
Playtesting session 1			
Coop		NCR	
Data model		NCR	
Data model		NCR	
Gameplay	Feedback: Con: Requires more work to turn a 'fantasy football-like' game into a full gameFeedback:Con:	NC	

	Requires more work to turn a ‘fantasy football-like’ game into a full game	
Data model	Feedback: Con: Does not fulfil behavioural sub-objective ‘facilitate interaction with simulation’ since you are not manipulating data toFeedback:Con: Does not fulfil behavioural sub-objective ‘facilitate interaction with simulation’ since you are not manipulating data to	NC
Playtesting session 2		
	Feedback: not all cards had the same direction of effectnot all cards had the same direction of effectChanges: force all cards to have a positive or negative effect on the target outcome	REP (everyone agreed)
	Feedback: didn’t understand why points changed between rounds (why some cards no longer scored, like players perfomring in fantasy football scoring one week but not the next so they get a point the first week but not the next)didn’t understand why points changed between rounds (why some cards no longer scored, like players perfomring in fantasy football scoring one week but not the next so they get a point the first week but not the next)Changes: introduce variables in each round; need some kind of updating epxlanaitonf or why values change, unclear why values change - why didn't I win. Could do as policy game, pick policies with ebst evidence base and have evidence shift (player suggested)	REP (everyone agreed)
	Feedback: wasn't clear what you learn from the gamewasn't clear what you learn from the game	REP (everyone agreed)
	Feedback: feels like players are isolated and just hoping next card will help youfeels like players are isolated and just hoping next card will help youChanges: needs to be a emchanic where things you do can inhibit people; or trading cards with eachother, or uno trump cards (player suggested)	REP (everyone agreed)
	Changes: if point is to show how different methods can lead to different results, can have a double ending where one is based on realty rhough process but truth in real world is different; double learning goal, that things are valiable but also what's most acceptable knowing we can't map onto real answer	REP (everyone agreed)
	Feedback: kite board, with different outcomes, rather than just depression but alternatives. Might find that tohers' cards pull you to center, or unforseen consequences based on depression specialty. Adds complexity to chose first couple roundskite board, with different outcomes, rather than just depression but alternatives. Might find that tohers' cards pull you to center, or unforseen consequences based on depression specialty. Adds complexity to chose first couple rounds	NC - meaning behind this statement lost and I was not able to clarify it with playtesters

Table 5.4 Table of changes for final playtesting of the data game

Feedback (and fix for Playtesting session 2)	Priority (session 2)	Status (session 2)	Issuer
Playtesting session 1			
Data (categorisation) - Are these categories to do with MR? I don't understand. It's confusing			P1
Visuals - P2 - Scrap the wriggling, it's confusing when everything changes. P4 - The moving isn't useful, it's like putting it in a box and shaking it. P3 - Could slow the animation or have a replay function (P2 agreed).			P2, P3, P4
Visuals - P4 - The colour of the edges is not clear what it means. P3 - The black on red is giving me a headache.			P4, P3
Kinetics - When you drag a ndoe it should stay there, particularly since edges want ot hide between nodes			P2
Visuals - Too many options are confusing			P2
Visuals - [About the node growing after an intervention and this being confusing] It's getting bigger does that mean its worse?			P4
Visuals - Need to factor in colourblindness			P3
Visuals - I would like to see summary statistics			P1
Visuals (views) - [About the node view] Why is this niceer? How is this view any different from before?			P2
Visuals (views) - [After making an intervention which only worsened health accidentally] P2 - Now we made a mistake, how do we learn from it?P3 - If something unexpected happens, you wonder why it is? P2 - Need to have a way to drill down into things. Nodes should be a pie chart made up of the different traits making up the overall category node value.			P2, P3
Data - Should have real ranges for values and sliders. When you alter one value you can see what changes on the values. (P2, P3 agree)			P1, P2, P3
Visuals - Could use stock-market like changes & up/down arrows or red highlights to indicate changes. (P2 agrees) and adds that the flexibility of this approach is that it can apply to all data; it is also something that msot people are familiar with even if they dont have stocks			P3, P2
Kinetics - Should swap around the left and right buttons because intuitively increase/decrease are the other way round in your mind. & Should take into account the valence of what you are doing (good/bad)			P1
Data - Binary traits can only be increased by 1, either they have it or they do not.			P1
Data - Not many traits make sense for intervention. "Is this a problem?": P2 - Perspective of pilocy makers is what can I do? SO outcomes are the end goal but not the intervention. P3, P1, P4 - Disagree: Think it's interesting to look at outcomes.			P1, P3, P2, P4
Gameplay - Could put investment into different policies. (P3 agrees)			P2, P3
Visuals - P2 - Needs to be more simple. For exmaple, science museum games can be very simple, give you a few variables and can produce different solutions from it. #Similar to my idea for combining intervention options into a few selectable options which affect multiple nodes with one button click. P2- Could also have an advanced mode where it shows all the nodes, not just categories to interact with.			P2
Gameplay - P4 - Could chose the outcomes/exposures for your game. P3 - It would be like setting up a workspace for an experiment, selecting variables. P4 - could have hypothesis free and hypothesis driven modes. P2 - i would choose to collapse the whole bio block because they don't mean anything to me			P4, P3, P2
Data - Linear links HAVE to do the same thing on each propgation. What it did before it must do again, by definition.			P2

Audience - For engagement tool: People might think that this data is a fact. be careful that you don't teach the people the wrong things.			P1
Gameplay - Starting conditions for be interesting, either modelling the UK or exploring different countries.			P3
Visuals (views) - Would be interesting to see sub populations. and the effects of covariables			P3, P2
Data - Given the data, the only way you can model individual level data is to do it as a population. David Spiegel gave a talk on representing deaths in charts using people figures. Could work as a comparison of the population with and without intervention. (P1 & P4 agree)			P2, P1, P4
Scenario - Could have an alien world setting to dissuade people from thinking this is an accurate model of reality			P1
Data - The aspect of time is ignored, have to be careful about how it is represented, especially with generational effects and survival analysis. also can be used to justify feedback loop ignored (takes time)			P2
Visuals (views) - Could give explanations of nodes on hover			P1
Playtesting session 2			
It was not clear that you click rather than drag on traits to intervene on them (44159). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - Tested by reviewer	P1
It was not clear that to dismiss one of the tutorials you had to press the fire icon to continue (44159). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - "	P1
It wasn't obvious to you when you were meant to 'Enact a Policy' and why (44159). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - "	P1
Tutorial does not explain arrow thickness (44160). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - Tested by me	
The game ends at level 6 which takes about 20 minutes to finish putting a 20m cap on playtime (44160). Fix: Give players the option to replay it and increase level cap by 1	3	Fix published -	
You were not clear about what to do with the goal and level sections of the user interface (44159). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - "	P1
It is not clear that the username is just for display purposes and is not saved (44159). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - "	P1
The interactive visualisation shows the game level UI which is confusing as it serves no purpose (44159). Fix: Auto-hide UI and only show in game	3	Fix published - Tested by me	P1
After unlocking the ability to make additional intervention(s) it is not clear that you have to make more interventions before the enact policy screen is shown (44158). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - "	P1
Players can ignore the game tutorial and interact which can lead to lost information (44159). Fix: Ask players toclick through the tutorial dialog boxes in the tutorial	3	Fix published - "	P1

Dialogs would sometimes go out of the screen and would not fit on the page (44159). Fix: Revise how dialogs are drawn on the page	3	Fix published - Tested by me	P1
The relationships in the data were still confusing because the player brought in their own knowledge (44160). Fix: Improve typewriting in the tutorial to better warn players about weird relationships	3	Fix published -	P1
The introduction blurb is a lot to read and participants might skip it (40507). Fix: Condense and improve typewriting to better communicate key ideas in tutorial	3	Fix published -	
The leaderboard was not clear whose policy was the best (44161). Fix: Move best policy above players if it beats it (a la real leaderboards)	3	Fix published -	
Tell players that effects propagate and they will be shown this (44161). Fix: Tell players that effects propagate and they will be shown this	3	Fix published -	
Further clarify on the leaderboard page why you have not won (44161). Fix: Make clarity on you did not improve trait message in leaderboard	3	Fix published -	
Encourage players to make more interventions at level 3 (44161). Fix: Encourage players to make more interventions at level 3	3	Fix published -	
Line thickness is not a reliable indicator for judging the effects of interventions (44159). Fix: Show line thickness in legend key (a la mirana)	3	Fix published -	P3
You were not clear on why you were shown the leaderboard, and what information was displayed on it (particularly when you achieved the best policy and it didn't show that any player beat you) (44159). Fix: Rework leaderboard for legibility	3	Fix published - "	P1
Instad of encouraging players to reach the end of the game, the message 'level 6 here I come' was confusing because it sounded like suddenly were at level 6 (44159). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - Awaiting issuer comments	P3
Double clicking on a node triggers an error message which is confusing and disrupts gameplay (44159). Fix: Improve typewriting in the error message to better communicate that only a single click is required to make an intervention	3	Fix published - Tested by me	
It was not clear from the tutorial what the effects of interventions are (44158). Fix: Improve typewriting in the tutorial to better communicate this	3	Fix published - Issuer(s) reported fix effective	Multiple
Users could give personally identifying data in the username (44158). Fix: Users assigned anonymous IDs	3	Fix published - Tested by me	
It was not clear that the tutorials should be completed before continuing with the game (44159). Fix: Add focus effect to tutorial dialogs to catch players eyes	3	Fix published - Tested by me	P1

It was not clear on the login page that you were meant to type in a username (44160). Fix: Improve typewriting on the login page to better communicate this	3	Fix published - Tested by me	P1
It is not clear what some traits are (e.g., eveningness and neuroticism) (44158). Fix: Add a help section	3	Fix published - Tested by me	Multiple
Players often feel lost and don't know what to do despite instructions from tutorial (44160). Fix: Add a help section	3	Fix published - Tested by me	
Showing other players could be a deception issue (44161). Fix: Remove competition with players	3	Fix published - Tested by me	P2, P3, P4
First intervention screen interrupts propagation of effects (44161). Fix: Delay tutorial to pop up later after player has seen some effects shown	3	Fix published - Tested by me	
Add my email to help section for players to email without referring back to qualtrics (44161). Fix: Add my email to help section for players to email without referring back to qualtrics	3	Fix published - Tested by me	
Participants can make interventions in the test using the interactive visualisation (44161). Fix: Participants can make interventions in the test using the interactive visualisation	3	Fix published - Tested by me	
It takes a long time to show all effects (44161). Fix: Consider shortening effects animation	3	Fix published - Tested by me	
The music abruptly stops on starting the game (44158). Fix: Revise startup code	3	Fix published - Tested by me	P3
The game does not initiate after pressing play (44158). Fix: Revise startup code	3	Fix published - Tested by me	Multiple
The error messages are jarring and interrupt gameplay (44158). Fix: Revise error handling	3	Fix published - Tested by me	
On reset sometimes the nodes would not reset and be transparet / coloured still (44160). Fix: Revise the reset code	3	Fix published - Tested by me	
The sound effects are inconsistent and only happen on some interactions (44159). Fix: Revise audio handling	3	Fix published - Tested by me	
The icons are not visible always against the background (44155). Fix: Add white background to images	3	Change not scheduled - Tested by me	P3
Being assigned a goal which you cannot achieve is frustrating - "In normal situation, starting a game like that, I would've given up and closed" (44159). Fix: Revise goal setting code	3	Change not scheduled - Tested by me	P3
It is not clear how many interventions have been made and how many are remaining (44159). Fix: Add intervention count UI element	3	Change not scheduled - Tested by me	
The planets no longer show up in the background (44158). Fix: Revise view code	3	Change not scheduled - Tested by me	
Fix box scaling (44161). Fix: Fix box scaling	3	Fix published - Tested by me	
Nodes go off the edge of the screen sometimes (44136). Fix: Revise drawing of SVG on page	3	Fix published - Tested by me	Multiple
The propagation algorithm does not display effects in the order which would be most intuitive (44136). Fix: Identify the most intuitive system; Revise propagation code	3	Change not scheduled - Tested by me	P3

5.4 Player feedback from the experimental study conducted in chapter 6

As part of a questionnaire handed to participants in the study, I will conduct in chapter 6, participants were asked to recall their experience in their own words. They were provided a box to given answer to “In your own words please describe your experience with the software. For example, did you have a strategy? Did anything prevent you from achieving what you wanted? Did you find any effects memorable? Did you have any opinions about the presentation?”. Answers to this question are presented below along with themes I assigned in a brief thematic analysis to understand the kinds of experiences players had with the game. Counts of feedback in each of the theme categories I identified are presented in Figure 5.4.

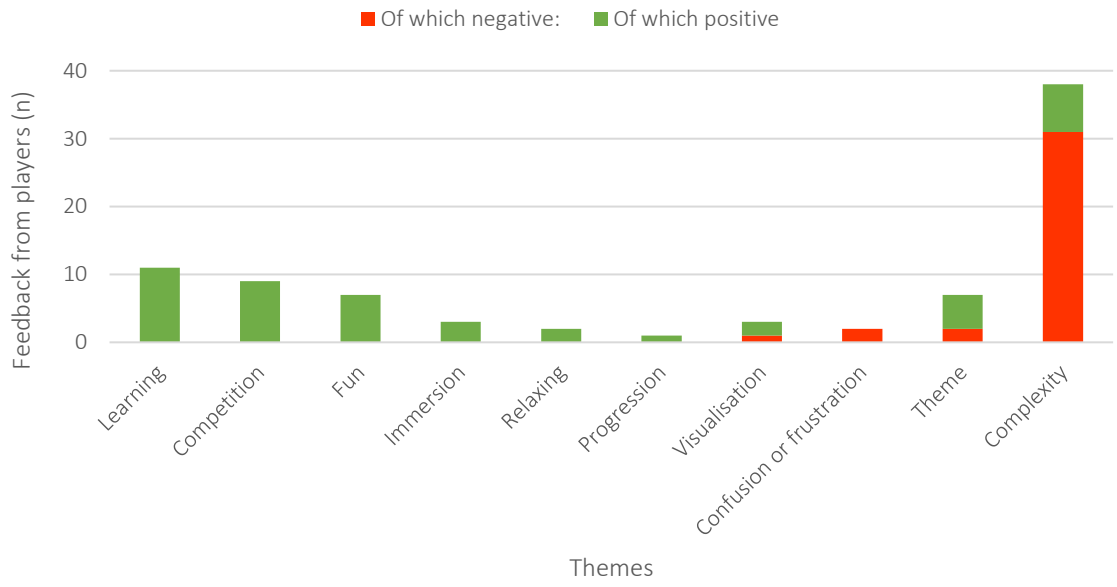


Figure 5.4 Analysing feedback from 90 players of the game reveals they learned about complexity in the game (complexity & learning), enjoyed competing with an AI agent (competition, fun) and found the audio-visual theme entertaining though distracting (theme, fun). 82 responses were flagged with at least one of 10 themes, so some participants gave feedback that was flagged with multiple themes, and some players feedback was not flagged with any theme.

Table 5.5 Player feedback from playing the game in the experimental study (chapter 6)

#	Themes	Feedback
1	Visualisation; specific effect	I didn't followed any specific strategy and nothing prevented me from achieving what I wanted.I liked the fact that I could see how each component had an effect

		on something else from a game.I could find more memorable the effects that I was already aware of like education reduces smoking.
2	Scoring (confusion)	I think insomnia and loneliness are the most prevalent factors that will affect others. Sometimes I am not entirely sure why the computer did better than me and how come the computer got the higher percentage affected. I do not quite understand how the percentage affected is worked out.
3	Aesthetics; fun	I enjoyed the game and the way it was presented, I tried to have some sort of strategy but ended up just choosing things I thought might be correct.
4	Aesthetics; fun; learning	Wonderful presentation, very interesting and fun to interact with the game. I began remembering what trait lowered and increased other traits. The soundtrack to the game is very well done.
5	Immersion	It was a somewhat immersive experience that distracted me for a while.
6	Structure	i tried to lower/increase the things i think would help me increase/decrease the things the game asked me to
7	Competition (social learning); aesthetics	As I gained more intervention uses per round the AI would use them in a different way to me that would bring a greater effect so I tried to follow it's strategy. The presentation was nice and I like the sound design.
8	Complex	I found the presentation of the stimulus quite complex t first but was able to adjust to it rather quickly. My strategy was to look into the trait that needed changing and then figure out which other trait would best lead me to do this
9	Complex	At first I found it difficult to understand, but I developed a strategy pretty quickly.
10		Some of the questions would appear multiple times which improved how well I memorised the answers
11	Complex	i had not a clue what was going on i am not going to lie to you
12	Complex	I did not really enjoy the software because I do not think it was clear enough about how to win. This may of been because I did not have strategy though.
13		I think it was a good game and I was motivated but found it difficult on a phone as couldn't really see the instructions which made knowing how to play it a bit difficult
14	Immersion; complex	It was captivating but also quite confusing to know what the goal is and how to achieve it. However, it was interesting to see what affected what in the game and it was quite fun. The instructions could have been clearer and the presentation as well because it was all overlapping so there was a lot going on.
15	Pregression; Aesthetics; reloading	It was strange considering the interventions for each goal didn't always make much sense. The first time I played the game it crashed/stopped working so I lost all my progress which was irritating. I liked the level up part, it made me motivated to keep going and progress through the levels. The actual presentation of the game was well-made and exciting, I can see how it would be an enticing method of learning
16	Complex	I feel like the pulsating icons/arrows distracted me when trying to work out the interventions. I did not have a strategy.
17		I had no strategy, just look to see which arrows were most prominent etc. It made the relationships between the factors a lot more memorable.
18	Competition; Aesthetics	It was interesting getting to grips with the simulation I quickly learnt how to reduce behaviour and it became easier to spot pathways to reduce patters. The presentation of the game was really interesting and I would definately play again, I especially liked the interaction with the computer that I was competing with something to make a positive effect.
19	Aesthetics	I liked the sounds and animations
20		I just tried to follow the links to see which one would have the bigger outcome and affect on the named factor
21	Complex	I found the software a little confusing to use. I just clicked a lot because my original strategies were undermined.

22	Complex	I found it slightly difficult to understand at the beginning. I didn't particularly have a strategy, I just tended to choose the interventions linked by the arrows. An effect i found memorable was the impact education had
23	Frustrating	I found it quite frustrating because the computer kept winning by intervening with insomnia, depression, loneliness and worry even when I could not see an arrow connecting the goal to these factors, which made me very very frustrated
24	Complex	At first it was a bit hard to navigate since I felt like the game's design is not the clearest. I did get the hang of it after a couple of rounds and noticed myself trying different strategies. The effect's would have been memorable if the game's presentation was a bit more clear.
25		I would place the icons slightly further apart as it felt slightly cramped
26	Simple; Aesthetic	I was trying to figure out how to use the map to choose the best way to solve the issue. I thought the presentation of the software was done really well, very clean and nice to follow
27	Prior knowledge	I found that sometimes I tried to use the arrows to make my decision and then the computer came up with a more successful intervention that I could have come up with if I'd use common sense instead. So it successfully brings your attention to interventions that could be used in our lives.
28	Complex	I found it confusing what I could and coulnd't clicke and how some of them weren't linked to certain goals
29	Competition; Complex	i learned a lot from playing the game.. i was especially surprised by things that i had no idea about such as "insomnia lowers down heart disease". i didn't really have a strategy and at first, i was confused on what i was supposed to do. but then i got the hang of it and i did pretty well. there were times that were abit frustrating when playing the game because computer beat me to it in some parts which made me feel dumb. but hey, i learn something new everyday. i also liked the design of the game software. the colours were bright and the display of the symbols/images were clear. overall, i had a fun playing the game
30	Relaxing; Aesthetic	Tried to find the root cause but often found that the best solution was the one directly influencing. Education became my go to when selecting solutions most of the time. The presentation was fun and the soundtrack noise was soothing/ enjoyable
31		I just mapped out the arrows and looked for the biggest effect
32		analysis of all the arrow colour and directions in relation to the specific task I had at each level
33	Aesthetic; Complex	I liked the space theme and my task. However, it was hard to tell what I had just done. For example, I'd forget which trait I just intervened on and it was tricky to figure out why I was successful or not in the task. The icons would move and be highlighted which was confusing and often frustrating when I was trying to finish the task.
34		I think some of the effects were memorable however I didn't really have a strategy
35		Some of the parts of the game I didn't fully understand but I did get it by the end.
36	Competition (social learning)	I found the effects memorable, and it would be a good way to educate younger children. Initially I was not as confident with finding answers, but after the computer got it right a few times and I could see the response it became easier to work out what kind of thing I should be picking.
37		Found it difficult, but it was memorable
38	Complex	Was confusing, ended up doing trial and error for most of the game
39	Simple	It felt a bit simplistic and obvious what is expected of you. I also found it a bit visually confusing, not always being able to see what the arrows were pointing at. Having to watch the effects spread through the network after each intervention got a bit annoying.
40	Complex	Was interesting, harder than I thought

41	Competition (social learning)	I followed the paths logically and learnt from the computer's answers
42	Complex	I couldn't rechoose another option to intervene with if I had changed my mind, and when some objectives were to lower a specific trait, intervening raised them.
43		I thought it was a fun strategic game that introduced me to different ways to promoting good traits and diminishing bad traits that one may encounter in daily life.
44	Complex; gamers	i found the layout very confusing but this could be because im not used to playing games at all
45	Complex	I had to play the game for a while until I completely understood what the arrows did the the effect of interventions, for example, for bad things the intervention would affect the other things in an opposite way to the arrow colours. If an arrow would reduce something, doing an intervention on a negative variable, would increase that thing.
46	Complex	I didn't really have a strategy and it took me a while to get the hang of it
47	Complex	I found it quite confusing, when the computer beat me in a task it was not really explained why. However, I developed a sort-of strategy of trying to find the thickest lines or correlation between traits and work with those, or look at traits that had another linkage in common. The presentation was confusing, but I think that was the point.
48	Prior knowledge	Thinking logically helped me the most, and after a while you tend to start memorising what affects what which makes it easier
49	Aesthetic	Looked at the arrows to see associations and payed attention to the computers choices. The software was clean and well presented but I did quickly get bored as it felt like I was doing the same thing over again with no real purpose.
50		I found not having a set strategy useful as I just went along with the experience and allowed myself to get lost.
51		The software was really intriguing and an interesting way of approaching mental and physical health. It was difficult to see some of the arrows as were small or hidden by other arrows.
52	Fun; visualisation	I really enjoyed the software, especially the layout showing how interlinked all the various traits are and the multitude of effects each intervention can have.
53		initially I did not have a strategy, but as the game processed I started producing patterns for the same issues. at the start the software was confusing, I didn't grasp the concept especially once upgraded to making two interventions but as I was exposed to it more, I got the hang of it. it was fun as it was highly interactive and engaging
54	Complex	I didn't like the game and I don't feel like I learned much, the whole thing was a bit confusing for me and I was at some point only looking at arrows and how they should work instead of thinking about the actual effect and learning from the game
55		it was a good game although i found it challenging sometimes.
56		Enjoyed finding how to reduce/increase certain traits although I didn't have a proper strategy apart from trial and error. I clicked the wrong button once which prevented me from achieving what I wanted.
57	Aesthetic	I liked the presentation
58	Learning	started to memorize what influenced what and used that in future strategies, very engaging
59	Complex	I found the system confusing but was useful to have the computer there. I wasn't able to click education
60	Complex; Aesthetic	I did not understand how the game worked, so by the end I was just playing it randomly. The game itself looked very interesting but the instructions need to be clearer.
61	Tutorial	The tutorial got me to be familiar with the task and increased my accuracy in later levels.

62	Complex	it was very complicated
63	Complex	I did have a strategy to use the arrows to guide me. It was frustrating because I didn't understand why certain policies would have certain effects. Loneliness on wellbeing seemed odd. Presentation was alright.
64	Fun	Engaging. It actually interested me when different interventions effect different traits etc
65	Complex	Tried to follow the instructions as best as I could but got confused a few times
66	Competition (social learning)	I made a mental note of the computer's winning responses and made sure to choose those options when they came up again. It would have been nice to be able to have a closer image of the game, or an interactive zoom option. Additionally, It would have been motivating and interesting to see the scores tallied between the computer and the player. I thought the music was a good balance of soft and continuous, providing an immersive experience without being distracting.
67		I used trial and error, but it was useful
68	Complex	I found the software challenging to use. As i thought i was understanding it I would then go a step back and not understand
69		I used the arrows to guide which intervention to select. I also thought of which would apply most in the real world. I liked the interactive diagram it was easy to follow
70	Aesthetic; Progression; Competition	I really enjoyed the presentation of the task and the music and overall appearance. I felt a sense of achievement when I did beat the computer and this did motivate me to do better on the next task. I was a bit confused on some tasks that I did not manage to have an effect on the objective even though I felt the options I has selected would have an effect.
71	Aesthetic (didn’t fit)	I tried to reduce one item in order to have a domino effect to achieve the desired effect but this did not always work. Education was generally a go to as it affected many different traits. I though the space background was both entertaining and distracting/didn't fit the theme of the task.
72	Complex; Learning	At first I was quite confused as to what I was doing but after level 2 I understood more. I found that for worry, depression and loneliness I understood it more as I know what increases and reduces it from personal experience, which also means I found these effects the most memorable.
73	Aesthetic	i really liked the music and the look of the software - i found it very engaging and it worked very smoothy. my strategy was to follow the arrows as much as i could. it was sometimes difficult when you couldn't select a certain one. i also tried to memorise the times the computer beat me and did what the computer did if the same goal came up again.
74	Complex	Some effects did not seem correlated at all, and I did find myself confused at first.
75	Complex	I had a strategy to choose the factor I thought would be influential first, then others that link to that. I did find the design of the game a little confusing, though.
76	Complex	I found the game difficult to understand, and couldn't work out a strategy that worked for me.
77	Complex	To be honest, I found the software a bit confusing. Sometimes I won even though I accidentally lowered/increased a specific trait/behaviour when I was supposed to do the opposite. Some effects did not really make sense to me and sometimes it was not possible to affect a trait in a specifc way - or it was and I did not understand how to. I did like the concept of competing against a computer and having to find the best interventions etc, though. I feel like if I understood the game better, I would have memorized the effects. I always judged based on the causal relationships. As soon as it was possible to select multiple interventions, I tried to see what affects a specific trait directly and indirectly (and how).
78	Competition (social learning)	I just tried to think of what would make the most impact and kept note of what the computer used. I waas surprised how some things linked more than others.
79		EDUCATION ALWAYS IMPOVED INTELLIGENCE

80	Fun	The format was a good motivator
81	Strategy; Aesthetic	In terms of strategy, I tried to make the biggest changes in the least moves possible. I liked how the effects weren't always that obvious and you had to think about how slight changes can make an impact on more than one thing. thought the game layout and system was great :)
82	Aesthetic (didn't fit); Complex	It took me a little time to get the hang of the game, and sometimes it was a challenge to influence something indirectly. Placing the game in space is an interesting choice, I'm not sure why it was done, although it is an appealing background, but perhaps does not aid clarity. I like the music.
83	Immersion; Complex	I found it very mesmerising but challenging at times, I pressed the icons that I thought was related to the task.
84		I have a strategy when the experiment go into later. I don't know how to conduct it at the beggining
85	Competition (social learning)	My method mostly involved memorising the computer's tac tics. Thought some of the game was a bit confusing e.g., what does eveningness mean? I was also mislead/confused about using interventions for emotions - i didn't realise you could intervene on 'wellbeing', 'depression', 'loneliness' etc. until a lot later in the game (though learning). It was interesting to see which combinations were the most effective.
86	Complex; Aesthetic	I wasn't clear on why I was progressing so there was no strategy. I liked the look and feel just wish I'd be able to understand it better.
87	Complex; Visualisation (complex)	I found it difficult to understand how to play, and what was asked of us. The graphics didn't make too much sense, I was confused by the arrows and I wasn't sure what I had to do. I was discovering the game as i was playing. Overall, it was stimulating because I have learnt some things about public health policies and which are best for a specific issue.
88		I tried pick the interventions that are most effective
89	Complex	At first, it is a little bit confused about the blue line and read line as well as how these traits are influenced by each other. I tried to remember some answers (best strategies of computer) and analyse it, that is really helpful. I am impressed that alcohol could improve evenings, education would increase many good traits. One thing that really confused me is the interaction of increase and decrease, as these lines are very complicated.
90	Nothing at stake	didn't feel anything was at stake so had no trouble guessing without mercy
91	Learning; Humor; Aesthetic; Calm	To start with I was intervening on things that were closely linked to the objective, but after a while I realised that intervening on more indirect things such as loneliness would have a bigger knock-on effect. I thought the presentation of the game was quite nice, visual design was pleasing, there was humour in the instructions and the music was calming.
92	Complex	I found it distracting that the effect of one intervention would flash all over the place while I was trying to choose another
93		I kind of form some strategy but it doesn't always work. About level 5 I finally realize some policy are memorable.
94	Complex; Aesthetic; Fun	I tried to remember the answers. I was a bit confused about which traits impacted on each other as there was a lot of detail in the diagram. But the presentation and sounds etc made it enjoyable.
95	Fun; Aesthetic	INtriguing - music made it more of an adventure. Fun to play with.
96		I looked at the strength and sequence of arrows
97		It is a good idea to play with the game to in order to increase fancy and fun.

6

6.1 Power analysis

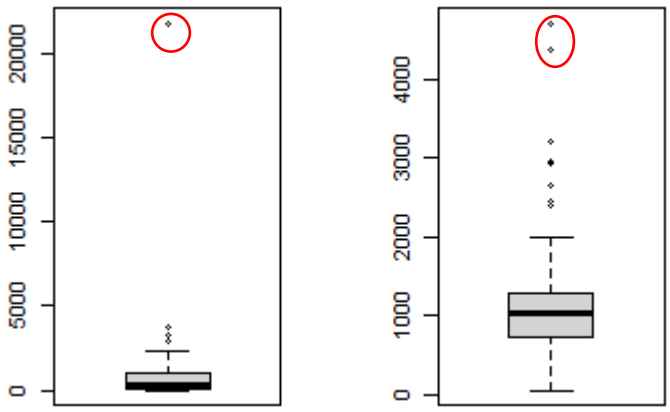
Table 6.1 Power analysis to determine sample size required to detect a minimum meaningful difference in learning outcomes at a range of sensitivities

Learning assessment score			Sensitivity		Required n participants	
Control group mean*	Standard deviation*	Experimental group mean	Power	Alpha	Per group	Total
13.84	+/- 1.77	14.84	80%	0.05	49	98
				0.01	73	146
				0.001	107	214
			90%	0.05	66	132
				0.01	93	186
			95%	0.05	81	162
				0.01	112	224

Note: Power calculation for sample size required to observe a 1 question/score difference in an MCQ assessment of learning (considered a minimum meaningful difference). * = Estimation of assessment scores obtained from pilot testing (n=6). Power was calculated using ClinCalc software (https://clincalc.com/), version Jul 24, 2019, based on conducting two-sample t-test analysis on continuous outcomes across two equally sized comparison groups.

6.2 Excluded participants

This section contains two figures which help illustrate how some participants were excluded. Figure 6.1 demonstrates that three participants either spent exceedingly long durations using the software or on the assessment. This was taken as an indication of inattention. Figure 6.2 demonstrates that nine participants spent exceedingly short durations completing the assessment, and received poor scores, which indicates they were inattentive as well.



Time spent using the software (s) Time spent on the assessment (s)

Figure 6.1: Left: Boxplot of time spent (s) using the software. Right: Boxplot of time spend (s) on the assessment. Excluded participants indicated with red circles (durations over 80 and 60 minutes respectively).

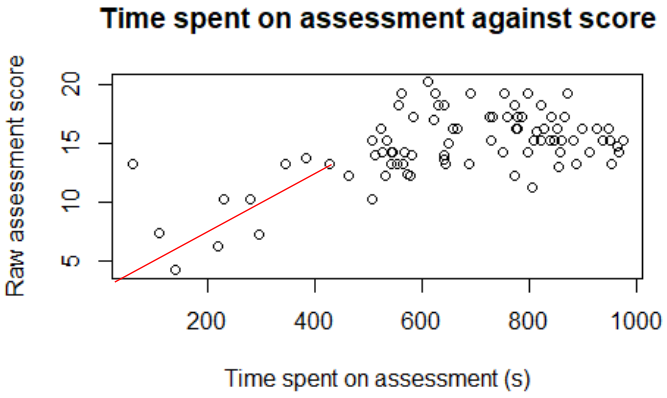


Figure 6.2 Participants who spent extraordinarily little time on the assessment also performed poorly indicating participants who spent less time to complete the assessment were also careless. A red line illustrates the trend among nine participants who were excluded (durations under 7 minutes = 420 seconds).

6.3 Questions in the MCQ learning assessment

All questions in my MCQ are listed below. Correct responses are marked with green.

1. Q1 Select from the list below all the traits present in the network (5 out of 6 correct answers required):
   - ❑ Correct answers: Education, Heart Disease, Wellbeing, Eveningness, Diabetes, Coffee intake
   - ❑ Incorrect answers: OCD, Public transport use, Phone use, Videogaming, Weight, Diet, Drug use , Social anxiety
2. Q2 Does insomnia share a **direct** relationship with eveningness?
   - ❑ Yes/No
3. Q3 Please indicate how different arrow colours in the visualisation represent how different relationships affect the prevalence of traits:

- ❑ Red arrows represent: Increases / Decreases
- ❑ Blue arrows represent: Increases / Decreases

For the next few questions please consider the direct relationship between coffee intake and intelligence

4. Q4 What is the direction of this relationship?
   - ❑ Intelligence affects coffee intake / Coffee intake affects intelligence

5. Q5 Compared to the effect size of other effects in the network, how large is this effect? (Remember the visualisation does not show relationship strengths)
   - ❑ Larger than average / Smaller than average
6. Q6 Is this relationship responsible for an increase or decrease?
   - ❑ The effect is responsible for an increase / decrease
7. Q7 Imagine that a fast food restaurant opened up and people started going there so much that they put on weight. This would increase their BMI. What effect would this have on smoking? (Remember the visualisation does not show relationship strengths)
   - ❑ Smoking would increase by a relatively large amount / relatively small amount
   - ❑ Smoking would reduce by a relatively large amount / small amount
8. Q8 Would either of the following public health interventions directly increase exercise? Please consider only the immediate direct effects of interventions.
   - ❑ Increasing intelligence would increase exercise / Increasing education would increase exercise / No, neither of the interventions above would increase exercise
9. Q9 Which intervention would most reduce heart disease directly? Please consider only the immediate direct effects of interventions.
   - ❑ Increasing exercise / Reducing diabetes / Increasing education
10. Q10 Would the effects of an intervention to increase education be on the general mental and physical health of the population? For the purposes of this question please treat increases in coffee intake, BMI, eveningness, smoking, neuroticism as bad even if this is not intuitive to you (you can see whether traits are good/bad by using the Trait key under the help menu of the visualisation)
    - ❑ Its effects would be only good / mixed / bad
11. Q11 Which trait causes the greatest effects on other traits in the network? (Remember the visualisation does not show relationship strengths)
    - ❑ BMI / Education / Intelligence / Depression
12. Q12 Imagine a scenario where students go to University. One might expect that education and alcohol consumption would both increase. What would be the combined direct effects of this on eveningness? Please consider only the immediate direct effects of interventions.
    - ❑ Increasing education and alcohol consumption would both increase eveningness / decrease eveningness / would cancel out and there would be little/no effect

13. Q13 Select the combination of interventions whose direct effects would most increase wellbeing. Please consider only the immediate direct effects of interventions.
    - ❑ Reduce neuroticism, depression, and insomnia / Reduce worry, depression, and insomnia / Increase exercise and reduce eveningness
14. Q14 Imagine that Universities were closed and the students just went home. This would reduce education. Given the example of the relationship between education and intelligence (shown above), what would happen if education was reduced?
    - ❑ Insomnia would increase / not be affected / decrease
15. Q15 What effect would reducing depression have on worry?
    - ❑ It would have no effect / reduce worry / increase worry
16. Q16 When considering the whole network of effects, including all indirect effects, does depression have an effect on coffee intake?
    - ❑ Yes / No

For the next two questions please consider the scenario: Imagine that a new brewery opened up and people started drinking more alcohol.

17. Q17 For this question we would like you to focus on only two of the effects of increasing alcohol consumption: it increases BMI and reduces education. Education then has a knock-on effect on BMI. We would like you to identify what this effect is and use this to estimate the combined overall effects on BMI:
    - ❑ Increasing alcohol and reducing education would both increase BMI / have opposing effects on BMI which would cancel out so overall there would be no effect / would both reduce BMI
18. Q18 For this question we would like you to focus on two different effects of increasing alcohol consumption: it reduces eveningness and reduces education. Education then has a knock-on

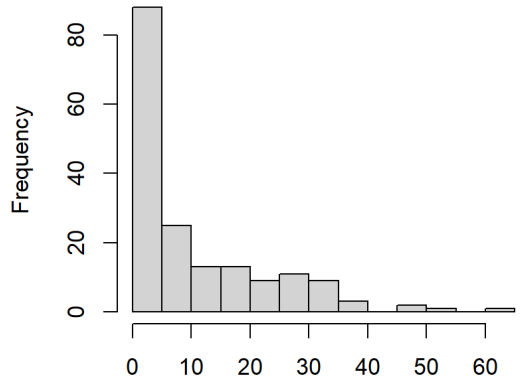
effect on eveningness. We would like you to identify what this effect is and use this to estimate the combined overall effects on eveningness:

- ❑ Increasing alcohol and reducing education would have opposing effects on eveningness which would cancel out so overall there would be no effect / both reduce eveningess / both increase eveningness

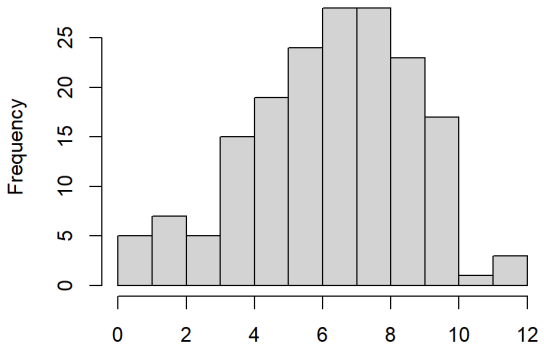
19. Q19 True or false: 'The size of effects decreases for each step in a pathway since each step is propagating a smaller proportion of prevalence change'
    - ❑ True / False
20. Q20 Select the intervention which would indirectly increase wellbeing:
    - ❑ Increasing depression / Decreasing loneliness / Increasing insomnia
21. Q21 Consider what would happen if depression increased. What is the furthest point in the network which will be affected by this?
    - ❑ Its effects will reach wellbeing / insomnia / coffee intake
22. Q22 If an intervention reduced depression, what would be the biggest source of change to wellbeing?
    - ❑ The direct effect of depression on wellbeing / The indirect effect of worry on depression / They would both be equal

6.4 Measurement distributions

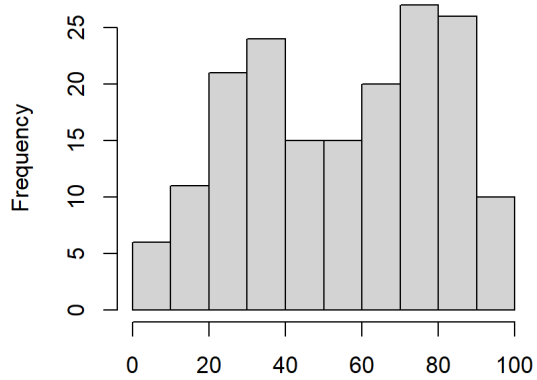
Duration of use
(minutes)



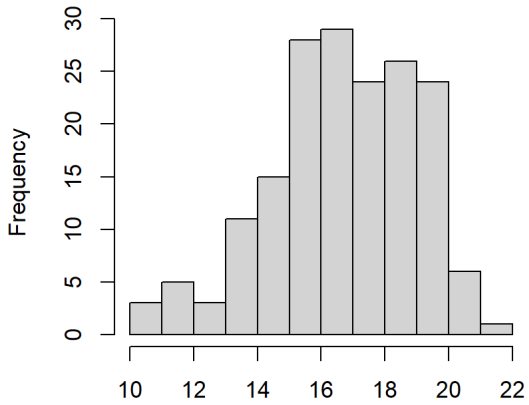
Playfulness
(0-12, higher values indicate more playful experiences)



Usability (%)



MCQ score (0-22)



In-game scores

(The individual scores for each intervention players made through the course of gameplay. 0-100%, higher values indicate interventions were more effective at solving problems.)

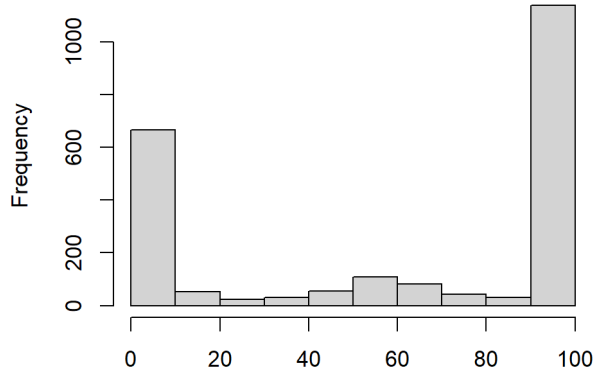


Table 6.2 Distribution parameters

	Bartlett test of homogeneity of variances			Shapiro-Wilk normality test			
	Control vs Game			Control		Game	
	K-squared	**df**	***P***	**W**	***P***	**W**	***P***
Duration of use	109.5	1	$1.29x10^{-25}$	0.719	$1.09x10^{-11}$	0.92	$4.59x10^{-05}$
Playfulness	0.3	1	0.605	0.97	0.045	0.93	$4.05x10^{-04}$
Usability	0.2	1	0.645	0.96	0.008	0.94	0.001
MCQ test score	2.4	1	0.120	0.965	0.021	0.94	$3.79x10^{-04}$
In-game score						0.72	$1.20x10^{-51}$

Note: In all cases Shapiro-Wilks testing was conducted to determine where non-parametric tests should be applied in analysis. Significant values (P<.05) indicate a sample is not normally distributed (none were normally distributed). *****Bartletts test of equal variance was used to ensure that, where Mann-Whitney U tests were applied, game and control conditions showed measurement responses with equal variance. This was important because this test makes that assumption, and all comparisons using this method passed the test by returning non-significant P values.

I also conducted an ordinal mixed effects linear model using a Cumulative Link Mixed Model (https://rdrr.io/cran/ordinal/). This is a method of estimating effects on an ordinal response, while accounting for mixed effects, such as players contributing multiple data points over the course of play. It makes the chief assumption that the relationships between responses across ordinal categories can be described with a single coefficient (the proportional odds assumption). This is to

say that the effect of an explanatory variable is proportional or consistent across the different levels of an ordinal dependent variable. The summary statistics in Table 6.3 demonstrate that odds ratios are relatively constant across each level of minutes in game (SDs 0.12-0.23), and trials (SDs 0.04-0.08).

Table 6.3 Descriptive statistics for odds ratios describing the likelihood responses would fall into ordinal categories of score (0,1,2) across all levels of predictors (minutes in-game and trials).

	Min	Max	Mean	SD
Minutes in-game				
0-1/2	0.00	0.43	0.29	0.12
1-0/2	0.00	1.18	0.44	0.18
2-0/1	0.62	2.12	0.88	0.23
Trials				
0-1/2	0.31	0.48	0.42	0.04
1-0/2	0.25	0.56	0.38	0.07
2-0/1	0.67	0.96	0.76	0.08

6.5 MCQ psychometric properties

6.5.1 Investigation of internal consistency

The psychometrics of these tests were then inspected. Each question in the assessments was designed to assess an individual learning outcome but mastery of core competencies should help participants answer multiple questions. Some internal consistency was therefore expected in the MCQ. Cronbach's alpha was used which is a measure of internal consistency which returns a score from 0 to 1 where higher scores indicate greater internal consistency (Tavakol & Dennick, 2011).

The MCQ showed relatively low internal consistency (a=.43) and so I inspected the individual questions in the questionnaire (Figure 6.3).

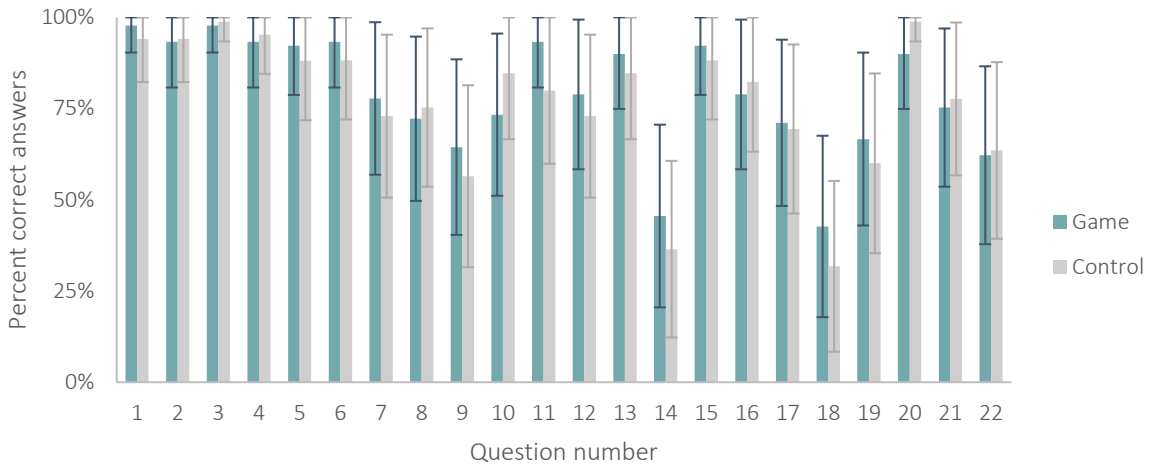


Figure 6.3 **Correct answer rate across questions in the MCQ by condition. Error bars represent standard deviations.**

Most questions appear to have shown a ceiling effect where correct response rates went frequently above 80% and this left little room for variance among participants. Consequently, differences on this questionnaire were smaller and more difficult to detect, ultimately requiring a larger sample to detect minute differences.

Additionally, scores on the twenty-five individual questions were often weakly correlated with the total score (correlation coefficients from .05 to .45) and with each other (Table 6.4).

Table 6.4 **Correlation matrix for answers on the MCQ**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Total	0.1	0.4	0.1	0.4	0.3	0.2	0.5	0.4	0.4	0.2	0.2	0.5	0.4	0.2	0.3	0.3	0.1	0.3	0.2	0.1	0.5	0.3
1		-0.1	0.0	0.2	0.0	0.0	0.0	-0.1	-0.1	0.0	0.2	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	-0.1	0.2	-0.2
2			0.0	0.0	0.0	0.1	0.2	0.2	0.0	0.2	-0.1	0.2	0.3	0.2	-0.1	0.0	-0.2	0.1	0.1	0.0	0.3	0.0
3				0.0	0.1	0.1	0.0	0.1	0.0	0.0	-0.1	0.1	-0.1	-0.2	0.1	-0.1	0.0	0.0	-0.1	0.2	-0.1	-0.1
4					0.1	0.1	0.4	-0.1	0.1	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.1	-0.1	0.0	0.2	0.3	0.0
5						0.0	0.1	0.1	0.1	-0.1	0.0	0.0	0.0	0.1	0.2	0.0	-0.1	0.1	0.0	-0.1	0.1	0.2
6							0.1	0.2	0.2	0.0	0.0	0.2	-0.1	-0.1	0.0	-0.1	0.0	0.0	0.0	-0.1	0.0	-0.1
7								0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0
8									0.1	0.2	-0.1	0.2	0.2	0.0	0.2	0.1	-0.1	0.1	-0.1	0.0	0.2	0.0
9										0.0	0.1	0.1	0.2	0.1	0.1	0.0	0.0	-0.1	0.0	-0.1	0.0	0.1
10											-0.1	0.1	0.1	-0.1	-0.1	0.0	-0.1	0.1	-0.1	0.1	0.1	0.1
11												0.0	0.1	0.0	-0.1	0.0	0.0	-0.1	0.2	0.0	0.0	0.1
12													0.2	0.0	0.1	0.1	-0.2	0.0	0.1	0.1	0.1	0.1

13	0.0	0.0	0.0	-0.2	-0.1	0.1	0.1	0.1	0.1
14		0.0	-0.1	-0.1	0.1	0.0	-0.1	0.0	-0.1
15			0.2	0.0	0.0	-0.1	0.0	0.1	0.1
16				0.1	0.1	-0.2	0.0	0.5	0.0
17					0.1	-0.1	0.2	0.0	0.0
18						0.0	0.0	0.0	0.0
19							0.0	-0.1	0.2
20								-0.1	0.1
21									0.1

Taken together this evidence indicates that this measure did not work as intended since a ceiling effect appeared which reduced variance among participants. I also investigated the possibility that my MCQ contained multiple sub-components measuring different constructs instead.

6.5.2 Exploratory factor analysis

Performing an exploratory factor analysis (EFA) did not reveal any components that predict large proportions of variance in MCQ scores. EFA was conducted using the Psych package in R (https://www.rdocumentation.org/packages/psych/versions/2.2.9). A questionnaire can be scored in many ways and one method is to create sub-scores by grouping certain questions together. EFA can be applied to the results of questionnaires to identify sub-scales which explain participants' responses. I started by creating a correlation matrix describing the relation of responses on each question in the questionnaire to each-other. Eigenvalues were calculated to describe the proportion of variance explained by grouping questions into sub-scales, referred to as "factors". These are visualised in a scree plot to determine the optimal number of factors to extract to explain variance in responses. Examining Figure 6.4, indicates that the optimal number of factors is between 3 and 7 factors, at which points relatively steep slopes on the graph indicate diminishing returns in terms of explaining variance in responses on the questionnaire.

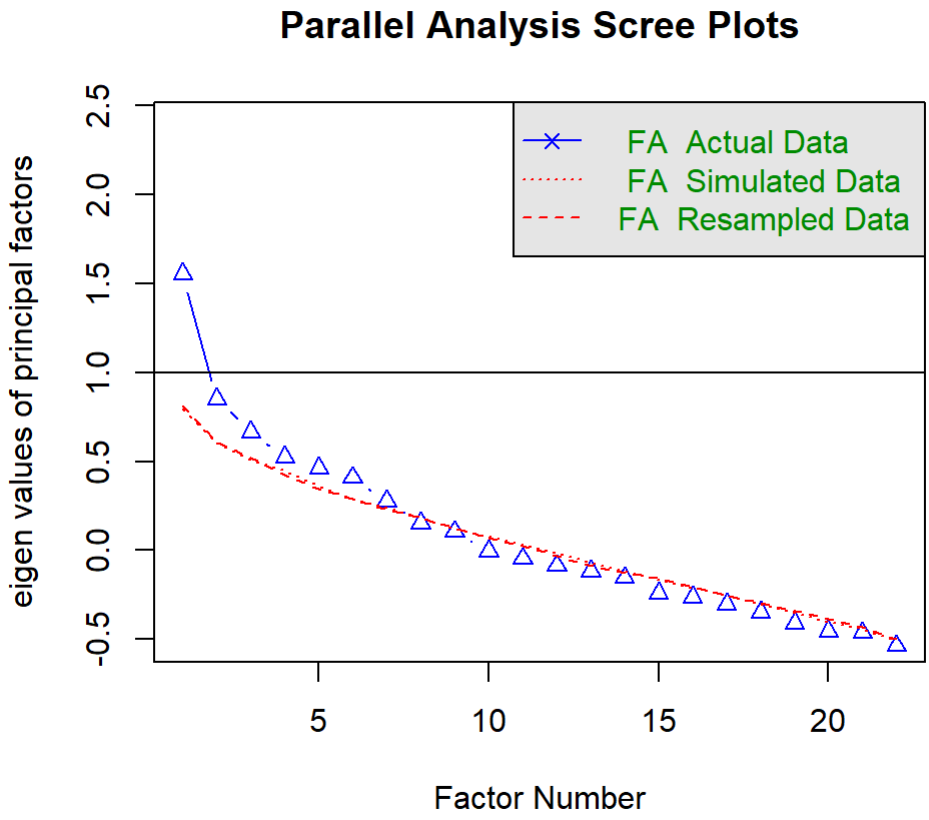


Figure 6.4 Scree plot of eigenvalues to determine how many factors to extract from the MCQ. Steep slopes in a scree plot can indicate points where adding more factors gives a significant diminishing return (e.g., from 3-4, and 7-8). Comparing eigenvalues with a randomly generated dataset of the same size provided additional guidance since eigenvalues below the red line indicated a lower proportion of variance explained than would be explained by random chance (calculated using the fa.parallel function in Psych).

I selected to extract 3 factors from the MCQ scores, as this is the first point in the scree plot where a plateau occurs, indicating the first instance of diminishing returns. I then performed EFA to identify groups of questions which explain the most variance in MCQ responses, and which may form sub-scales. Oblimin rotation was used to extract factors since responses on MCQ questions were related, and this is the typical procedure in this case.

Three factors were identified representing sub-scale factors comprising 2-4 questions in the MCQ. Factor 1 contained questions 2, 7, 12, and 13 and explained 11% of the variance in responses. Factor 2 contained questions 4, 15, 16, and 21, and explained 8% of variance. Factor 3 contained questions 8 and 11 but scored question 8 negatively; incorporating incorrect responses could not be a valid

way of measuring network knowledge so this was discarded as an inviable sub-scale. Therefore, only factors 1 and 2 were considered valid sub-scales though neither explained a substantial proportion of variance. This is demonstrated in Figure 6.5, neither component produced strong groupings of individuals that would have indicated that participants could be sub-grouped.

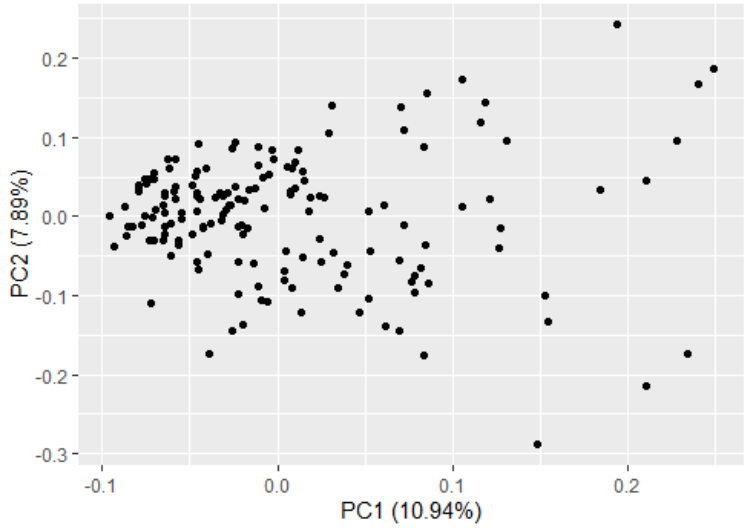


Figure 6.5 Participants did not fall into strong groupings based on two sub-scales in the MCQ (factors "PC1" and 2 "PC2").